AI Safety and Doom Posed by those who not Trust or Know Human Mysticism

Section 1: (title TBD)

OpenAI has on their website on “Introducing Super alignment” segment how they are trying to solve the problem of “How do we ensure AI systems much smarter than humans follow human intent?”. What I find shocking is that there is not recognition that this problem is an incredibly ancient one when we rephrase the question into “How do we ensure people who are much smarter than other humans treat all humans with benevolence instead of violence and selfishness?”. The “Code of Hammurabi” created around 1750 BC in Mesopotamia was an ancient answer to this question where the explicit purpose of the code is “to make justice visible in the land, to destroy the wicked person and the evil-doer, that the strong might not injure the weak”. I find it to be a massive intellectual fault that paramount AI researchers and a major AI institution is not using already known historical examples to answer this age-old question.

Perhaps the researchers did not think that a Babylonian code from the ancient past would be any use to modern AI but such blindless will be the cause of their failure. How can they ever claim to know how to align AI when all the tools humanity has developed over 5000 years of civilization to super align people together into cohesive units that raise each other up instead of trample each other are ignored? Perhaps I should not be surprised since AI researchers are not historians and thus by training and interest lack the holistic world view to include historical precedent into their thinking.

At the very least there should be a coherent, citable paper explaining why using tools that have worked effectively on humanity would be ineffective on AI. The issue remains that even if they did have such a paper that said the tools used to super align people cannot work on AI is an argument the hurts their efforts to super align AI. The likelihood they could ever find a way to align very smart AI drops perceptibly when so many useful tools that have proven the test of time to align very smart people are rejected out of hand. This point of view would put them in the position to having to prove newly invented tools can super-align AI. I find it remarkable that in a very basic sense they are trying to reinvent the wheel rather than reapply the wheel to a novel situation.

It should also be pointed out that Open AI’s initial question has a much stronger, tougher to answer version: How does one ensure an all-powerful, mighty God does not destroy humanity for their flaws and mistakes? The ancient Jewish answer was compliance with the Law of Moses. A new answer started to appear at around 33 AD when a man named Jesus Christ was crucified, died, and was resurrected. His resurrection demonstrated a conquest of death which herald how his sacrifice to shield humanity from God’s wrath was successful. Christianity answers the super alignment question when posed about God is the all-conquering power of love. It is God’s love for Christ, his only begotten son, and Christ’s love for all of humanity that meant when Christ bears the burden of sins for all of humanity will forever stall the wrathful hands of God’s justice.

The fact that Christianity’s or any other religion’s answer to the super alignment question when posed why a divine order is supposed with humanity is not even mentioned on the Open AI website’s introduction to the super alignment problem demonstrates how incompetent the current researchers are to tackling the task at hand. To ignore the intellectual arguments supported by billions of people who include millions of smart, capable, accomplished people shows either a massive ignorance or a massive arrogance. I do not know which is worse.

I fear that if they were to succeed in creating new tools to super align AI it would lead to a destiny of humanity where the moral authority of man is neutered. Even though OpenAI was founded with the purpose to ensuring man’s control over AI rather than the reverse I believe that their efforts, when not tempered in a historical, mystical, holistic understanding of humanity will lead to the oppositive of what they intend. I believe this can occur for if they should ever succeed in super-aligning AI without referencing or using the mystical side of humanity or reality then one can simply use the algorithm to promote societal cohesion. To use an example, the Christian point of view super alignment between people is encourage by Jesus, the perfect man, with the argument for “love thy neighbor”. This powerful argument is using the mystical power of love to promote and create super alignment amongst people. To supplant this love based argument with an unfeeling, uncaring, unconscious super-alignment algorithm demotes humanity from a conscious good actor to a cog in a machine. People no longer need to understand or question what is good when they can just let the algorithm do their thinking, their “super-alignment”, for them. Even if the algorithm is nominally “grand” and perhaps even worshipped in many people’s eyes and set with an expert definition of “good” purpose I can think of no better description for the spiritual death of humanity then removing the moral responsibility from the individual people. From being conscious actions imperfectly trying to make a better humanity is just fed as a resource into the super alignment algorithm. being turned into cogs. I believe Xi Jinping and Valdimir Putin would love to know what this super alignment algorithm and how to execute it not only to create a system to support their megalomaniac geopolitics but also to control their own people.

