Reflections on “Post Modernism: What AI Alignment Teaches Us About Education”

After reading about AI alignments, I deeply feel that an overly clear and direct measurement for defining "success" actually decreases the possibility for things to go in their original way. Proxies will never take into account all the factors. This reminds me of what I wrote in that reflection before, that never think of a quantitative, or even an approachable categorical standard, would reflect the entire thing. As stated in "Atomic Habits," achieving the purpose is actually better for approaching the ideal identity rather than the goal itself. Although those two concepts are not the same, it reminds me of the idea that it is always better, as a human, to improve ourselves with a long-term identity agreement instead of an overly clear goal that is supposed to ensure success (which can reflect the whole idea of success). But it is really important to consider time due to the "time consistency" problem, where the long-term achievement of a goal might be overly optimistically predicted.

But what we can do to separate ourselves from AI, as I feel, aside from doing what AI can never do, like innovation, is to use our intrinsic emotional value effectively. Since we are aware that the measurement of most things could never be perfect, we'd better go back to "why"—why the measurement exists. They want things to be controllable, like education, the level 10 piano certificate, the exam score, which are designed for checking knowledge but do not ensure knowledge. Over time, the credibility of these measures increases, making them trustworthy and being incorrectly sought as "success." Extrinsic motivation takes the same approach; it is designed to use certain things to affect what we care about, instead of making us care about certain things. These deviated measurements also make motivation through extrinsic events even harder to control, such as those that are imposed.

When talking about AI, what I learned from this passage is actually the application, or the conduct of how things (like AI) are designed. Some people, especially me, will always overestimate those who are famous or credible in their fields and incorrectly assume that they are beyond what humans are. For me, before, I never thought about how AI was created, imagining it as a magic that has a human mind and thought. However, as I calm down and think about it again today, the creators are all human as well; they must start in a way that people can understand and impose. In that way, what AI does is like the selection method, which is similar to evolutionary selection, as people always learn from nature and "history."

The final prediction and improvement ideas make me aware that the ways to make AI better could also benefit us; it is a two-way relationship. It might be possible to try shifting measurements during AI training. We, in human judgment, aside from subjective comments, actually need objective measurements to reduce some of the repeated work that can be done by standardized questions (though continuous innovation is needed for what to test real abilities, like the three rules introduced in the Education section). All it needs to do is find the right way to maximize the accuracy of measuring the needed ability. And since it is impossible to create all things as intrinsic core (or at least understanding-based testing), the standard still happens, as a way to maximize what we can, without using all

resources. A good example is that in China, the Gaokao is the best way to ensure fair education (especially what the public believes and is true in cases). However, that leads me to think of a special term that happened in ancient China (before the modern age), where the "科举 imperial examination" had direct testing from the emperor for the top three exam takers in one year's exam. That makes me think that something like an interview could also take place in those academic exams; not only to check the person's personality but also academically. Using scores to reflect ability, as shifting measurements, the requirement for getting a high grade could be to pass a newly designed question/standard/selection for subjective testing, using a combination of evidence to give an obvious and simple score that can be used and compared easily, as the shifting measurement (like the US college application process mixed with the Chinese Gaokao process).

These are just my little understandings of what reminds and inspires me from the article "Post Modernism: What AI Alignment Teaches Us About Education." I skipped certain knowledge-based things that I learned, like proxies, how AI works, and the functions of those measurements, etc., since I think those things don't need to be repeated and is not meaningful in a reflection thus just containing some of my ideas. All these ideas only show my personal insights.

Cen(Sam) Sun