Free Will and Machines

Kevin Li

May 14, 2021

As intelligent machines become increasingly capable, the scientists who design them are also faced with an increasing responsibility to make sure these machines do not pose threats to humans. In Human Compatible, Stuart Russell offers a set of guiding principles towards this task, in order to frame how we approach large-scale AI systems. The first of his principles is that an intelligent machine's only objective should be to maximize the realization of human preferences. The first part of this paper will attempt to clarify what he means (or could mean) by "human preferences," in the context of Harry Frankfurt's idea of first and second order desires. On the basis of this clarification, I will argue that so long as machines use human behavior to understand human preferences, they have no capacity to understand second or higher order desires. The second part of the paper considers the consequences of such a conclusion, where machines are given the objective to maximize human preferences, without any consideration for their second or higher order desires. I argue that machines of this design are at risk of violating free will, in the way that Frankfurt uses the term. Lastly, this paper will consider Russell's notions of uncertainty and maximization, in an attempt to revise his principles (or at least clarify how they should be interpreted) to avoid the problem of violating free will.

To make sense of Russell's idea of "maximizing human preferences," I will first outline Frankfurt's distinction between different kinds of desire. When someone has a desire towards an action, they are said to have a first order desire. A second order desire, in contrast, is a desire to have a particular first order desire. For example, I might have a first order desire to eat spicy food, in which case I truly want to eat spicy food. Let's say, for sake of contrast, that I can't handle spicy food, in which case I don't want to eat spicy food (or don't have the first order desire to do so). If my friends are all going out to eat hot wings together, and I want to feel included, then I'd have a second-order desire for spicy food—I don't want to eat it, but I want to want it. I have a desire to have a first order desire.

Another distinction that Frankfurt makes is between a desire and an *effective* desire. When a desire compels someone to act, it becomes an effective desire. This effective desire, in turn, expresses that person's will. Whether someone wants a particular desire to be effective or not is a question of their second order volition, a term that Frankfurt uses to mean a desire for some first-order desire to be effective. For instance, consider my two first order desires, which are to go to the gym (because I want to be strong) and to sit on the couch (because I am also lazy). If I choose to sit on the couch, it becomes my effective desire, since it is the desire that moves me to act. My second-order volition, however, might be to go to the gym. It isn't my

effective desire, but I want it to be—it's the desire that I want to act on.

With these terms now clarified, we might ask how Russell's idea of human preferences fit into the picture. He argues that he means to use the term in the broadest sense possible, to include "everything you might care about, arbitrarily far into the future" (173). If this is true, it seems that preferences would include all different kinds of desires. They seem to include both effective and ineffective desires, since the term "everything you might care about" will include the desires that you don't act on. They also seem to include second order desires and volitions, since the term "arbitrarily far into the future" implies that preferences may not just be your immediate desires, but desires you wish to have at some other point.

I argue that Russell's idea of what counts as a human preference (from the perspective of a machine) must be smaller in scope. Take Russell's third principle: "The ultimate source of information about human preferences is human behavior" (176). If machines should ultimately learn a human's preferences through their behavior, then they can only learn preferences through effective desires. In other words, human behavior only captures effective desires, and so the desires we have but do not act on will not reflect in our behavior, and therefore won't be accounted for by machines. Consider a machine who is learning about a drug addict's preferences. If they observe the addict's behavior (which is repeatedly taking the drug), they are likely to conclude that the addict has a desire to take the drug, and that it is the preference that should be satisfied. The addict may have a second order volition to not take the drug, but a machine who ultimately decides on a human's preference through their behavior can only conclude that the addict wants to take the drug. So long as we accept Russell's third principle, the intelligent machine has no possibility of learning desires that have not been acted on. His notion of "human preferences," in this context, can only mean effective desires.

One way a machine might be thought of as understanding second order desire, though, is if it were to *ask* a person what their second order desires are. In other words, rather than prompting you to act on your first order desires, a machine might prompt you instead about which of your desires you want it to satisfy. In this case, it seems like a computer is still only observing behavior, but the behavior being observed is an indication about a second order volition. An argument like this could be resolved in two ways. The first is to just claim that any behavioral alignment you have to a second order volition makes it an effective first order desire by virtue of acting on it, even if it is only to indicate a preference for it. In Frankfurt's terms, one could argue that any second order volition becomes a first order desire the moment in which it

affects your behavior—it is only a second order desire until it has compelled you to act in some way. A second way to respond is that even if a machine can understand some of your second order desires, machines can never understand *all* of your desires—there will always be desires you haven't acted on (or effective desires you act on but wish not to,) that a machine can't prompt, and so the question that remains about what they will do in that scenario is still consequential and worth contemplation.

