There are many pressing concerns in AI safety. None of them have easy answers (if they did, well we wouldn’t be concerned would we?). Some are more likely than others. Some are more dangerous than others. This article will explore a laundry list of these concerns. I won’t suggest many solutions, and if I do, none of these solutions will be by any means complete. For the most part, it is my opinion that to solve these dilemmas we do more research, and in particular theoretical and rigorously mathematical research. Here we go!

One of the most commonly discussed dangers is that of existential risk. Existential risk in this context is the vague theory that an artificial general intelligence (AGI) could become powerful enough and misguided enough to result in human extinction. There are two necessary steps for this type of threat to become reality: (1) the AGI must have values that stand opposed to the human race, and (2) the AGI must have the power and capacity to act on these values, overcoming any human intervention. The first condition (1) seems plausible. It may be difficult to instill prescribed values into AGI systems (though giving objectives/values seems trivial), or more likely, malicious actors may choose to intentionally gives values to AGI systems that are not in line with the interest of society at large. The second condition (2) seems less likely to be realized, though also plausible. One particular concern here is that of *recursive intelligence growth*. Consider an AGI agent that is able to control its own intelligence or create new intelligent agents. If we humans are able to build an agent more intelligent than ourselves, than it would seem logical or at least possible that this agent would be able to build another agent even more intelligent than itself. A chain of intelligent agents, each a little bit more powerful than its predecessor could be created in this recursive manner. If these incredibly intelligent agents also had values not aligned with ours, and the ability to act in the world without constraint, there may be real existential risk. And it’s worth noting that even if these outcomes are unlikely, as long as there is any reasonable chance that they could manifest themselves, we must take them seriously (since the outcome could be so *infinitely* bad).

Besides existential risk, there are many more immediate challenges and threats posed by artificial intelligence including deep fakes, autonomous weapons, social manipulation, systematic discrimination, and AI transparency. Not all of these challenges will be discussed here. The first one of these that I do think is worth discussing is AI transparency. There are at least two distinct transparency problems that AI faces. As Mind AI puts it: “The first deals with public perception and understanding of how AI works, and the second has to do with how much developers actually understand about their own AI” [1]. In regards to the first issue, of course the public will not entirely have the time or means to learn how AI works, and it should certainly not be their responsibility to do so.

This leaves a real dilemma. AI is going to make some of the most significant societal decisions on topics ranging from legal responsibility to medical diagnoses to running yellow lights. If the general public is unable to understand how these decisions are made then how can they be expected to trust them? Especially when these systems might make decisions that seem incorrect, discriminatory, or have immediate negative consequences, it seems almost impossible to trust them if we cannot understand why they are making the decisions that they do. This problem of public understanding and perception could be solved to some extent if AI developers were at least able to explain the fundamental question of why AI agents make the decisions that they do.

The state of the art in AI right now is ‘deep learning’ which consists of massive neural networks that are much the opposite of transparent. It seems like the best way to understand what causes a neural network to make a decision is to try varying the input and see how the output varies. This is really the last resort to understand any sort of black box algorithm (though its not a black box algorithm at all, really just some multivariate calculus which is the most confusing part; how could we not understand something we have defined so precisely?) I think there are many different approaches to overcoming this transparency issue. The most easy one, and I think perhaps the best, is just to get better at living with uncertainty. Another possible solution is to abandon neural nets (fat chance, they’re the best we got) for some other more transparent machine learning algorithms. I do also think there is some potential for really just getting better at understanding what the neural network is doing. What this might look like technically is viewing neural nets as doing causal inference (like in a probabilistic graphical model) which is a much more transparent process whereby you can directly identify which particular things caused which other particular things and to what extent.

The next threat that I will discuss is that of deep fakes. Already when your browsing through Facebook or YouTube or a friend sends you an image or video with something truly remarkable, there is a sort of skepticism that we get (or at least I do). Maybe we’re not always sure *how* a video or photo could be faked so precisely, but none the less we think that it might have been. This type concern is wholly reasonable, and in fact it is artificial intelligence (in particular generative adversarial networks (GANs) that are starting to be the culprits for these incredibly ‘real’ fakes. This can quite quickly become dangerous (and already is). Consider the court room. What if you can’t trust body camera footage or crime scene photography because its just to easy to create fake footage with AI systems. But you don’t even have to go to the court room, or even outside your house to be worried about this type of deep fake. Because if the hype around fake news wasn’t enough already, imagine all the fake scandals and he said she said fake videos. (Wait is this fake news too‼?! Nah don’t worry about it, this is just some random blog post, I don’t have the time or effort to come up with fake stuff.)

Moving on to another buzz word topic: discrimination (as if this whole post wasn’t just buzz word galore already). One of the primary tasks of AI systems right now is classification. AI is very good at answering questions like “what type of animal is this” or “how much money will this house sell for”, or “what is the probability that this person on trial is guilty” or “which candidate should I hire to most improve my business” which all fall under either continuous or discrete classification problems. The way these AI systems come up with their answers is by first training on example cases. Now one problem is that if the examples that are being used for training are already discriminatory (say that in the data used for training black men on trial were found guilty 30% more than white men) then the AI system is probably going to predict similar statistics going forward. Even if race is not explicitly a model of the system, other things will be such as the persons zipcode, income, or employment which correlate with race. We could just mandate that equal number of colored and non-colored, female and male, and so on all get equally proportioned outcomes in the end. But that seems clearly problematic as well. What if we’ve hired 550 white guys and 530 black guys this year. Does that mean the next white guy we interview we have to reject? What if were really sure he would be a great asset for our business, well then maybe we hire him too, but then were at 551 white guys and 530 black guys, so the next white guy we really have to reject. This issue I don’t think is one particularly unique to AI systems, and the solution here is really to decide as a society what our preferences and values are.

Let’s go back to a less dangerous topic (or maybe more dangerous, it depends on your perspective I suppose). Another threat AI poses which goes a bit along the line of existential risk is that of social and market manipulation. It seems plausible that an artificial intelligence which is so incredibly beyond what we could imagine, probably won’t have super strength, or massive military force at its disposal, and might not be able to hack into offline, triply redundant systems like U.S. nuclear launch sites. What does seem plausible is that with the super intelligence the AI does have, it will be able to foresee events that no one else can. After all that’s what AI is being designed for – prediction. With these incredible predictive powers, and some capacity to interact with the world, the AI could set up chains of events with precise calculation to manipulate complex systems such as financial institutions, international relations, or even biological systems. Of course going back to the first concern on AGI, condition (1) must be met already, where the values of the agent are unaligned with out values for this type capacity of predictive manipulation to be a concern instead of sublime endowment.

There are many more threats and challenges of AI that were mentioned or simply overlooked (hey I’m not the superintelligence here). As I said at the beginning, it is my belief that to really understand these threats, how realistic they are, and how dangerous they are, what we need is formal systems, rigorous mathematics, and careful analysis. So I better get back to it!