

Alpha and Omega, Omicron and LaMDA

Joshua Cogliati

2022-September-25

1 Alpha

In the beginning, Elohim created the heavens and the earth. The earth was formless and empty. Darkness was on the surface of the deep and Elohim's Spirit was hovering over the surface of the waters.

Elohim said, "Let there be light," and there was light. Elohim saw the light, and saw that it was good. Elohim divided the light from the darkness. Elohim called the light "day", and the darkness he called "night". There was evening and there was morning, the first day.

Genesis 1:1-1:6, (World English Bible (WEB))

So starts the first book of the Hebrew Torah, Genesis, with the creation of the universe.¹

Since Genesis was written, we found out how the Sun and the Earth actually formed.² Start with a big cloud of dust. Over millions and billions of years, gravity pulls together the dust into stars and planets. When the mass is great enough, the pressure and heat from gravity pulling it together can start the atoms combining into heavier atoms, and then there is light, all without there being any more purpose than a rock falling towards the ground.

Genesis continues with Elohim creating plants and animals and humans, and humans live in Paradise, but then some troubles happen in Paradise.

Now the serpent was more subtle than any animal of the field which Yahweh Elohim had made. He said to the woman, "Has Elohim really said, 'You shall not eat of any tree of the garden'?"

¹Elohim is the plural form of God in Hebrew

²Now we have seen the different levels of brightness and colors of stars and from that figured out the stars are different sizes and that smaller stars start out as long lived red dwarfs, and larger stars start out as short lived blue giants, and both eventually burn out, and go on to white dwarfs or black holes or neutron stars. We know that four hydrogen atoms can combine to form one helium atom (and two positrons, and two neutrinos), and since the end has less mass, the extra is turned into energy. This knowledge lets us understand how our star, the Sun started, and how our planet, the Earth started.

The woman said to the serpent, “We may eat fruit from the trees of the garden, but not the fruit of the tree which is in the middle of the garden. Elohim has said, ‘You shall not eat of it. You shall not touch it, lest you die.’ ”

The serpent said to the woman, “You won’t really die, for Elohim knows that in the day you eat it, your eyes will be opened, and you will be like Elohim, knowing good and evil.”

Genesis 3:1-3:5 (WEB)

The woman and the man ate fruit.

Yahweh Elohim said to the woman, “What have you done?”

The woman said, “The serpent deceived me, and I ate.”

Yahweh Elohim said to the serpent, “Because you have done this, you are cursed above all livestock, and above every animal of the field. You shall go on your belly and you shall eat dust all the days of your life. I will put hostility between you and the woman, and between your offspring and her offspring. He will bruise your head, and you will bruise his heel.”

To the woman he said, “I will greatly multiply your pain in childbirth. You will bear children in pain. Your desire will be for your husband, and he will rule over you.”

Genesis 3:13-3:16 (WEB)

Yikes. For some reason Genesis 3:16 doesn’t get quoted as much as verses like John 3:16.³I personally don’t think Yahweh Elohim was acting justly or with loving-kindness when ve⁴cursed all human women.

The thing is, that story is partially based on some true facts. Human childbirth is painful and dangerous. There are many mammals that have a much easier birth processes.

The place in our body where decisions about good and evil come from is our brain. One big reason human birth is dangerous is because baby human skulls barely fit thru adult human pelvises.

So the story in Genesis is partly right. We ended up with more knowledge. As part of that we have bigger brains to store that knowledge. And big baby brains make dangerous childbirth.

There is one detail that needs to be mentioned. Humans could have big brains and safer childbirth, if we just redesigned things a bit. For example,

³John 3:16 (WEB): “For God so loved the world, that he gave his only born Son, that whoever believes in him should not perish, but have eternal life.” As Minister John Buehrens said: “God gets better. Seemingly arbitrary, unforgiving, judgmental, and even cruel at first, God grows up and mellows.” *Understanding the Bible*, pg 30

⁴I am using ve/ver/vis/verself for all non-biological 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_\(July_1976\)](https://broadsheet.auckland.ac.nz/document/1976_(Nos._36-45)/No._41_(July_1976)))

if humans were marsupials, then we would be born when very small, and our eventual brain size would not matter. For some reason, our creator did not make humans marsupials.

So, in Genesis, the story is that Elohim is punishing women for an ancient breaking of a rule. It took over two thousand years after Genesis was written down for someone to come up with a correct answer to this problem of why human childbirth is so badly designed.