To revisit Russell's first principle for beneficial machines, I ask: what does it look like for a machine to maximize the realization of human preferences, if the machine's idea of human preferences is only their effective desires? In other words, what would happen if a machine tried to realize a human's preferences as manifested in their behavior, with no regard to their second order volitions? The machines will maximize a preference which is based on the desires that have been acted on in the past, because, for Russell, that is the ultimate source of knowledge about preference. The problem lies in the case where a human's second order volitions, or what they want to be the desire they act on, is not in line with their previous desires. Take the addict case from above. A machine whose job is to either give him drugs or not give him drugs will give him drugs, because that is the preference that has been learned from behavior. The addict has a strong second order volition to not take the drug, but the machine doesn't know this, or can't deduce it from his behavior. As a result, the machine is realizing the wrong preference. Something even worse happens when the machine is tasked not only with realizing the preference, but maximizing this realization— if the machine has confidence that the addict likes the drug, it figures out that the most efficient way to realize this preference is to directly feed it into his body, which is faster than having him take the drug himself. To explain this another way, the addict has two choices: to take the drug or to not take the drug. The intelligent machine is confident that, because the addict has always taken the drug, their preference is to take the drug. Since they are tasked with *maximizing* this preference, it becomes inefficient for the machine to even offer the alternative at all. If the machine is a good machine and does its job as Russell wants it to do, it must factor in the cost (in time or in resources) in giving the addict the choice to not take the drug. What's at stake here is a violation of Frankfurt's freedom of will. His idea of freedom of will is "the satisfaction of certain desires—desires of the second or of higher orders—whereas its absence means their frustration [...] those suffered by a person of whom it may be said that he is estranged from himself, or that he finds himself a helpless or a passive bystander to the forces that move him" (17). Freedom of will is the freedom for a person's second order volitions to become

their effective desires. In other words, if a person wants to be compelled to act by a certain desire, they can do so. In our situation here, a person may want to act on a desire (a second order volition A), but if that desire is different than what they have previously acted upon (an effective desire B), it is possible that a machine will prevent them from acting on A in an attempt to maximize the realization of B, all still within the limitations of Russell's principles. To make this problem clearer, I'll offer another example. Consider a mobile app that provides a curated news feed to a user. The preference being maximized, in this case, might be a user's preference for either liberal or conservative-leaning articles. Say that a user initially has a desire to read liberal news and acts on this desire. They click on mostly New York Times articles, and spend less time looking at articles from Fox News. The app, which has been tracking the user's behavior, picks up on the fact that they prefer reading liberal news to conservative news. Since the app is tasked with the job of maximizing the realization of preferences, the user's feed is now only populated with articles from sites like the New York Times and The Guardian. Conservative articles will only distract from the articles that the user actually reads, so the app decides that the best way to *maximize* preferences is by taking away any articles from Breitbart and Fox. Say that the user has recently heard about the dangers of media bias, and they form a second order volition towards reading more conservative counterpoints— it is a desire that they have not acted on yet but wish to. The app doesn't realize that the user's preferences have changed and continues to exclude conservative news sources from the users' feed. The user wants to read from Fox News, but the app has made it impossible for them to do so (because the app thinks they want to read from somewhere else), violating their freedom of will.

Here's one way the situation could easily have gone differently. Let's say that the app discovered the user's preference for liberal leaning news, but instead of completely removing conservative news from their feed, the app simply reduced the frequency in which they popped up. A Fox News article would still show up on the user's feed every couple of weeks. When the user forms the second order volition to read conservative news, they begin to act on it—at first quite infrequently, because the conservative articles don't show up often, but sooner or later, the machine picks up on the fact that the preference has changed, and starts to curate a more diverse news feed for the user. The difference between this situation and the original one is that the intelligent machine did not act as if they were certain that the user only read liberal news (because if they were sure, the best way to realize the preference would have been to get rid of the conservative news articles all together). Being

unsure about the user's preferences allowed the machine to adapt when the user's behavior started changing (as they started acting on different desires). As a result, the app made it possible for the user to act on their second order volition, respecting their freedom of will. This is what Russell's second principle seems to account for, which is that "the machine is initially uncertain about what human preferences are" (175). The clarification I'd like to make here, though, is that initial uncertainty is not enough— if humans can develop second order volitions at any point in time, then it seems that machines should always be uncertain as to what human preferences are. In other words, humans have the capacity to change what desires they act on, which means that their behaviors change, and therefore their machine-read preferences will change. If a machine is ever too certain about the preference, they risk the possibility of making it too difficult (or almost impossible) for a human to act on a different set of preferences. In order for a machine under Russell's principles to respect free will, it seems that it should always have some level of uncertainty about human preferences.

