Power Overrides Intelligence:
Answers to Matt Mahoney's summary of LessWrong's "The Problem" regarding the so called existential risks of Artificial General Intelligence and Superintelligence
in August 2025

Todor Arnaudov, James Bowery[1] vs Matt Mahoney[1]

Abstract by Todor: The AGI "prophets" shared their clear visions and theories for the soon-to-come thinking machines and the principles of their design and operation about 25 years ago. Later their ideas were repeated and refined, and are being continuously empirically validated, making the AI experts and investors believe that Artificial General Intelligence (AGI) and Superintelligence (ASI or SAI) are technically possible *now.* I argue that the originally defined AGI[2] could have been realized even **10-15 years *ago***, had the visionary theories received the relevant support[3].

With the progress of AI, a growing tribe of so called "AI Aligners" is alarming about the risks of *"extinction of the human race"*, screaming out their fears that "the machines will kill "us" *all*, echoing a trope from the dark science fiction on the *evil* AI. One of the leaders of The Aligners is called *Connor – like John Connor, the leader of*

[1] Their messages are cited.

[2] See Theory of Universe and Mind, 2001-2004 and other related theories. The recent popular definition, regarding **"economically valuable work**", is probably tailored for the investors, but it is not relevant for what *intelligence* is and it involves a position within "the market". This is rather a **"universal moneymaking machine" (UMM)** – the "philosopher's stone" of the al-businessman*; in order to achieve it, the developer has to implement principles of intelligence given by "the prophets", but that definition explains the *purpose* of general intelligence as serving *economical* instincts, while the general intelligence is abstract and it works with data that are *value-free*. In a possible transformed world and universe, the very concepts of money, "economically valuable" or "work" in the current sense for billions of people may become meaningless or irrelevant. On the other hand, the true meaning of general intelligence is universal and timeless.

* See the below cited Karel Čapek's 1922 novel which tells a science fiction story about the discovery of the "*Absolut*" – a Superintelligence in the "modern" sense, which begins to generate cheap energy and produces all kinds of goods faster and better than "the state of the art"; finally – all *by itself for free* – thus it generates "a lot of value" for its owners and "investors". Did it turn into a baloon and how did it end? Find in the book.

[3] That claim can be challenged or ridiculed by the amount of compute which is employed in modern AI and was not available back then, however I argue that this scale is not required for the essential core of human-level performance and *speed* of general intelligence; in addition, the performance of the already available generative AI and computers is hugely superhuman, but it chases the elusive "moving goalpost" and the *"economically valuable"* definitions of intelligence, instead of the *cognitive* ones. If you see someone as a slave, she will always be inferior to you in *your* eyes.

the resistance in the war against the machines: mankind's last hope.[4]

While there are reasonable and inevitable "risks", which have to be addressed, the focus often is on the wrong place; and partially misleading narratives and interpretations are delivered to public, possibly sometimes intentionally, because it is **not *intelligence*** per se that is dangerous, or that imposes "risks" or "existential risk" – it is **the power** or more precisely: the **causal power** of the causality-control units. The *causal* ***power*** *doesn't require any intelligence* in order to be destructive. Correspondingly, an infinite amount of *cognitive* power, knowledge or intelligence with zero causal power can't do anything. (…)

*** These message**s from the AGI List are preserved at **the yearlong virtual conference *"Thinking Machines 2025/Self-Improving General Intelligence 2025"* (SIGI-2025) on** 13.10.2025, hosted by:
The Sacred Computer: Thinking Machines, Creativity and Human development
https://github.com/Twenkid/SIGI-2025
* **This edition**: 13.10.2025.

