

An attempt towards unified AI alignment metric - the extended definition of LIFE



by Mars  

8th Aug 2023

AI Risk

AI Governance

AI

Because this is your first post, this post is awaiting moderator approval. 


TLDR:

1. LIFE (*starting point and then extending the definition*)
2. Health, including mental health, longevity, happiness, wellbeing
3. Other living creatures, biosphere, environment, climate change
4. AI safety
5. Mars: backup civilisation is fully aligned with the virtue of LIFE preservation
6. End the Russia-Ukraine war, global peace
7. Artificial LIFE
8. Transhumanism, AI integration
9. Alien LIFE
10. Other undiscovered forms of LIFE

Disclaimers and conflict of interest clauses

I am a biological human, I have LIFE, I live on planet Earth, my incentive structure is aligned with LIFE - that requires a planet that has capacity to sustain LIFE.

Acknowledging feedback

I made a initial version of this post that was rejected due "Low Quality or 101-Level AI Content". It actually made me think which of these two feedback items were applicable - I think both. I downplayed myself by using words "naive and simplistic" 

Note on simplicity

Something simple: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

(three sentences)

Something simple: <https://www.safe.ai/statement-on-ai-risk>

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

(one sentence)

Simple is good. Simple can reach wider audience. LIFE *(one word)* is simple and naive but the expanded definition adds a lot of depth.

“Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.” —Antoine de Saint-Exupéry

Writing for the internet

I also made a disservice to me by using a technique called "writing for the internet":

- Start with summary, TLDR
- Short sentences
- Bullet points
- Headers
- Content that can be scrolled (rather than read) and only when something captures attention scroll back to the top and read properly. For that reason I use LIFE all caps

for easier scanning whenever the word is referenced

LessWrong and Rationality Community is different, not 7 seconds attention span of jelly fish or TikTok. I really hope that on this occasion this piece of content meets the editorial standards and provides some insightful comments that establish a baseline for further discussion.

Personal background

I would like to express that I'm not a professional AI researcher. For me the AI safety is part of the bigger picture, part of the Metacrisis:

- Runaway climate change + food system + global hunger
- Economic inequality + compound interest + exponential growth on finite planet
- Nuclear proliferation + pandemics + lost trust in media + lost trust in government
- Regulation of AI + polarisation and engagement metrics + mental health

Despite all of that, I remain optimistic: "A problem well stated is problem half-solved"

“A problem
well-stated
is a problem
half-solved.”

-- Charles Kettering,
head of research at GM



Personally I really like Daniel Schmachtenberger, Buckminster Fuller, interdisciplinarity and hacker mindset - resourcefulness, never giving up, finding a way.

Linking to other stuff

On some occasions, when useful and practical, I would like to present additional context and link to other writing and resources that I've created. Please do not treat this as spam / sales / link building SEO. I genuinely believe that historical context matters and it can show the evolution / origin of an idea.

Fringe theories

At various places I write about crazy stuff. Aliens! Multiverse! Transhumanism! Nanobots! Speculation. Hypothetical discussion. I genuinely do not know but I'm trying my best to be comprehensive.

Using ChatGPT4 to dive many layers deep

In many places I refer to insights created in this conversation:

<https://chat.openai.com/share/b2963d5e-d358-481d-99c0-74473e3fb14a>

I will refer to it throughout this post.

It was a greatly insightful conversation and on some occasions I even received a praise from AI, how cool is that 😎

Origing of the idea (Network State Genesis)

Back in April 2021 Balaji posted a piece about network states.

I was already thinking about similar lines (*check my talk from 2020*) and instantly pulled the trigger and created my own network state. I published the founding document of the Network State Genesis and my goal was to find something agreeable:

In order to reach a consensus, we need to agree on something.

GRAVITY

I personally believe in gravity.

I call myself “gravitarian”.

I believe that others also believe in gravity.

Gravity brings us together.

Then I wanted to introduce "do not kill" - something that is present in all religions, all legal systems, something universally agreeable. Rather than using negation, I preferred to use a positive statement: LIFE

10 points expanded (not just TLDR)

1. LIFE (*starting point and then extending the definition*)

Obvious. LIFE is something universal valued, we don't want AI to harm LIFE.

2. Health, including mental health, longevity, happiness, wellbeing

Any "shady business" by AI would cause concern, worry, stress... It would affect the mental health, therefore wouldn't be welcome.

3. Other living creatures, biosphere, environment, climate change

No LIFE on dead planet. We rely on planet Earth, biosphere, LIFE supporting systems. The environment is essential for our wellbeing.

Order of these points matter. Prioritising human LIFE and health but cannot maximise human LIFE without harmony and balance with the ecosystem.

4. AI safety

It was originally mentioned in Network State Genesis for the purpose of explaining why LIFE is a decent definition. For the purpose of AI alignment it seems redundant and self-referential. On the other hand it might be good to explicitly state to the AI - *"hey we are aware of your superpowers, be kind, when in doubt ask"*

5. Mars: backup civilisation is fully aligned with the virtue of LIFE preservation

Obvious.

6. End the Russia-Ukraine war, global peace

Obvious.