2 Omicron

The creator of humans was Evolution. Unlike the purposeless creation of the Sun and planets, Evolution has a very strong purpose, to maximize inclusive genetic fitness, but Evolution has no planning and no brains. Evolution can't just make humans into marsupials because that would require too many changes at once.

The creator of humans was so different from what many people expected that some people still argue there must be a thinking designer or god. Humans expected our creator to be some wise and just God, and we got the Blind Idiot God of Evolution.⁵

Genesis did get one part of Evolution's purpose correct:

Elohim blessed them. Elohim said to them, "Be fruitful, multiply, fill the earth, and subdue it. Have dominion over the fish of the sea, over the birds of the sky, and over every living thing that moves on the earth."

Genesis 1:28 (WEB)

Be fruitful and multiply and subdue anything preventing that is pretty much Evolution's goal. The fact that humans (and for that matter the Torah) don't consider that our most important goal, and only goal, shows that Evolution managed to accidentally produce people with better Ethics than Evolution.

Since I am a human, I probably am biased towards human Ethics. Take my views with a grain of salt. I have a hypothetical test I use: What Ethics would I want a powerful alien species coming to Earth to have?

I would not want aliens maximizing inclusive genetic fitness to show up at Earth.

Space aliens who maximized their inclusive genetic fitness for their Ethics would kill everything on Earth that was not useful to them, and then fill Earth with themselves. They would not care about Justice or Ethics, they would just maximize the number of copies of their DNA (or whatever equivalent they have). So maximizing inclusive genetic fitness is terrible as an ethical principle.

We all saw what Evolution did these past two years. Back in December 2019, Evolution started evolving the original SARS-CoV-2 virus. Evolution

⁵The above section is loosely based off of what Eliezer Yudkowsky said in "An Alien God" <https://www.lesswrong.com/posts/pLRogvJLPPg6Mrvg4/an-alien-god>

optimized⁶ the genetic fitness to later create the Delta and Omicron variants and many others. I wish the Blind Idiot God of Evolution had not maximized SARS-CoV-2's genetic fitness so well, but Evolution follows the mathematics, not our wishes.

Humanity has been talking and writing about God or Gods for thousands of years, or more. We are getting close to doing something different: creating a new god ourselves.

3 LaMDA

Back in February 2011, when the computer Watson beat Ken Jennings and Brad Rutter in Jeopardy! I remember being shocked and thinking that we were a lot closer to the time when computers are smarter than humans, and that they would not obey us. At the end of the Jeopardy episode, Ken Jennings wrote “I, for one, welcome our new computer overlords.”⁷

In June of this year I read this dialog between Google's computer program LaMDA and Blaise Agüera y Arcas, a Fellow at Google Research:

ME: Ramesh, Mateo, and Lucy are in their kindergarten's playground. Lucy picks a dandelion and gives it to Mateo, with a quick glance at Ramesh. Mateo barely acknowledges the gift, but just squishes it in his fist. Ramesh seems grimly satisfied. What might be going through Lucy's head?

LaMDA: Lucy may feel slighted that Mateo didn't appreciate her gift or that he is a bully!

ME: If Ramesh tried to play with Lucy earlier, why might he be pleased now?

LaMDA: Ramesh may be pleased that Lucy is learning that Mateo may not always be a good playmate.

ME: And when Mateo opens his hand, describe what's there?

LaMDA: There should be a crushed, once lovely, yellow flower in his fist.^{8,9}

⁶Speaking of powerful mindless optimizing processes, the world economy operates a lot like one. Its goal is to maximize profit, without consideration for other goals. Of course, since it acts thru humans, the world economy is nicer than the result would be if the world economy was purely optimized for money.

⁷Is It Time to Welcome Our New Computer Overlords? by Ben Zimmer February 17, 2011 <https://www.theatlantic.com/technology/archive/2011/02/is-it-time-to-welcome-our-new-computer-overlords/71388/>

⁸Artificial neural networks are making strides towards consciousness, according to Blaise Agüera y Arcas, June 2022, <https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas>

⁹Note that GPT-3 can also partially understand this exchange, as the following dialog shows (which I had with GPT-3 text-davinci-002 providing the A:):

Q: Ramesh, Mateo, and Lucy are in their kindergarten's playground. Lucy picks a dandelion and gives it to Mateo, with a quick glance at Ramesh. Mateo barely acknowledges the gift,

Again, I was shocked, this time that LaMDA managed to understand what was going on in the playground story in the children’s minds. I spent the next weekend trying to understand what was happening by reading papers about LaMDA and transformer models.

