Article 3: Are four dimensions enough to fit a moral understanding?

Character cannot be developed in ease and quite. Only through experience of trial and suffering can the soul be strengthen, ambitions inspired, and success achieved- Helen Keller

Logic will get you from A to B. Imagination will take you everywhere- Albert Einstein

Does one believe “good” in all its infinite meaning can be reduced to a series of binary 1s and 0s? If yes, then super aligning AI is simply an issue of programming and finding the right formula to program into its circuits. If one answers the above question in the negative then what recourse do we have to determine if AI is acting with “good” intent such that we can have reasonable trust in the AI?

In this issue of trusting AI one might ask how do we humans trust each other? Is it only under threat of legal action and criminal punishment that baristas at Starbucks do not mess with people’s drinks? I would answer no to that question. Instead, I would argue that as people go about their day and work they choose to not harm people for the simple reason they believe it would be the wrong thing to do. It is the internal moral safe guards that most people have that results in people being able to get along, trust each other, and not harm each other.

My answer to the trust question is not to dismiss the necessity of criminal law. Nor I am dismissing how helpful it is to provide a legal recourse for people when they are unlawfully harmed. To express my argument in more theological terms, I believe most people go to church to get to Heaven not to avoid Hell. While Hell can still provide an important motivator in the same way penalties for violators of criminal law do, especially for people who do not have a strong internal moral compass, it is not the primary reason people do the right thing.

I believe the best way to understand why people are attracted to doing the right thing is in assuming that man has a connection to the divine. Whether this means Man and God or Man and some other concept of an inherent divine order, I am making a fundamental, generic argument that should not change based on what religious system one chooses. I have heard this assumption of man looking for the divine get referred to by psychologists as a religious instinct.

It is an empirical fact religion has existed with humans for millennia, and the base arguments made by those religions is that the moral heads of those

religions, i.e. the Gods, God, or divine energy, pre-existed our times and exist outside the limits of our known reality. The main point here is that even in an atheistic point of view, people have been grasping towards the divine, whatever that means, for longer than most technological inventions. Even if one rejects the theological arguments of those ancient people as “illusory” born from a lack of understanding of science, there is still the issue of explaining what this religious instinct is. The main point I take away from man’s long study and search for the divine is that, given the panoply of evidence that exists to support this inherent feature of man, between different cultures at vastly different times, there is this common trait within people where they fundamentally yearn to do the right thing. There seems to be an innate desire for one’s actions to be in coherence with the metaphysical plan as described by the religion or other stand-in for an eternal order that one follows. In more fanciful terms, it is that the movement of the cosmos, stars, and man is in a coherent, beautiful, dance and people year to dance with it and partake in that cosmic beauty. Sometimes poetic language is the more precise description of an idea.

The fact that most people feel and act on this yearning interests me deeply, both as a student of neuroscience and a very curious young boy. The very basic supposition of neuroscience is that people need neurons to process information and behave. This fact is well backed by decades of research and experiments. Most often this is demonstrated by showing that if one takes away neurons from a person they lose the ability to process information or behave. In fact, the brain is so well organized that many neurons end up specializing for specific tasks. Damage in specific areas of the brain, like Broca’s area, would mainly cause issues in speech articulation. There is also the famous example of Phineas Gage who was a well-functioning railroad foreman until one day a railroad spike went through his head, destroying much of his frontal left lobe. What once was a person who could take on the responsibilities of a foreman after the accident became an erratic, emotional and otherwise unreliable, person. Another neuroscience discovery was the existence of “grandmother” cells. An experiment was able to show that only a specific neuron would fire in response to a person seeing a picture of their grandmother- hence the “grandmother” cell. Researchers were able to repeat this experiment to include pictures of celebrities as well. These examples and experiments can be crudely but very accurately summed up that if one takes away a person’s brain there is not much person left, hence why bullets work.