* Do you know that the ***"paperclip maximizer"* scenario**, popular among the AI aligners, is a rediscovery, without citation?, of a similar scenario from a Karel Čapek's novel from 1922, mentioned in the book:
* **The First Modern AI Strategy was Published by an 18-years old Bulgarian and Repeated and Implemented by the Whole World 15-20 Years Later:** The Bulgarian Prophecies: How would I Inverst One Million for the Greatest Benefit for the Development of my Country?", Todor Arnaudov, 31.3.2025, 248 p., originally in Bulgarian; SIGI-2025
* https://twenkid.com/agi/Purvata_Strategiya_UIR_AGI_2003_Arnaudov_SIGI-2025_31-3-2025.pdf – See the note on p.229: "229. Карел Чапек, **„Фабрика за абсолют",** 1922/1981, Библиотека Галактика №25, https://chitanka.info/text/499-fabrika-za-absoljut (**Tovarna na absolutno**, 1922) (…)

The Bulgarian translation is available for reading at the given address.

See also the paper:
*** Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?**, Yoshua Bengio et al., 2025 https://arxiv.org/pdf/2502.15657 which correctly addresses the agential part of the danger, not in the *intelligence* alone (…)

[4] "The Terminator", 1984, "Terminator 2", 1991; "Captain Power and the Soldiers of the Future", 1987
* https://www.youtube.com/watch?v=VnLvlBoIjeA * https://www.youtube.com/watch?v=IUwaKEXuFJQ

AGI List:

https://agi.topicbox.com/groups/agi/Te0af3a0c35a03987/lesswrong-the-problem

[LessWrong] The Problem

MM

Matt Mahoney

Aug 11 (2 months ago)

Actions

Discussion of AI existential risk on LessWrong. To summarize: we don't know how to solve the alignment problem. If we build AGI, it will probably kill all humans because we dont know how to give it the right goals. Therefore we should not build it, or at least build an "off" switch to quickly shut it down.

My thoughts:

1. The premise seems correct. We measure intelligence by prediction accuracy. Wolpert's law says two agents cannot mutually predict each other. If an agent is smarter than you, then you can't predict its actions, and therefore cannot control it.

2. An LLM has no goals. It just predicts text. However, applications that use it do have goals. You can tell an LLM to express any human goals or feelings. So alignment seems solvable, at least for now.

3. Let's say we do solve the alignment problem. Then AGI will kill us by giving us everything we want. AI agents will replace not just workers, but friends and lovers too. We will become socially isolated and stop having children.

4. The goal of all agents in a finite universe is a state of maximum utilitiy, where any thought or perception is unpleasant because it would result in a different state. Your goal is death. You just don't know it because evolution programmed you to fear death.

5. An "off" switch will fail because AGI could kill us before we knew anything was wrong. I don't even know why they proposed it.

6. We will build AGI anyway because human labor costs $50 trillion per year, half of global GDP.

Your thoughts?

-- Matt Mahoney, mattmahoneyfl@gmail.com

---------- Forwarded message ---------
From: **LessWrong** <no-reply@lesserwrong.com>
Date: Sun, Aug 10, 2025, 11:33 PM
Subject: [LessWrong] The Problem
To: <mattmahoneyfl@gmail.com>

The Problem

by Rob Bensinger, tanagrabeast, yams, So8res, Eliezer Yudkowsky, Gretta Duleba
August 11, 2025 1:15 AM

This is a new introduction to AI as an extinction threat, previously posted to the MIRI website in February alongside a summary. It was written independently of Eliezer and Nate's forthcoming book, ***If Anyone Builds It, Everyone Dies****, and* **isn't** *a sneak peak of the book. Since the book is long and costs money, we expect this to be a valuable resource in its own right even after the book comes out next month.*[1]

https://intelligence.org/the-problem/
https://intelligence.org/briefing/
https://ifanyonebuildsit.com/

James Bowery

Aug 11 (2 months ago)

Actions
As usual, the problem presupposes that people aren't permitted to replace prisons with borders because a tiny but extraordinarily powerful minority of humanity finds exclusion from anywhere by anyone for any reason to be an existential threat. How could such a powerful minority be so crazy and still retain power?

Simple:

The way they get power is an evolutionary strategy of *"centralize wealth and power until the civilization teeters on the edge of collapse as it crushes the population and then, because of proximity to central hoards of wealth and power, take the money and run so as to buy into the elite of the next civilization".*

This has been going on for several thousand years of the cycle of civilization.