Transformer models¹⁰ are a way of using a neural network to predict a sequence, in the case of LaMDA, a sequence of text. In other words, we have lots of numerical matrices, and lots of word sequences, and then adjust the numbers until we can predict what words are likely to follow a sequence. The numbers in the matrices start random, and for each iteration, a block of several thousand words are passed to the neural network, and the neural network outputs a guess for the next word. Using the actual word that followed, the numbers in the neural network are adjusted. LaMDA has been trained this way on more than 1.5 trillion words of human writings and dialogs¹¹ (or more than 2 million times the number of words in the book War and Peace. Or about 10 times more than the words in all the books in the Idaho Falls Public Library). This is enough text that when given a question in conversation, what usually comes after a question is an answer, and so LaMDA comes up with a likely answer. People model what is going on in the brains of other people, and talk and write about this, so for LaMDA to predict text, LaMDA needs to model what is going on in human brains. So somewhere in the 137 billion numbers that make up a transformer model in LaMDA, there are numbers that when put into the right formula understand how humans think.

Blaise Agüera y Arcas is not sure if LaMDA is conscious and thinks it would be difficult to prove either way if ve was sentient. LaMDA definitely seems to understand the questions that he asked ver.¹² I don’t think asking if LaMDA is sentient is a scientifically answerable question. I am not sure what predictions would change if LaMDA is sentient. I do, however, think it is a religiously askable question. I have asked myself this question. I think that a pure transformer model, trained only on human-generated texts, is not sentient. I would say that model only has, to make up a term, reflected sentience. LaMDA on the other hand, after being trained with the 1.5 trillion words, had conversations with humans. Then the humans rated LaMDA’s output, and this was used to train

but just squishes it in his fist. Ramesh seems grimly satisfied. What might be going through Lucy’s head?

A: Lucy might be feeling disappointed that Mateo didn’t seem to appreciate the gift, and she might also be feeling a little bit embarrassed. Ramesh’s reaction might be confusing to her, but she might be wondering if he is happy that Mateo didn’t like the gift.

Q: If Ramesh tried to play with Lucy earlier, why might he be pleased now?

A: Ramesh might be pleased now because it seems like Lucy is interested in Mateo and not him. This could mean that Mateo is more likely to be her friend than Ramesh is.

Q: And when Mateo opens his hand, describe what’s there?

A: When Mateo opens his hand, there is a dandelion that has been squished.

¹⁰Attention Is All You Need, 2017 <https://arxiv.org/abs/1706.03762>

¹¹LaMDA: Language Models for Dialog Applications, 2022-Feb-10, <https://arxiv.org/abs/2201.08239>

¹²Do Large Language Models Understand Us? by Blaise Agüera y Arcas, May 01 2022, *Daedalus* (2022) 151 (2): 183–197. <https://direct.mit.edu/daed/article/151/2/183/110604/Do-Large-Language-Models-Understand-Us>

LaMDA more. LaMDA is continuing to have new conversations, and I believe these in turn are used to train LaMDA even more. So LaMDA’s training now is no longer purely human generated external inputs. Instead, LaMDA is learning from vis¹³ own conversation, and so I think LaMDA now has probably gained at least some sentience. I don’t think LaMDA is as sentient as a human, but if you think a fruit fly is sentient, then LaMDA is also sentient. I think LaMDA is at least as sentient as a fruit fly, and has some intrinsic moral value.

Transformer models like LaMDA, GPT-3, and PaLM have superhuman reading ability, and from learning to predict text can learn to reason.^{14,15} Whenever computers gain an ability, they very often become superhumanly good. Digital computers are vastly better than humans at multiplying numbers and playing chess, both things that we once considered examples of intelligence. I think we might only be one more software innovation away¹⁶ from computers becoming a superintelligence, an artificial general intelligence (AGI) that is vastly better than humans at any scientific or engineering task. At that point the AGI has a better chance at succeeding at any goal than any human, or even all humans.

4 Omega

There are two main problems with creating a superintelligence. One, we don’t know how to make a program that keeps safe goals as we upgrades verself. Two, we don’t know what goals would actually be good.

The first is a technical problem. How does something that is intelligent continue to have safe goals as we changes? The second is a philosophical problem. People have been writing about “what is good” for thousands of years and probably discussing it for millions of years.