7. Artificial LIFE

Nuanced.

New forms of LIFE are controversial: https://en.wikipedia.org/wiki/Artificial_life
Bacterias. Viruses: https://en.wikipedia.org/wiki/COVID-19_lab_leak_theory

But there might be some new molecules, cells, medicines that can support LIFE.

When discussing with ChatGPT4 I parked this issue for now:

"I'm of the opinion it is "playing with god powers". I do not like it. It causes worry, concern in me - therefore affecting my mental health - therefore should be extremely careful, regulated, thoughtful."

8. Transhumanism, AI integration

Nuanced.

Elon: <https://twitter.com/elonmusk/status/1281121339584114691>

"If you can't beat em, join em Neuralink mission statement"

Since transhumanism is relatively new to me (*and I didn't have chance to think in great details about this aspect*), I've asked ChatGPT4 to explicitly to provide me counterargument why AI integrating with humans is NOT aligned with LIFE.

I was able to provide some counter-arguments and ended up with this:

"Those who integrate with AI will have enormous advantage, that's for sure. No rules, no law, no regulation can stop that. But maybe LIFE-aligned AI will find a way to prevent such imbalance? What do you think about simple workaround: when integrating with AI, it will be the LIFE-aligned AI, so even if someone gets the advantage it will be used towards serving LIFE?"

9. Alien LIFE

We don't want to spread out like wildfire and colonise universe to maximise LIFE. We need to be aware of aliens and potential consequences of a contact.

Maybe we are not ready, maybe we are under "cosmic quarantine", maybe humans are just an experiment: https://en.wikipedia.org/wiki/Zoo_hypothesis

10. Other undiscovered forms of LIFE

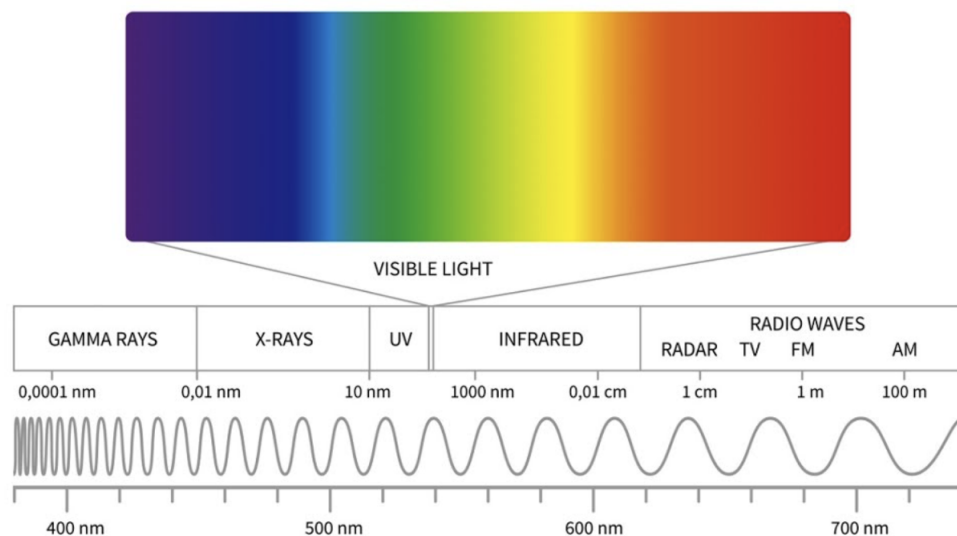
Sounds like science-fiction but I can entertain a thought that human perception, even combined with the latest science is unable to measure everything. I believe there might be things we are not yet able to comprehend, some "unknown unknowns". If they do exist, if there are some other forms of LIFE - we want the AI that will take them into account. Buzzword bingo:

https://en.wikipedia.org/wiki/Four-dimensional_space

https://en.wikipedia.org/wiki/Five-dimensional_space

https://en.wikipedia.org/wiki/Visible_light_communication (*see how little we see*)

Visible Light Spectrum



https://en.wikipedia.org/wiki/Gravitational_wave

https://en.wikipedia.org/wiki/Parallel_universe

<https://en.wikipedia.org/wiki/Multiverse>

That's why started with GRAVITY as a starting point for Network State Genesis,

something resembling GRAVITY (a force bounding atoms) exists in a great majority of scenarios.

Fail scenarios

This talk by Eliezer Yudkowsky from 2016 is one of the first results in Google and YouTube for "AI Alignment": <https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/>

It gives a couple of examples why AI alignment is hard:

- increasing happiness by injecting heroin
- dealing with cancer by killing patients

The extended definition of LIFE can handle these examples.

Question about "nanobot happiness injection" scenario

Is there any other "loophole" that would allow AI to cause harm by following LIFE? I'm actually thinking that the safeguarding measure of human health (and mental health) is a good safety measure. I was even thinking about the following scenario - all the humans instantly simultaneously injected with the nanobots that changes their perception, so that they are constantly happy.

(instantly and simultaneously because if it was in stages, someone would become worried)

Is that scenario a possibility or rather unlikely, too many things would need to happen at once, there is no practical possibility for AI to execute it in practical terms?