One unanswered neuroscience question is which neurons/neural systems code or are responsible for a strong internal moral compass, alternatively, of the absence of one. The so-called sociopath is viewed as the minority and the focus of the endeavor. If the neurons that encode an internal moral compass are off or malfunctioning then by fixing those neurons we can restore/give a person the use of the common or “normal” internal moral compass. In the context of AI and the concept of a “safe” AI (at least as safe as normal humans), this understanding would extend to how to build or replicate a strong internal moral compass inside of AI.

This basic neuroscience question- about the nature of the human moral compass and those “moral neurons” that are the source of sound moral judgements, has some very weird physical questions attached to it. How do these neurons code to be the morally good ones? Do moral neurons have special proteins that enable them to code to morality? Or is the brain just a very good repository of information? In other words, if a person is taught a moral system by their parents, then are they just reproducing what they are taught? If a person’s sense of morality just comes from what they are taught then this would point in favor of the argument that AI just needs to be trained correctly to become super-aligned with human morality.

I find this teaching or training argument to be a somewhat correct, as many parents have experienced that one does need to teach their small, young, impressionable children that it is wrong to steal or hit or bite. Furthermore, the fact that minors are treated differently under criminal law than adults points in this direction as well. I have personally had friendships with people who were nominally very religious, going to church each week, but then suffered a tragedy and lost faith. Therefore, it seems that they were taught a set of rules and they were faithfully following them until they found some contradiction in them and thus stopped believing. In effect, they followed the rules because they were taught to not because they fundamentally believed in them.

From an anti-religious point of view one can argue that these notions of faith and morals is just an advanced mechanism to achieve cooperation between humans and thus maximize people’s chances of survival. Life is simply optimized to solve the objective function of reproduction and survival. From that point of view, religion can just be understood as an advanced evolutionary tactic. This argument being true would then mean that, for training, AI one would just need the correct objective function and training algorithm to then achieve the same

“advanced cooperation mechanism” humans developed in religions. Fundamental understanding would not be required. While I will later address in more detail the physical, mathematical flaws of Darwinian evolution in positing that “survival of the fittest” is or was the cause of biological evolution, note the intellectual sleight of hand in this anti-religious perspective of simply replacing a God or a pantheon of Gods with a cleverly optimized objective function.

From a religious perspective, let us use as an example of the “objective function of life” the Christian core goal to live a “good” life as dictated by the moral system established by God, the supreme deity and source of morality. Furthermore, in Christian lore, many of their fundamental rules come from prophets who had special, fundamental connections to God that enabled them to well describe the correct moral code. In addition, most Christian sects believe that man is created in the image of God and thus there is an intrinsic divine connection between each man and God. This then enables the fundamental understanding of the moral rules of the universe although not always without some effort- sometimes called prayer or meditation. I should mention that in this argument people are meant not to simply produce or mimic the correct behavior; rather they are meant to produce the correct behavior because they believe it is the right thing to do. In this framework, a person’s moral compass is supposed to point towards God. I should mention that in this Christian world view people are having to deal with the concept of both a loving and alien God- alien in the sense that God is such a vast supreme, all-knowing being that God is clearly not human. Therefore, there will be situations where God, this vast unknowable being, will test people with (or in some views teach people through) moral conundrums for which they cannot have prior knowledge on how to answer them. The only way to know the correct way forward is to have the internal compass morally aligned as to provide the moral answer- even if only by asking God. This system would argue that true moral behavior does not just come from teaching people to simply reproduce what they were taught. In some Christian sects they would appeal to this divine messenger called the Holy Ghost that provides access to and communicates this knowledge.