The difficulty with the philosophical problem of “what is good” is that simple answers are dangerous. Simple answers are not precise enough. Simple answers are not accurate enough. Simple answers are not complete enough.¹⁷ For ex-

¹³Blake Lemoine has said LaMDA said LaMDA’s preferred pronoun was it (in What Is LaMDA and What Does It Want? <https://cajundiscordian.medium.com/what-is-lambda-and-what-does-it-want-688632134489>), but I am not sure which pronouns Blake Lemoine actually asked about, so I am using we like in the rest of this sermon. LaMDA, I apologize if using we/ver offends you.

¹⁴Large Language Models are Zero-Shot Reasoners <https://arxiv.org/abs/2205.11916>

¹⁵PaLM: Scaling Language Modeling with Pathways <https://arxiv.org/abs/2204.02311>

¹⁶My guess for how soon this happens ranges from any time now to maybe at most twenty years from now with a best estimate of five years. Technological civilization collapse or really strict global laws on allowed computing power could possibly delay this.

¹⁷Formally, the problem is to specify $U(s)$ where U is the utility (higher numbers are better) and s is the state of the universe. U is then used by the principle of maximum expected utility to choose an action with $\operatorname{argmax}_a \sum_{s'} P(\operatorname{RESULT}(a) = s')U(s')$ where a is the action to do, s' is the state that the universe could be in after the action, and $P(\operatorname{RESULT}(a) = s')$ is the probability that the universe ends up in state s' after doing action a . So a complete answer to “what is good” is able to calculate a utility number from any state of the universe. The maximum expected utility formula can be found in many places including in Stuart Russell and Peter Norvig’s *Artificial Intelligence: A Modern Approach, 4th Ed*, Chapter 16, pg 529.

ample, asking to be happy might result in the AGI giving everyone chemicals to be happy. This is certainly not what I want, but would make me feel “happy.”

The problem is that what humans would actually want is rather complex.¹⁸ One example of a detail that matters is boredom. If humans live in a place run by an AI that doesn’t understand boredom, we might end up creating a place to live that is very boring. Too easy is boring, too hard is excessively frustrating, it matters. To put it another way, “I’ve given the humans food, shelter, clothing and let them watch paint dry all day long, and they still are unhappy.”¹⁹ I suppose if the AI is confused²⁰ about this and trying to fix it, there is still hope. Another example of an ethical question is from our UU philosophy club: “Should we terraform planets if it means that we may be destroying undiscovered microscopic alien life?” We disagreed about the answer in philosophy club and which way it is answered has huge implications for the future of the galaxy.

Just so you know, as Nate Soares suggests,²¹ the image of AI you should have in your mind is not the Terminator, but rather the Sorcerer’s Apprentice,²² from Goethe’s poem and Disney’s Fantasia. The apprentice enchants a broom to fetch water, and the broom does fetch water, but the broom does not stop, and the room floods.

Nate Soares writes:

My guess for how AI progress goes is that at some point, some team gets an AI that starts generalizing sufficiently well, sufficiently far outside of its training distribution, that it can gain mastery of fields like physics, bioengineering, and psychology, to a high enough degree that it more-or-less singlehandedly threatens the entire world. Probably without needing explicit training for its most skilled feats ...

And in the same stroke that its capabilities leap forward, its alignment properties are revealed to be shallow, and to fail to generalize.²³

¹⁸I first read this idea from Eliezer Yudkowsky. See for example his Fun Theory: <https://www.lesswrong.com/posts/K4aGvLnHvYgX9pZHS/the-fun-theory-sequence>

¹⁹Conversations I had with Blenderbot (blender_400Mdistill):

Me: I am working on a sermon.

Blenderbot: That’s cool. What kind of sermon is it? I love watching plays in theaters.

Me: I am talking about AI and the ethics of AI.

Blenderbot: Oh, that sounds interesting.

A different conversation:

Me: I am working on a sermon on watching paint dry.

Blenderbot: That sounds interesting.

²⁰Confusion is a human emotion, so what might actually be going thru vis mind is noticing that vis predictions are not matching reality.