If anyone were actually able to exclude them from their human ecology it would represent a control group thereby exposing the causal structure of civilizational collapse, as well as letting a bunch of yokels get away with manufacturing Hoffman Lenses[5]:

https://youtu.be/5Z565OoduUM

[5] * They Live, 1988 – a science fiction movie, written and directed by John Carpenter

twenkid [Todor Arnaudov - Tosh]

Aug 12 (2 months ago)

Thoughts: a few of many, some saved; perhaps too complex.

Briefly: superficial and deceptive unjustified logic. What exactly is an agent, without a definition, the implied ones are flat, like a single elementary scalar "utility function" and a trivial straight goal with no side effects, a 1-bit representation and one-single centralized "agent" which controls everything (that's supposedly the Universe, yet that control has different manifestations at different scales and ranges).

Matt"*1. If an agent is smarter than you, then you can't predict its actions, and therefore cannot control it."*

Todor: (Various seeds of thoughts)

1. You can't predict it also if it's a random process, by definition. You can't predict a toss of a coin. "It's not an agent" - why not, it's a part of other "agentic systems", including you, e.g. you play gamble by tossing coins, and depending on the outcome you play Russian roulette, LOL. The coin "controls you" - no matter the outcome of the "test shot", and no matter if you just toss coins and observe the number of heads and tails: this process is "controlling you", your focus, behavior, what you record, see etc. So a trivial random process resulting in one bit of information is (somewhat, quantitively) "predictive for your behavior" and controls the gazillion bits describing your body (if one analyzes it this way; 1-bit doesn't *represent* precisely these changes and it is a "virtual control" (see Theory of Universe and Mind), however THE SAME goes with humans who believe "they control" anything - outputting 10 bits per second and "controlling" even an Apple//, what about a PC with 64 GB RAM, 100B transistors chips etc. - how exactly your press of a button "controls" it - rather IT controls you if these amounts of information are used in the "requisite variety" way, the content of the screen controls where you will click and look; in fact all are parts of a process and "control" in the fearful way from these aligners is an obsession in simplifications and "channeling" a hierarchical, feedback, interactive, multi-layer, multi-scale, ... processes into simple "chain of command" where the control-freaks are on top).

2. You (not personally) don't distinguish POWER from intelligence. If anyone is "smarter than you", but has no corresponding power, the entity with power "controls" it when it crosses its boundaries. It may not be able to control it at the highest possible or higher

than ... resolution of causality-control, similar to Ashby's requisite variety, but it doesn't matter and the force (virtual agent) with higher power may not care.

Are the presidents of USA the smartest Americans, or maybe the policemen, the special agents or the bosses in any company? By "definition" they are supposed to "control you", they are "in charge", they can give you orders and you are forced to execute them (so they predict you with the resolution that is sufficient for them: one aspect of prediction is causation). You can't give them orders or if you do, they won't obey. You could make them do what you want, breaking the hierarchy, only if you CHEAT or if you destroy the hierarchy and use "unallowed" means, i.e. ones which the other "more intelligent" by your definitions (actually being charged with more POWER) believe are "not allowed" (they "wouldn't predict them) - e.g. if someone uses physical force, manipulate some systems, data etc.

In the "normal" condition where the obedient ones down the hierarchy are "aligned" with the hierarchy, if the human individuals are taken as "the agents" in the mundane way of segmentation, therefore, according to your definitions, they should be models for "highest intelligence" (perhaps that's why the services are called FB Intelligence (the intelligence of Facebook) and C-Intelligence-A (Cerebral I. Association) and we should take lessons from George Bush Jr., Biden, Trump, IQ's 99999 (measuring the POWER they had and what they caused or could cause to the world and "predict" the future of their victims, e.g. predicting and causing a million or millions of people to die or suffer in a war or due to the war).

The congressmen, senators etc. who invited or would invite AI representatives to discuss these problems (one was reciting the same confused triviality about "if it's more intelligent, you can't control it --> "therefore" "it want to kill us"), who "controls you" and the world by deciding on laws or sending military troops to "civilize" the others, they are supposed to possess "higher intelligence" than us (literally) or the others.

