How to Live in a Post-Intelligence-Automation World

As AI continues to advance, it is becoming clear that humanity is on the verge of paradigm breaking changes to structure, culture and thinking. With the pace of progress, it has become imperative to decide the structure and values individuals must have going into this new era, lest an irreversible collapse to an attractor state occurs, preventing individual flourishing. There are a variety of futures, some closer to human flourishing, some closer to utopia-dystopia, some involving transhumanist and biotechnological shifts. Depending on the future, one may require a greater degree of precommitment. Transhumanist futures will require some precommitment of identity, independence and privacy aspects, while perhaps requiring less in the direction of human-AI boundaries, pleasure-seeking behavior and deliberate thinking and action. Futures in which humanist AI shepherds humanity towards cultural richness perhaps require the least amount of precommitment due to said AI limiting most downsides of a technologically advanced future, but could need some commitment towards maintaining ideological independence, as to maintain agency as an essential component of being human. Then there are the Utopia-Dystopias, which can be seen from either angle depending on particular values. Such a future is one where everyone is equivalent to a heavy amphetamine user – which can be classified as an utopia if the criterion is solely pleasure. However, such a future is quite distinct from what is typically considered human flourishing. This type of future will likely need the most restrictive value set. The other most likely event set is in which extinction or near-extinction level events occur. This future is not one that can be prepared for and thus will not be discussed in this essay, and whose probability substantially depends on the pace of AI development, given that mechanistic interpretability techniques will likely be capable of superalignment within 5 years. I will address each of these rough “event sets” individually.

In a transhumanist future, it is likely that humans – which will likely become a more loosely defined, variable group – will maintain control of governance, will continue to have an economic role, and will be compelled to use their capabilities as a matter of survival necessity. Transhumanist adaptations can range from machine-human hybrids, neural prostheses and biological modification. The human brain as it exists now has a much greater hypothetical cognitive capacity than seen even in today’s geniuses – which we can guess from highly efficient structure in humans with normal intelligence and large amounts of missing neural mass – and is largely bottlenecked on quality of instruction and the limits of learning on slow, low-density input material.[12] As hopeful as this is, such a technology also enables people to effectively teach and spread harmful concepts, such as cult-like beliefs, ignorance of empathy and erasure of other “human universals.” Additionally, what it means to be human can become much more ill-defined, as the range of biological modifications and technology integration greatly increase the variance within humans that exist. This may complicate ethics,

1 **Exploiting this efficiency with better training mechanisms enabled by direct neural stimulation is just the beginning; small genetic modifications can have an enormous difference, which we know from the difference in humans from their primate ancestors. We only have two times as many neurons, which would lead to almost no change in ANN power-scaling law terms, and yet, the difference in intelligence is night-and-day. The design space for this learning machine is incredibly large and just starting to be explored – first by evolution, then by us in conjunction with AI.**

2 **See Karpathy’s recent interview on Dwarkesh Patel’s podcast**

identity and empathy. In light of such a future, it becomes important to determine the envelope in which one's own identity may shift in such a future. What range of modifications is acceptable, and to what parts of one's person should modifications be limited to? I personally am comfortable with intelligence enhancements – given such modifications do not substantially alter what I value, and BCIs – also given some boundaries. Excessive reliance on integrated AI, if that instead lies in our future, could hobble one's own problem-solving and creativity. If these traits are valuable – they seem to be an essential experience of being human – it becomes important to deliberately exercise this intellectual activity to maintain our identity. It also becomes important to remain open-minded about how we distribute empathy. In the current day, we can view transgender humans as a prototype of humans in a transhumanist future. We see that they are treated harshly and in a manner lacking empathy by a large portion of society, which results in concrete harm: suicides, suffering and dysfunctional humans. A transhuman future, once set in motion, is not something that can be stopped, despite how uncomfortable the questions raised in such a future are. In the frame of the current morality we have, the only way to avoid committing harm is to maintain an open mind to an expansion of what it means to be human; anything else will mean an irreparable shattering of humanity from one's frame of reference. Importantly, transhumanist futures will likely not face a total cultural collapse, as solutions will be found to avoiding the risks of superstimuli and maintaining the value of human cultural output, maintaining human intellectual relevance (albeit perhaps with a changed idea of what it means to be "human") and human-controlled governance structures that persist values and morality.

An unfortunately more likely scenario than a transhumanist future is a Utopia-Dystopia. On some level, despite all of the intellectual capabilities we have accrued, our desires are similar to rats in a Skinner operant-conditioning box. If we have some behavior that is rewarded, we will continue to exploit it until the rewards stop. This tendency is limited in the world that we live in. Reward doesn't always present as an if-then; it diminishes, changes criterion and requires complex action to obtain. AI and resulting neuroscience advances will allow us to create environments in which the rewards always continue, require no adaptation and are entirely passive – in the sense that one only feels like they are taking action, but that every element of experience is generated by AI, and the thought, action and response are all implanted into the mind through technology. While this can clearly be set up to generate immense pleasure by satisfying ranges of fantasies, it seems to be distinct from human flourishing. Human flourishing typically encompasses a rich experience of life, with elements of wonder, friendship, intellectual achievement, struggle and difficulty. Most simulated environments will singularly focus on only the positive side of these attributes, and will emphasize them to an extreme amount. However, simulated environments do not necessarily have to be avoided; we can create and live in environments that have the richness of life without losing our humanity.