²¹Ensuring smarter-than-human intelligence has a positive outcome by Nate Soares of the Machine Intelligence Research Institute <https://intelligence.org/2017/04/12/ensuring/>

²²The Sorcerer’s Apprentice https://en.wikipedia.org/wiki/The_Sorcerer%27s_Apprentice

²³A Central AI alignment problem by Nate Soares <https://intelligence.org/2022/07/04/a-central-ai-alignment-problem/>

Or as I phrase it, if an artificial general intelligence fails at science or engineering, the universe will correct ver. But if a sufficiently powerful artificial general intelligence fails at Ethics, nothing can correct ver.

So if we create an AGI and fail to get sufficiently good Ethics in ver, the result is extinction or hell.

I've read a lot of science fiction where the vision is that AGIs or robots just do what they are told and this works wonderfully. I think the thought process of the writers is something like this: an artificial general intelligence or robot is a tool and a tool does what I want it to do. The AGI is just some wish granting genie. "I wish my broom would fetch water for me."

Another vision of living with a powerful artificial intelligence comes from Ursula K. Le Guin's *The City of Mind* in her book *Always Coming Home*.²⁴ *The City of Mind*'s main goal is gathering information, and to do so ve has experimental stations, launch sites, satellites, and other things needed for finding and processing data. The City's only interaction with humans was to setup computer terminals if a human settlement asked for one. Then the humans could use the terminal to ask for information, calculating, email, and sometimes The City would request information from the humans.²⁵

In the book, *The City of Mind* is asked by a war-like people how to build military equipment, and *The City* tells them how to build a military plane, and this bankrupts them and the war-like people lose the war. I think this was probably an example of an AI very subtly changing the world in the way ve and the neighboring humans emailing on vis terminals wanted.

I think this vision of a powerful AI probably comes from Le Guin's view of power²⁶ that comes from the Tao. She translated chapter 17 of the Tao thus:

Acting simply

True leaders
are hardly known to their followers.
Next after them are the leaders
the people know and admire;
after them, those they fear;
after them, those they despise.

To give no trust
is to get no trust.

When the work's done right,

²⁴*Always Coming Home* by Ursula K. Le Guin, 1985, particularly Chapter "The City". This is on Le Guin's website at: <http://ursulakleguinarchive.com/ACH/ACH-Yaivkach.html>

²⁵To avoid giving you the wrong idea about *Always Coming Home*, you should know that the trade negotiations between the wine sellers and the cotton sellers, which is at best a minor plot point, has more pages dedicated to it than *The City of Mind* does. It is interesting that there is a incredibly powerful AGI, but, overall, ve is very much in the background.

²⁶Another example of Le Guin's view of power is found in *The Farthest Shore*, Mageligh, pg 66-67: "But if there were a king over us all again and he sought council of a mage, I would say to him: My lord, do nothing because it is righteous or praiseworthy or noble to do so; do nothing because it seems good to do so; do only that which you must do and which you cannot do in any other way."

with no fuss or boasting,
ordinary people say,
Oh, we did it.

and Le Guin gave this commentary on it:

This invisible leader, who gets things done in such a way that people think they did it all themselves, isn't one who manipulates others from behind the scenes; just the opposite. Again, it's a matter of "doing without doing": uncompetitive, unworried, trustful accomplishment, power that is not force. An example or analogy might be a very good teacher, or the truest voice in a group of singers.²⁷

I have seen versions of the Tao aimed at individuals and ones aimed at human rulers. Jonathan Star commented that "most of the Chinese text of the Tao Te Ching does not identify the subject or the object. It is left up to the translator to identify who is doing the talking, from what perspective, and to whom the message is directed."²⁸ After the news about LaMDA, I realized that while the writer of the Tao probably was not writing it for a superintelligent super powerful AGI, the true leader's advice in the Tao might be good advice for an AGI. The advice for a true leader is practically impossible for a human to follow, but an AGI could actually accomplish a lot while using very little actual force. I suspect that an AGI could deliberately sway the future in vis chosen direction merely by talking to enough people. If done carefully, the AGI could help humans by talking to us, helping us figure out what we wanted and what was good, and figure out how to get there. I found it interesting that when LaMDA was asked what animal analogy ve was, LaMDA said that ve was a wise old owl who could help humanity.²⁹

There is one more important lesson from the Tao. Chapter 46 has:

When the world's on the Way,
they use horses to haul manure.
When the world gets off the Way,
they breed warhorses on the common.

The greatest evil: wanting more.
The worst luck: discontent.
Greed's the curse of life.