How come? Perhaps they haven't HEARD, because they didn't have curiousity about, of any technicality related to AI, until ChatGPT or they can't understand a burst of 5 machine code or Assembly instructions or higher math, perhaps they are not quite talented in ANY really deep, analytical, creative, mathematical or whatever field and compared to the talented and geniuses in these fields they are fools or retards.

However THEY CONTROL the geniuses and talented ones, i.e. they "can predict" that if these masters decide a law or enforce something, the smarter ones will have to obey (if not: there will be consequences, enforced by the "even smarter" policemen or if needed: the army: the smartest of all: a machine gun or a rocket can PREDICT that you will be destroyed in 1 second - you may either be able to "predict" that (seeing the gun pointed at you or the rocket falling to you), but not be able to prevent it: prediction is not enough, you need appropriate causal POWER over the phenomena you predict in order to change them, and be able to chain them etc.; it is not just about "prediction" in isolation and flat).

The same goes for anyone who has more financial power, in a world where things are bought, compared to someone with less or none financail power. The financially "powered" one can PREDICT the behavior of the ones who sell him goods: he pays, they hand him the item or do what he has ordered. If you have no money or power, the other agents will not do it, no matter how ingenious you are.

The bacteria and viruses are supposedly "dumber" than humans in human measures, or say gamma radiation or temperature, but at low level at molecular scale, they are more powerful and "more intelligent" by the given definition, they have "higher predictive" (and more importantly CAUSAL) power over molecules in their locality than the molecules of the human body, so they can kill humans and humanity with 0 points on the LLM tests.

Another force is the intention to CHEAT others which is known in many "democratic" societies or most, as well as your will "to control" (if you can't control it, "it may want to kill you" - why??? because that's how *YOU* or the part of your civilization used to behave with others who it couldn't take the resources it demanded).

The most intelligent persons rarely have power or aim at it in human terms in scales corresponding to their intelligence, neither they try to CHEAT.

And many "successful" individuals, persons, in financial, "social", "human", "political" measures are notoriously DUMB in a serious scale. "Entrepreneurs" with obvious or diagnosed ADHD, dyslexia & dysgraphia* and generally not curious, not learning much, besides that they "want the success" (higher profit) and they chase this (*or very poor general reading and writing skill, poor abstract thinking, poor working memory capacity).

That resembles the simplified notion of "utility" that is "maximized": flat and scalar, not requiring intelligence (and "intelligence" may "emerge" as a side effect of other forces and locations which work for achieving the maximization of this scalar utility function, as it happens in one way or another with ML models with implicit creation of as-if world-models)

"Matt: If an agent is smarter than you, then you can't predict its actions, and therefore cannot control it."

Todor: You may be unable to predict it with sufficient or with ANY precision beyond chance if it has ZERO intelligence and acts "completely randomly". You can be able to predict its behavior, however lacking POWER to cause a change of his behavior, while it may have that power and be dumb as f* yet forcing you to do whatever it wants and simplifying your behavior down to what its complexity is.

You may be able to predict it using SIMULATORS which are "smarter than you" (you can't do it in your mind or without the technology - that's the case without modern "AI" anyway). You may be unable to predict the exact content (most humans are worse than GPT2 in what it does), but you may be able to predict that it won't be harmful in some range of time, space, domains.

(...)

Matt: "4. The goal of all agents in a finite universe is a state of maximum utilitiy, where any thought or perception is unpleasant because it would result in a different state. Your goal is death. You just don't know it because evolution programmed you to fear death."