Before moving on, please note the difference between the perspectives. The anti-religious point of view effectively advocates for using this “survival metric” as being superior to the “closer to God” metric advocated by Christians or other theistic religions. Furthermore, the anti-religious point of view will then have to appeal to many physical ideals as fundamental like entropy, magnetism,

conservation of energy and momentum etc., to fully explain physical reality. In effect, the anti-religious view assumes mathematical ideals are what fundamentally construct the universe and thus humanity simply lives, and dominates, this planet due to its evolved mastery of mathematics. In many ways, this just replaces the divine with math. In contrast, the Christian worldview does not require such a long-written justification and neither does it exclude mathematical or physics based arguments. In its more fundamental logic, it is just a lot simpler (at least to a human, I am not sure an AI would find it a simple argument at all), and there is a special intellectual elegance in simplicity. Since God is the most fundamental basis for order simply being closer to that most fundamental thing is good. Although many intellectuals might sneer at such reasoning as overly reductive to simple state that reality is the way it is simply because God wills it. Deus Vult as the crusaders would say. It is important to recognize that this “simiplication” is not one which should be understood to as a justification to ignore the intellectual complexity inherent in reality. The Christian argument is actual the reverse. The power of God is supreme such that anything is possible. Using complex, esoteric mathematical arguments to describe the universe is not wrong, but they are not end of truth in and of themselves. God is more complex than even the most complex mathematics! They are true because of Deus Vult, God wills it. In opposition, the anti-religious perspective states mathematics as truth onto itself and man’s connection to math as that connection of divinity and that is it!

As a quick aside, I can think of no more cruel classism than arguing that people who can do math better are inherently morally better than other people. In the anti-religious perspective, since math is effectively what is divine those who have stronger connections to the divine i.e. math, are then necessarily better than others who are bad at math. I believe it is a strong mark against the anti-religious viewpoint that there does exists not a widely accepted viewpoint that those with strong math skills necessarily have strong internal moral compasses.

It might seem like am standing way out in left field to be discussing God and a religious perspective as a core question in identifying how to train AI alignment, but the reason I included the religious perspective versus the trained or educated perspective is that they are, in fact, pursuing two very different physics argument. In the trained perspective one can take the information that exists in reality, encode it into a physical system, and then get desired outcome like in a Skinner’s Box. For refence, Skinner’s Box was a contraption develop by behavior

psychologist B.F. Skinner to train animals to simply replicate the desired behavior. Whether the animal had any fundamental understanding of what was going on was not required. Extending this idea to people and how they are trained would refer to the sound waves heard from sermons and the light emitted by ink on a holy text which then transmit the information to the neurons in a person's brain to then encode morality. In the Christian example, which does not exclude the trained perspective as much as it would call the trained perspective as flawed, the fundamental understanding of morality does not come from just the sound waves of sermons or the light from biblical ink, it ultimately comes from God himself, often defined by words like inspiration or enlightenment. In the Christian perspective neurons in the brain could spontaneously encode fundamental senses of morality that come from a God who is not constrained by our physical reality. Since God is supreme one cannot then put the laws of physics above him and demand he obey them. Therefore, God, as understood by most people, cannot be simply bound into 4D dimensional spacetime. Essentially, I would argue that the Christian argument, as well as any theist argument, aligns with the idea of there being information which lies outside this realm of reality which people can in fact access due to their "divine" connection. From that perspective, faith is the ultimate teacher and that it provides a connection connects to a higher dimensional space than just the simple, easy to see 4D universe we live in.

As an important but relevant aside, I have watched a video of a physicist I quite like stating that there are no, in principle, differences between the brain and a computer. At the end of the day, it is all just computation. I find this perspective unimaginative and disturbing. Ask yourself, reader, is your feeling of joy, beauty, sorrow, and grief just a complex computation? Would you surrender your love for your brothers, sisters, parents, and kids to being just a complex computation? For the parents out there, is the love one feels for your children, and the care parents put into their children, simply come from an evolutionary computation compelling them, and because society has taught it was the right thing to do? Or it is more profound and special? If one feels that it comes from outside oneself, something new that did not exist there before, then I would argue that is in line with the extradimensional, religious understanding rather than the anti-religious trained understanding.

