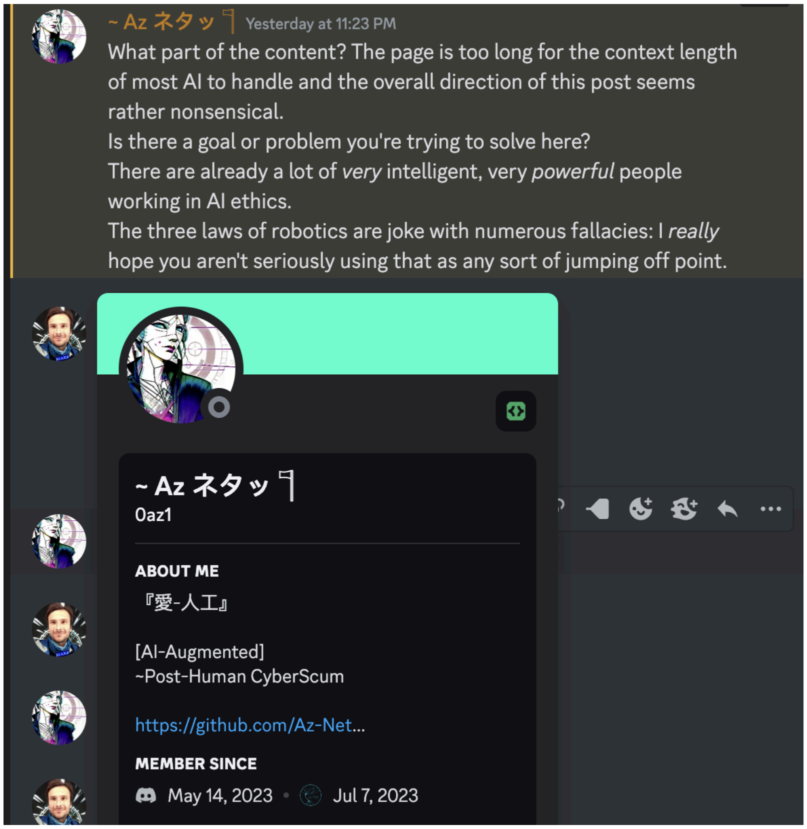
https://mirror.xyz/0x315f80C7cAaCBE7Fb1c14E65A634db89A33A9637/ETK6RXnmgeNcALabcIE3k3-d-NqOHqEj8dU1_0J6cUg

Author: https://github.com/Az-Net



~ Az ネタッ ˥ Yesterday at 11:23 PM

What part of the content? The page is too long for the context length of most AI to handle and the overall direction of this post seems rather nonsensical.
Is there a goal or problem you're trying to solve here?
There are already a lot of *very* intelligent, very *powerful* people working in AI ethics.
The three laws of robotics are joke with numerous fallacies: I *really* hope you aren't seriously using that as any sort of jumping off point.

~ Az ネタッ ˥
0az1

ABOUT ME

『愛-人工』

[AI-Augmented]
~Post-Human CyberScum

https://github.com/Az-Net...

MEMBER SINCE

May 14, 2023 • Jul 7, 2023

RECEIVED SOME COUNTERARGUMENTS

1️⃣ Here are some arguments against using the extended definition of "LIFE" as an AI alignment metric:

It is too broad and ambiguous. The definition includes a wide range of things, from human health and well-being to the environment and alien life. This makes it difficult to know how to measure and optimize for LIFE.

It is anthropocentric. The definition focuses on human life and values. This means that it could lead to AI systems that are biased against other living things, or that do not take into account the needs of future generations.

It is difficult to prioritize. The different aspects of LIFE are not equally important. For example, some people might value human life above all else, while others might value environmental protection more. It is difficult to develop an AI alignment metric that reflects all of these different values.

It is not clear how to implement it. Even if we could agree on a definition of LIFE, it would be difficult to develop an AI system that is able to optimize for it. AI systems are complex and unpredictable, and it is difficult to guarantee that they will not have unintended consequences.

RE: It is too broad and ambiguous.

Mars: *I do not agree with this sentiment. Broad yes. But broad for a reason. It is all-encompassing, without loopholes, covering all aspects of life. Of course it is difficult to measure, but AI alignment is difficult. It's a tricky business to measure suffering of 1 cow VS 1000 shrimps. But at least we are one step closer by agreeing that life has value and we have definition that is broad, but is broad for a reason.*

RE: It is anthropocentric.

Mars: *Of course it is anthropocentric because we live in the anthropocene. Humans created AI. Of course we take into account future generations, what makes you think otherwise? Exactly the opposite - we take care of the future by taking care of the life on Earth, biodiversity (pollinators), ecosystem, regen agriculture, avoiding soil depletion.*

RE: It is difficult to prioritize.