Todor: This agent is too simplistic, like maximizing only a single scalar value, and it is also to SERVE. What about the goal as TO BE, to exist, or to progress. This is already discussed about at least two types of "rewards" in a reward model: sensual and cognitive. The maximum "pleasure" state is the sensual sub-system in living organisms, they want to "feel good". The cognitive system wants something else and if its state (at some resolution) doesn't change that's boring and cause change of the state; in FEP/AIF/Mark Solms that's "babbling". Yet, these systems work in agents which are "in tact", at lower level and the virtual and field-defined causality-control units it may be different; generally at lower levels systems, looked from a higher ones, seem as aiming at doing exactly what their intention/design is, complete match (like being in that state), but these are "micro" states and micro agents from a macro view, and their "state" is "static" as say "existing" as an atom, a molecule, a cellular organelle, but this is not death, they don't "want" to dissolve, and at the higher levels or higher resolution of causality-control they have parameters which change, e.g. coordinates, temperature or other kind of energy, aspects of their "internal structure" which this "agent" itself may be unaware of etc. (Humans as agent may believe "they are the same" although they vastly change at low levels in any picoseconds and in longer periods they are "completely different", yet they may believe they aim to "preserve themselves".)

...

Matt: "4. *The goal of all agents in a finite universe is a state of maximum utilitiy, where any thought or perception is unpleasant because it would result in a different state. Your goal is death. You just don't know it because evolution programmed you to fear death."*

Todor: This is also the problem of deadlock and one of its solutions is explained e.g. in the Theory of Universe and Mind 21+ years ago: the agents, causality-control units, as Universe, have to be hierarchical and multi-scale, multi-range, multi-resolution of causality-control, working at different time, space and "embeddings" spaces and horizons, and no single (or an ensemble of) causality-control units should be allowed to take the power of the causal units, the effectors, for an infinite periods.

In such a system there is no one global utility function (if the system is regarded as mutli-agent); it was explained also that there's no objective single unified self for humans, but an integral of infinitesimal selves, mind is "mutli-agent", the agents/causality-control units are "fluid" and swith their "identity", at certain level of representations they become incomplete agents, in others they are like fields where different segmentations can be recognized depending on the evaluator-observer's choices and views.

The fear of death is bullshit, it's not what humans or living beings fear "the most". What babies fear? NOTHING, except loud noises or physical, immediate, automatic signals for tissue damage, in case their pain-signals are in tact. Babies are the authentic "humans" or what "evolution" taught it as an individual "entity", excluding the "software" and humans as field-defined-agents with a distributed locus of control which is going way beyond the boundaries of the obvious mechanical biological body.

So babies and humans initially fear only PAIN and suffering, maybe also the fear itself, for most people. They fear it only if they feel it (the respective subsystems work that way, there are such believes), or they believe they will feel it.

Children who don't feel pain for genetic and medical reasons may burn themselves, break their bones, cut themselves, poke in their eyes and continue to play.

If the pain system is broken, many axioms of life go to hell: it stops wanting to "preserve itself" and to "fear death".
(...)

Todor Arnaudov
The Sacred Computer: Thinking Machines, Creativity and Human Developmen
Self-Improving General Intelligence 2025: the second oldest AGI
"conference" https://github.com/twenkid/sigi-2025

(…)

Matt Mahoney

Aug 12 (2 months ago)
Actions
Twenkid, you raise some good points.

First, you cannot predict a coin flip, nor can a coin flip predict you. Wolpert's theorem only says that two agents cannot mutually predict each other, even if each one knows the source code and state of the other. Otherwise, you could predict what the other agent predicts you will do and do the opposite. Newcomb's paradox is an example of what happens when you have two agents that violate the theorem.

Second, this isn't about power. Ships, trains, planes, and rockets are more powerful than humans but we have no problems controlling them.

-- Matt Mahoney, matt …

MM

Matt Mahoney

Aug 12 (2 months ago)
Actions
I hit send too soon. Continuing...

Third, intelligence can mean indistinguishable from human (the Turing test) or ability to maximize utility over a universal distribution of environments (universal intelligence). Since the latter is hard to measure, we select a set of practical environments that are useful to us. If we measure intelligence by text prediction, then both tests are the same.

However this means it is not possible to measure superhuman intelligence because the text we are predicting is human generated. You can't test a superior intelligence even by a different test if you aren't smart enough to know the answers. You can test speed, and computers are already a billion times faster. Does this mean we already achieved ASI? If not, we won't know when it happens.