But how uncertain should these machines be? If too much certainty runs the risk of violating free will, too little certainty would make it impossible for the machine to decide on a preference to realize. To put it differently, there is a trade-off between Russell's second principle (or as I've proposed to revise it,) which is that the machine should always be uncertain about what a human's preferences are, and his first principle, which is that the machine should maximize the realization of preferences. For example, if a news app were always uncertain about which news articles a user prefers, it wouldn't be able to do its job, which is to maximally display whichever news articles the user prefers. On the other hand, if the app were too certain, we run into the earlier problem, that the user has no ability to change which desires they act on. If the machine needs to be uncertain but also act maximally, and there are situations in which they trade off, then Russell's principles need to answer for how to balance the two.

One way to resolve this problem is to prioritize maximization but offer some reasonable threshold for how much uncertainty a machine should always have. In other words, the machine should always maintain some baseline level of uncertainty as a necessary condition for it to operate, but the machine is free to maximize any preferences up to that level of certainty. For instance, we might require that the news app is always 5% uncertain about a user's preference. If the app believes that the user's preference is for liberal news, then the app should treat their preference as being 95% liberal news, and 5% for everything else. In more concrete terms, this might mean that the app guarantees that at least 1 in every 20 articles shown is uninfluenced by

the user's preferences (or at least what the machine believes to be their preferences). This prevents machines from violating free will because users are guaranteed some minimum amount of leeway to act outside their previously effective desires. In some sense, the guaranteed uncertainty offers a constant window of opportunity for a user to change which desires they act on, according to their second order volitions.

A question that remains in such a proposal is what that threshold should be, and how it can realistically guarantee people the opportunity to act on their second order volitions. An answer to this is hard to settle in broad theoretical terms, since to say "realistic opportunity" is to ask about the context of what the machine is doing, and what it is to be reasonably uncertain. For this reason, we might leave it to the people who build the machines to be more specific about what a threshold should look like. What we've accomplished here is not a claim that a certain threshold will solve our problem, but rather than the lack of one altogether is out of the question— that uncertainty *must* play some part to protect free will.

Another way to approach this tradeoff is to view uncertainty as itself a factor that can be optimized. If we think of uncertainty as the opportunity for a person to act on a change in preference, the ideal machine should actually allow for this to happen, such that it can better realize preferences. Consider two possible machines whose only goal is to maximize the realization of preferences. The first machine figures out what a person's preferences are, and then maximizes those preferences. If people's preferences change, our complaint would be that the machine violates their freedom of will (based on my argument earlier in the paper). However, the complaint could also just be that the machine isn't actually maximizing the realization of preferences, since it fails to account for and realize *new* preferences. A second and even more ideal machine would figure out exactly how often a person's effective desires change and introduce the exact right amount of uncertainty at the right times, so that it can correctly update what it thinks a person's preferences are and therefore continue to maximally realize them as time goes on. In this way, uncertainty is not a concern separate to the goal of maximizing preferences but is instead just part of what a machine takes into account in order to more effectively do so. This solves the trade-off problem not by changing Russell's principles, but by lending a more specific way of interpreting how a machine can go about "maximizing". If maximizing the realization of preferences already entails that machines use uncertainty to adapt to our changing second-order volitions, there might not be a problem in the first place.

To summarize— I have offered a way to think about what Russell means by "human preferences" through Frankfurt's notion of first and second order desires. Given

Russell's principle that human preferences arise from human behavior, I've argued that machines, under this picture, have no possibility of knowing a human's second order volitions, which in turn means that machines who purely maximize a human's past effective desires is at risk of violating their freedom to act on other desires in the future. The latter half of the paper looked at how Russell's notion of uncertainty is important towards maintaining freedom of will, and concluded with two proposals of how, given its importance, uncertainty can be reconciled with the ultimate goal of intelligent machines, which is still to maximally realize human preferences.

References

[1] Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy*, vol. 68, no. 1, 1971, pp. 5–20.

[2] Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control.* Penguin, 2019.