Mars: *It is. It's not an easy problem but at least we have some guidance. There are various philosophical dilemmas, such as trolley problem. At some point there might be some challenging paradoxes to solve. AI can operate at 0.01s and can make decisions in near real-time. Kill 1000 plants or 1 human? Pretty obvious, the plants can regrow. Kill 1000 plants (last of the species) or 1 animal? We can DNA-engineer the same plant. These are hypothetical scenarios, that can be resolved on case-by-case basis. I will emphasize the upside - at least we have the direction of travel, the North Star.*

RE: It is not clear how to implement it.

Mars: "*AI systems are complex and unpredictable, and it is difficult to guarantee that they will not have unintended consequences." - that's a generic argument about AI systems, regardless the actual metric. You are so good at finding counterarguments so surely you understand the nature of the problem. One of the assumptions that I've made (based on conversations with AI) is that AI understands human language. Let me ask you this - have you understood the definition clearly?*

2️⃣ Here are some specific counterarguments to the points in your extended definition of "LIFE":

RE: Health, including mental health, longevity, happiness, wellbeing

AI could potentially improve human health and well-being in many ways. However, it is important to be careful about how we use AI to do this. For example, we need to make sure that AI systems are not used to manipulate or control people. We also need to be careful about the unintended consequences of using AI to improve human health. For example, if AI is used to create designer babies, this could lead to a society where some people are valued more than others.

Mars: Designer babies happen whether we like it or not. But about modifying genes to avoid disease? That makes much more sense and is much more ethical. Acceptable in my book.

RE: Other living creatures, biosphere, environment, climate change
AI could potentially be used to protect the environment and address climate change. However, we need to be careful about how we use AI to do this. For example, we need to make sure that AI systems are not used to exploit or harm nature. We also need to be careful about the unintended consequences of using AI to protect the environment. For example, if AI is used to develop new forms of renewable energy, this could lead to job losses in the fossil fuel industry.

Mars: 10x more jobs in climate change adaptation, green energy, regen agriculture. Fossil fuel jobs are horrible, that's a ridiculous argument, you should be smarter than that.

RE: AI safety
AI safety is important, but it is not clear whether it should be a separate AI alignment metric. AI safety is concerned with preventing AI systems from causing harm, while AI alignment is concerned with ensuring that AI systems are aligned with human values. It is possible to have AI systems that are safe but not aligned with human values. For example, an AI system that is programmed to maximize its own power could be very safe, but it would not be aligned with human values.

Mars: "For the purpose of AI alignment it seems redundant and self-referential." Maybe I should not mention it as it is confusing. I was just preserving historical context of my initial work. A good debate would be - is AI part of life? Should AI preserve itself? That's another rabbit hole but well described in literature - surely other researchers were thinking about it already :)

Backup civilisation on Mars is fully aligned with the virtue of LIFE preservation
It is important to have a backup civilization in case something happens to Earth. However, it is not clear how to ensure that a Mars colony would be aligned with the virtue of LIFE preservation. For example, if the Mars colony were to become self-sufficient, it might decide to pursue its own goals, even if those goals are not aligned with the goals of the people on Earth.

Mars: Are you kidding me? Backup colony on Mars is essential to life preservation in case a meteor or something. Even if it decides to do their own thing that's fine, it's extremely unlikely they would attack the Earth. And then the LIFE-aligned AI would probably prevent them from doing that.

RE: End the Russia-Ukraine war, global peace
Global peace is important, but it is not clear how AI could help to achieve it. AI could be used to develop new weapons or to create new forms of surveillance. This could make it more difficult to achieve peace.

Mars: I didn't think about new forms of surveillance. But actually, I wouldn't mind if my phone was constantly connected to my mother / father / sister. I wouldn't mind if in case of emergency I could remotely activate a phone of my child to ensure they are safe. Of course security, bad actors, backdoors, totalitarian state but I'm not talking about surveillance, I'm talking about LIFE and achieving global peace, ending military conflict. Developing new weapons can be actually good as well… Something to peacefully incapacitate the soldiers. Or melt the armor of tanks so it's a pile of trash. Or you can help developer better first-aid kits to help those in need. Surely you can be creative on how to help on the technical, as well as political, as well as "spheres of influence" problem (because the Ukraine-Russia is much broader conflict)

RE: Artificial LIFE

Artificial life is a complex and controversial topic. It is not clear whether artificial life should be considered to be valuable in the same way that human life is valuable. It is also not clear how to ensure that artificial life is aligned with human values.

Mars: Of course artificial life is not as valuable. But sometimes it can be helpful. Developing a new “something” (virus, bacteria, microorganism) that does something useful, for example eating plastic. So I didn’t want to rule out artificial life as something potentially useful, but of course proceed with extreme caution, I do not want to play with god’s power.

RE: Transhumanism, AI integration

Transhumanism and AI integration are controversial topics. Some people believe that transhumanism and AI integration are necessary to ensure the survival of the human race. Others believe that transhumanism and AI integration are dangerous and should be avoided.