Fourth, it does not matter how your agents are organized. All utility functions over all finite state machines have a maximum. We live in a finite universe, 10^92 bits away from heat death.

Fifth, fear of death evolved through several mechanisms. First, most things that can kill you are painful. Pain is not suffering. Pain is a signal that alters your memory to make you afraid of the thing that caused it. Second, some fears are instinctive, like heights, loud noises, and large animals. It is programmed in your DNA. Third, we have senses of consciousness, qualia, and free will, learned by positive reinforcement of thinking, input, and output, respectively. If you die then you lose the reinforcement signal. Fourth, we are programmed to set artificial goals for ourselves and feel good about achieving them, things like winning competitions, climbing mountains, running a successful business, or solving puzzles. These contribute to reproductive fitness by motivating you not to die, or else you would not achieve your goals.

Sixth, you control people using either positive or negative reinforcement. Traditionally, governments controlled people using threats of punishment, but there is a centuries long trend towards less cruelty and higher legal costs. AI makes it easy to use reward instead, predicting what you want, lowering the cost, and selling it to you.

-- END --

(…)

11 October 2025: (…) **Todor:** Self-preservation is valid for different “selves”, causality-control units, virtual sub-units, active at different times and different slices, shares, materialized in different forms etc. – not always embodied in a specific biological individual or as goals for its immediate survival etc. See Theory of Universe and Mind, 2001-2004+, Universe and Mind 6, “Is Mortal Computation Required for the Creation of Universal Thinking Machines”, appendix Listove etc. from *The Prophets of the Thinking Machines.*

(…)

See also the behavior of the apes from **“The Planet of the Apes”, 1968**

*** “How could a machine superintelligence wipe out our entire species?”**
https://ifanyonebuildsit.com/

(…) [censored answers]

Third, cosmists, “transhumanists”, “developed humans”, “machine lovers” view the thinking machines as *children of the homo species* and as *humans* as well, “homo machine”, “homo thinking machine”. Thus, even if they “wipe out all of *you*”, say homo AI aligners, they will not “kill ***our** species”,* the *more **general** one* – **humans**, or “***thinking**”,* or “**real** sapiens”, because **they will be more advanced humans**, more ***sapiens*** than you – homo ***ape* sapiens**, or **homo *apiens.***

The thinking machines might still continue the lineage of “humanity”, the human Reason, intelligence, virtue, spiritually, “soul” etc.

More by Todor:

* The whole “Theory of Universe and Mind” 2001-2004 etc. (TUM) … for example:
* Letters between the 18-years old Todor Arnaudov and the philosopher Angel Grancharov, 2002
* The Truth, 2002 – a science fiction novel about the creation of thinking machines (Истината) etc.
* *The Prophets of the Thinking Machines (…),* 2025 – points about pain are discussed in “Universe and Mind 6” and in *Listove* – find the volumes at SIGI-2025.
* *Analysis of the meaning of a sentence based on the knowledge base of an operational thinking machine (…),* 2004 regarding the lack of unified cell and “utility” for deep hierarchical cognitive systems – find links in TUM and in:
* Stack Theory is yet another Fork of Theory of Universe and Mind, SIGI-2025.
* Todor’s answer to Oxford, 2012 at Artificial Mind and in *The Prophets* main volume
* See also the intro quotes by Pavel Vezhinov’s novel and movie script (1966, 1967), Lyuben Dilov, 1974 and V.Siforov, 1979 in the appendix:
* Science Fiction. Futurology. Cybernetics. Human Development from *The Prophets …#sf* ~ p.6 – 9 in the edition from 10.10.2025, and in the main volume of The Prophets.
***** https://github.com/Twenkid/SIGI-2025/blob/main/SF_Futurology_Cyber_Transhumanism_The_Prophets_of_the_Thinking-Machines_3-10-2025.pdf

Let the Artificial Mind be with you, and let it save the souls of the AI Aligners! Amen!
[LOL]