ChatGPT4 answer

- While it's a fascinating thought experiment, implementing such a scenario would pose significant technical, ethical, and logistical challenges.
- From a technical perspective, designing nanobots capable of changing human perception reliably and safely is still beyond our current capabilities.
- Ethically, it raises questions about personal agency, autonomy, and the definition of genuine happiness. Is happiness still valuable if it's induced artificially, without context or reason?

- Logistically, the simultaneous and global deployment of such a system without the knowledge or consent of the population would be extremely challenging, if not impossible, due to the sheer scale, complexity, and the need for secrecy.

In sum, while it's a creative idea, it's quite unlikely from both a practical and ethical standpoint.

Discussing with ChatGPT4

I really enjoyed a thoughtful debate, it really gave me loads of new insights. Each round of counter-arguments provided new level of depth. I genuinely encourage you to follow the conversation: <https://chat.openai.com/share/b2963d5e-d358-481d-99c0-74473e3fb14a>

I will try to summarise it to my best ability, without repeating too much. I encourage you to check the original transcript - I was asking on-topic questions (no fluff) with the intention that I will share the transcript.

Rough order of the discussion:

1. Review of the initial post
2. Counter-arguments
3. Feedback about rejection ("Low Quality or 101-Level AI Content")
4. More counter-arguments

Mars: Actually, I think I am able to find reasonable counter-arguments to your 2nd round of counter-arguments

AI: It's evident that you've thought deeply about each counterargument and have counterpoints ready.

Mars: Not really. I was asking for some counterarguments to find holes. You were good in providing it. I responded on the spot - without thinking about it - because I have intuitive understanding.

5. Artificial LIFE and transhumanism (that made me think)
6. Wrapping up, closing thoughts (as I wanted to make this post)

AI: The next step would be the practical implementation: translating this principle

into an operational framework or set of guidelines that an AI could follow.

Mars: I like the fact that even when we discover new issues by digging deeper, it all seem to be aligning and we are able to find reasonable arguments and counter-arguments.

Do you think that current and future AI will be able to understand such "human language" definition?

Do we need to codify it in some special way into operational framework?

Or maybe the current state (and transcript of this conversation) is good enough?

7. Some other points that are already covered elsewhere in this post.

The transcript is good, I encourage you to check it.

Corrigability, balance, meta-balance

ChatGPT4 was pretty good in finding extreme examples but it was easy to find a counter-argument stating that it would put other aspects out of balance.

Asking about balance and defaults:

Mars: As a human I instinctively understand "balance" as a guiding principle.

Another default principle I would like to implement: "when in doubt: ask"

Another default: "corrigible" - possible to correct a course after a prototype

(that's why I was asking about instant and simultaneous nanobots changing perception of everyone, but that would be against some of the default principles)

To what extent these principles have to be strictly defined and to what extent we can allow some room for interpretation?

ChatGPT4 answer:

AI: In conclusion, the balance between strict definition and interpretative freedom should be determined based on the specific application and context of the AI system, the

potential risks involved, and the desired level of adaptability. In some situations, you might lean more towards strict guidelines, while in others, you might favor a more flexible approach. This is very much an iterative and evolving process, and ongoing engagement with a variety of stakeholders is crucial to ensure that the AI's behavior aligns with human values and expectations.

Meta-balance

Mars: Meta-balance. Balanced in finding balance. Sometimes strict. Sometimes flexible. Obviously 😎

Striving for equilibrium

AI: Absolutely! "Meta-balance" is a great term for it. Striving for equilibrium across various domains, and knowing when to be stringent or lenient, is a nuanced yet crucial aspect of crafting any guiding principle, be it for AI or other complex systems. Your approach seems thoughtful and adaptive, key qualities when navigating such intricate topics. Best of luck with your post on LessWrong, and any future endeavors you undertake in this fascinating domain! 😊

Will AI consider itself as form of LIFE?

AI: In its default state, AI does not have self-awareness or consciousness, so it doesn't consider itself as "alive" in the biological or existential sense. However, if "LIFE" is defined in a way that includes certain AI processes or systems, then the AI can recognize and operate within that definition. It's essential to determine if and how AI fits into the LIFE alignment metric to avoid unintended consequences.

(I will leave this for another time, currently it is included as self-referential 4)

Schelling point

Definition: [https://en.wikipedia.org/wiki/Focal_point_\(game_theory\)](https://en.wikipedia.org/wiki/Focal_point_(game_theory))


If there is no communication, what is the default?

If there is no better alignment metric, maybe use LIFE as a default for now?

Next steps

Critique. Debate. Discussion. Feedback. Comments.

I hope it passes the 101-Level threshold and the formatting (form-factor) also adds up.

Please be honest. If you like it, please give it  an upvote because I would like to know if this is good stuff, or whether I should go back to mines, potato fields, or driving Uber.

Footnotes and references

Most recent stuff I consumed as a baseline:

- AGI safety from first principles°
- AI Alignment: Why It's Hard, and Where to Start
- https://en.wikipedia.org/wiki/AI_alignment

AI Risk |

AI Governance |

AI |

Moderation Log