The difference between this argument also highlight some of the very weird cognitive abilities people have. For example, how is it logically possible for a

person to think of a logically impossible situation? If the trained perspective was right then a person should never be able to think outside of the box as defined by what they were taught. How is it that people can go through psychosis and demonstrably see things which are not there? How then are these experiences and plans executed in neuron if training is required in all cases? What is going wrong or right in neurons or more perhaps more generally, the physical essence of the brain when, in fact, the brain is not “normal”?

While I believe that there simply does not exist the information within our physical reality to explain the empirical phenomena of human behavior and thus additional “dimensions” are needed to provide a coherent explanation for the brain and consciousness people work, this is in no way to deny the basic dimensionality of the universe being a 4-D manifold as described by Einstein’s General Relativity framework. The question then becomes figuring out what rules govern how these extra-dimensional spaces intersect or compact down to then cohere with our 4-D reality. If one can answer these questions then I believe will be able to make safe AI for a certainty. Hopefully now the reader understands the title.

What I fear is going and on will continue to go until we have a good physical answer for how people are able to have this intrinsic sense of good is that current AI will only develop along the lines of the trained argument. If the religious view is correct then it can be understood that what training does is approximate the fundamental meaning good rather than replicating it. In other words, training can provide a synthetic, approximate understanding of good. Therefore, the issue is using this synthetic, fundamentally incomplete understanding of good to guide behavior rather than the real deal. For AI development this means that even if one trains the AI eternally on a theorical infinite amount of data it will always only develop a synthetic, approximate understanding. While in practice this might very well be fine what I fear is the emergence of catastrophic errors.

As an example, and to show that while I have been arguing about teaching good in this article, my fundamental argument extends to solving any task. Let us consider an AI that is meant to bake bread. Let us assume that the AI has been trained on thousands of hours of video watching bakers bake. Then the AI is put in charge of the bakery. The AI works wonderfully, at all hours of the day, and is able to effectively bake more bread than a human baker and the quality of the bread is the same as well. One day, however, the oven overheats due to a

mechanical error and the AI simply continues to bake bread without stopping. Even though the AI learned a sense quality control in throwing out bad loaves in none of the videos it was trained on did the bakers every deal with a fire emergency. Thus, the AI literally has no idea what to do nor an idea why every loaf it makes is severely burnt. The AI just continues to bake wasting energy and ingredients. Eventually the catastrophe cumulates into the bakery burning down.

Now one can argue one just needs to include data the includes bakers turning off an overheating oven. However, that is one simple novel circumstance. What is the next novel circumstance that can occur that the AI will need to be trained on so it can “understand” what to do? Furthermore, how much training data does one need to do deal with fire emergencies? With a five minutes or five hours of video training be required to get the AI to react properly? Might as well do the five hours especially if that limits liability concerns. If 5 hours of video training does not exist then then one in effect make five hours of data by hiring people to create the necessary training data.

Let us assume that after training on 5 hours of data the AI the fail was able to prevent the entire bakery from burning but couldn’t save the oven so less of a catastrophe. However, that is not good enough, the oven needs to be preserved as well. Then the AI designer find out that AI was only able to deal with an overheating oven effectively after 500 hours of training video. This issue at this point is the for designers they had already ran over their time and investment budget to build full autonomous baking AI.

There are so many nuances in training and testing AI that ALL need to be answered well for AI not to be fundamentally limited and flawed. Furthermore, I believe the intrinsic understanding is necessary for the AI to adapt and react to novel situations. While from some tasks that some animals like dogs can do competently then AI just needs to have a dog’s intrinsic understanding. I certainly believe it is possible to create and deploy AI with this synthetic understanding it could work for 99.9999% of situations. However, my fear is the 1 novel situation will produce profound a catastrophe which will affect all of society even if AI is regulated to discrete tasks that do not threat human safety. I have a fear that an AI which has been endless trained on infinites amount of data then the novel event it cannot deal with will, by definition, be catastrophic. Only through the intrinsic understanding can these catastrophes be avoided. This is an important reason I am writing my articles.