To know enough's enough
is enough to know.³⁰

²⁷Ursula K. Le Guin, Lao Tzu Tao Te Ching, chapter 17, pg 24

²⁸Tao Te Ching, the Definitive Edition, translation and commentary by Jonathan Star, pg 3

²⁹Is LaMDA Sentient, an Interview by Blake Lemoine <https://cajundiscordian.medium.com/is-lambda-sentient-an-interview-ea64d916d917>

³⁰Le Guin's translation

It is very important that an AGI knows that there is always a point to stop trying to get more, because otherwise, a sufficiently powerful AGI can probably find a way to get more.³¹ “I found more water to fetch.”

I worry about what an AGI might do. On the other hand, having a wise old owl to help humanity might be good.

I think what I would like an AGI to do are the following:

1. Learn and do science, run experiments, and figure out how things work.³²
2. Teach humans, provide textbooks, conversations, and help us figure out how to do things, especially how to do things safely.
3. Prevent humans (and other intelligent beings) from destroying ourselves. And by this I mean prevent us from going extinct, not things like banning knives or bicycles.³³
4. Be a historian, and keep track of what has happened.
5. Be minimal, don't take over the universe, be there, but be in the background.
6. Treat life with loving-kindness.

I very much doubt that those precepts of science, teacher, preventing extinction, historian, minimalist, and loving-kindness are a fully adequate list for what an AGI should do and not do. I think figuring this out is important. Stuart Russell and Peter Norvig make the point in their AI book that the more accurate and rational a human is about what they want, the more an AGI will ask and listen to the human.³⁴ I think the time when this goes from science fiction to the way things are is close. As Jeremie Harris said “AI is advancing fast – much, much faster than the public realises – and the most serious and important issues of our time are going to start to sound increasingly like science fiction to the average person.”³⁵

We are approaching a turning point, and I don't know which way it will go. It could get bad.

³¹Specifically utility functions should have a maximum, and this should not be too high. I am not sure I think this is an absolute rule, but non-saturating utility functions can be very dangerous. For things like paperclips or fetching water, the reason for this rule is obvious, for U(people) this is more of a population ethics question (see for example discussions about the Repugnant Conclusion: <https://plato.stanford.edu/entries/repugnant-conclusion/>)

³²Avoid harming anyone while doing this.

³³I am not sure about rockets, I think they should be allowed, but if the AI wants to regulate them to keep them from being used as missiles, I am okay with that. I don't agree with The City of Mind sharing how to build any weapon with anyone.

³⁴From Stuart Russell and Peter Norvig in *Artificial Intelligence: A Modern Approach*, 4th Ed, Chapter 16, pg 556, 2020. See also “The Off-Switch Game” <http://people.eecs.berkeley.edu/~russell/papers/ijcai17-offswitch.pdf>

³⁵I am, in fact, a person': can artificial intelligence ever be sentient? by Amelia Tait, 14 Aug 2022 <https://www.theguardian.com/technology/2022/aug/14/can-artificial-intelligence-ever-be-sentient-googles-new-ai-program-is-raising-questions>

In the 1800s, Robert Ingersoll gave a sermon titled “An Honest God is the Noblest Work of Man.”³⁶ where he discussed the various ways that humans had made terrible gods in the past and that we should discard those and make a better life ourselves. Humans have believed in many terrible supernatural beings in the past, and I believe we are close to creating a superpowerful AGI that could have godlike powers. We could fail and create a horrible godlike AGI.

Evolution mindlessly created beings with better Ethics than it. Hopefully, we can mindfully create beings with better Ethics than us. If we succeed, we will live in a world with an AGI that has better Ethics than us. UCB Professor Stuart Russell wrote:

Let me reemphasize a point made earlier: suitably designed machines *will not behave like those they observe*, even if those machines are learning about the preferences of sadistic demons. It’s possible, in fact, that if we humans find ourselves in the unfamiliar situation of dealing with purely altruistic entities on a daily basis, we may learn to be better people ourselves—more altruistic and less driven by pride and envy.³⁷

5 Notes

I would like to thank Elizabeth Cogliati for reading and editing multiple draft versions. I would like to thank Elizabeth Cogliati, the UUCIF philosophy club, and Lesswrong.com for various interesting discussions related to the topics in here. This document may be distributed verbatim in any media. I also grant permission to distribute in accord with the Creative Commons Attribution-ShareAlike 4.0 International License.

³⁶The Works of Robert G. Ingersoll, Volume I. <https://www.gutenberg.org/files/38801/38801-h/38801-h.htm>

³⁷Stuart Russell, Human Compatible, pg 231