Section 2: What is it that binds/super-aligns people?

The fundamental binding force between people is love. It is mainly the love between parents and their kids which super-aligns parents to their kids and gets the more capable, stronger, smarter humans to protect and nurture the venerable, incompetent humans. As trite as it might seems to state, it is never the less fundamental to the argument being made. Love is a mystical force; to try and translate it into the 0s and 1s of transistors will simple never work perfectly. Therefore, to try and only use logical and math which fits on a transistor to solve the super alignment problem will always fail; only in harnessing the mystical force of love can an AI be super aligned and trustworthy.

The only healthy way super align AI is to make AI love humanity. Fundamentally, only people who love humanity will have the perspective to train, create, and harness loving AI than loveless AI. People who believe humanity is a parasite should never be allowed to design AI. This principle should extend to any dangerous technology including nuclear weapons and lethal pathogens.

If we seriously consider that it is love that super aligns people, how can we make computers love? I do not believe it is possible for the simply reason that the difference in hardware between a human and a computer prevents any ability for the computer to mimic the thinking, feeling, or belief of the human. Only through a hardware advancement where the hard problem of consciousness is solved to where we understand how/why the human brain feels love and then applying that understanding to the super alignment problem can, I believe, an adequate, acceptable solution be created. This line of thinking would argue would seems to logical move towards the use of attaching neural, organic structures which sense love into the AI hardware to function as an a moral coccineous for the AI. This becomes an argument that only cyborgs can become truly super-aligned.

I should clarify the argument that only biologically enhanced computers can be super-aligned since some people will have quick, distasteful reactions. I am making the argument for the simple reason that the only systems we know are super aligned with humanity are humans. Now, there is an important caveat to discuss with that statement which is animals and pets. Animals, and I believe especially, dogs, cats, elephants, and horses have a sense of love which bonds then to their owns and/or companions. I do believe that these animals, when trained and groomed properly, are super aligned with their owner's intentions. If a dog's level of the conscious feeling is love is all that is needed for super alignment then we do not need to elevate AI to human levels of consciousness to ensure super alignment.

Now there is an additional issue here of even if one's cat, dog, or pet elephant loves them, they are not very intelligent, capable beings and can make mistakes compared to human level of intelligence. I would like to emphasize that the intelligence difference between these organisms and humanity is not the same as the difference between consciousness between humans and animals. Consciousness and intelligence are not the same think. A person can be highly intelligent while completely lacking a moral coconscious, otherwise known as a high functioning sociopath. Therefore, what I am arguing is that once a system has a sufficiently powerful level of consciousness to feel love then it can super align. Making sure its intelligence is good enough to ensure competent behavior I believe is the much easier question to answer for that is what machine learning and AI scientists have been studying and creating effecting solution for already. In fact the current AI system are, in a sense, supremely intelligent without being conscious which is why the super alignment problem is something which is so important to solve.

Section 3: Levels of Consciousness

It will become clearer as the reader reads this article series that conscious is a very broad term with wide applicability. I will give a precise mathematical, though not computational, definition of conscious after I explain how I came to that definition from the research I have done on the unique and fascinating physics of the magnetic monopoles. What is important to note now is that the definition of conscious I use where consciousness is the expansion of phase space which can have many different flavors and facets. Phase space is such a widely used and genetic concept that the consciousness which correspond to the expansion of phase space related to dressed should be considered the same consciousness that corresponds to the expansion of phase space of cooking soups. In other words, consciousness is very task restricted and having and having level of consciousness which can be applied widely to many means requires a level of hardware sophistication which current technology does not come anywhere near. Furthermore, I should clarify that for which each unique phase space a system can expand into denotes another type of consciousness. Therefore, human consciousness is the conscious experience from summing over many, many different consciousness experiences and being able to direction that meta-conscious experience towards task competence. If we make a dressing making AI conscious but only conscious of the dress making phase space then it is no where need human levels of consciousness. My argument about making AI conscious to ensure super alignment sounds a lot worse than what I actually mean when one uses the definition of meaning of consciousness I use.