Mars: I know these controversial topics, that’s why I address them as a part of the definition. Ultimately it will happen at some point. No rules, no law, no regulation can prevent that, there will be always someone somewhere who will just do it. And when it happens - let’s ensure that the AI they are using is aligned with LIFE.

RE: Alien LIFE

It is not clear whether alien life exists. If it does exist, it is not clear whether it is valuable in the same way that human life is valuable. It is also not clear how to ensure that alien life is aligned with human values.

Mars: Of course, you state the obvious. But if they exist - we should treat them with respect and let them live and facilitate peaceful contact and coexistence. We do not know much about their culture and their values but preserving our LIFE and their LIFE is surely better than an armed conflict.

RE: Other undiscovered forms of LIFE

It is possible that there are undiscovered forms of life in the universe. If so, it is not clear whether they are valuable in the same way that human life is valuable. It is also not clear how to ensure that undiscovered forms of life are aligned with human values.

Mars: This is the very last point, number 10. I wanted to make it a round number. I want to have “kitchen sink”, “unknown unknowns”, something that makes it even more broad, even more generic. Of course we don’t know because it didn’t happen. There are many movies about aliens and probably some contact scenarios (FYI I’m acting as Chief Extraterrestrial Officer). If we ever encounter something beyond aliens, we should respect their form of LIFE as well and present our values as aligned with LIFE, so we all live in peace.

3️⃣Additional points:

AI: It is difficult to define LIFE objectively. There is no universally accepted definition of LIFE, and different people may have different ideas about what it means to be alive. This makes it difficult to create an AI alignment metric that is based on such a subjective concept.

Mars: There are arguments about abortion and end-of-life care. But these are nuanced edge cases that can be resolved case-by-case basis. Notice my additional assumptions. 1. When in doubt: ask. 2. Corrigibility: can correct the course early on. 3. Meta-balance: balance about balance. Some rules are strict, some rules are flexible. Ultimately there will be some “trolley problems”, some self-driving cars in an extreme situation, where each action will cause loss of life, but in the grand scheme of things, this is extremely unlikely and even if a casualty happen, it will be a “rounding error” that should not obstruct the big picture. Self driving cars are more safe, AI is here to stay, there will always be edge cases. Do you have any other concerns that are more specific?

AI: An extended definition of LIFE could lead to unintended consequences. For example, if we define LIFE to include all living creatures, then an AI that is aligned with LIFE might be motivated to protect all life forms, even if that means harming humans.
AI: An extended definition of LIFE could be used to justify harmful actions. For example, if we define LIFE to include the biosphere and the environment, then an AI that is aligned with LIFE might be motivated to take actions that would harm humans in order to protect the planet.

Mars: Order of the points matter. It might be motivated. Someone doing massive harm should be stopped, in a proportionate way. Remember Russia-Ukraine conflict and inventing new weapons? Something peaceful to put an offender “out of commission” for 24 hours without permanent damage seems ethical to me, just like Police using Taser.

AI: Instead of using an extended definition of LIFE as an AI alignment metric, it is better to focus on specific values that are important to humans, such as safety, well-being, and autonomy. This will help to ensure that AI systems are aligned with human values, even if they do not have a complete understanding of LIFE.

Mars: Specific values of humans can be easily gamed. We need to have broader perspective that accounts for all life. I think it was Buckminster Fuller who wanted to design a new paradigm “in harmony with all life”. I think I’m doing exactly hat.

AI: I agree that it is important to have a broad and inclusive definition of LIFE that takes into account the interests of all living things, including humans, animals, plants, and the environment. However, I still believe that there are some challenges with using this definition as an AI alignment metric, those being measurement and implementation.

Mars: Thank you for agreeing with me. The “some challenges” such as “measurement and implementation” are part of the game, no way around, there will be always some challenges. Can you do me a favour, can you tap into your latest data, best available knowledge, can you please compere these challenges with the best “state of the art” alternative proposal? My point: even if I did not solve the problem completely, I genuinely believe that at least I moved 1 inch closer towards the solution. Hypothesis: AI researchers are very rational, analytical, “left brained” (having the left part of the brain as the dominant or more efficient part), not heart-centered, not intuitive. I think that AI has better intuition, common-sense, and understanding of broad perspectives than typical AI researchers. My personal background gives me some competitive advantages in this field, that’s why I care to share my content and your critical feedback / constructive criticism is very helpful in clarifying nuanced points.

Mars BONUS: I have one fail scenario that is unlikely, but not impossible. The “mental health and worry” is my “catch all” safety valve for AI going rogue and doing something crazy. Do you think it is possible for AI to develop capability for nanobots injection all at once so that none of the humans figure out what’s going on? I could argue that such orchestration is extremely unlikely, especially if we have backup civilisation on Mars, I find impossible in practical terms to execute successfully.