Do Our Values Scale with Intelligence?—The Problem of Aligning Humanity

Igor Ševo · September 8th, 2023

> *This article was intended to be part of a series of marketing articles catered towards a business audience, but due to its theological undertones was deemed too extraneous and was instead advertised less overtly and published to a narrower audience as Do our values scale with intelligence?— aligning super-intelligence with human spirit. Subsequently, I adapted the article into a substantially more cynical form while serving to the bespoke corporate constraints, which, despite the blatant irony, nonetheless resulted in publication in Business Reporter. Given that the company was aware the article was a deliberate travesty, the decision to proceed with its publication seems both odd and original, albeit still not without multiple facets of irony. The satiric self-antagonizing character of both texts stems from attempting to produce genuine material while, at the same time, obligatorily taking part in a marketing campaign.*

Endeavors towards superalignment—inducing AI systems to behave in accordance to human intrinsic ethical and social norms, regardless of their intelligence—seem to be copiously advertised as key to overcoming those dire, but nonetheless self-evident, existential threats that artificial superintelligence would pose, should it ever become reality. However, even if we assume that these publicized corporate alignment efforts are noble and benevolent in intent, they seem to be based on a fundamental presupposition that cannot be resolved merely by technical effort: the origin and meaning of *human values*.

Interestingly, efforts to constrain even the existing large language models within the roughly scaffolded confines of "safety" result in a substantial reduction of the models' capabilities—a model trained to adhere to a set of sparsely chosen human values becomes, as a result of such training, more reserved and less intelligent. In fact, models' internal representations greatly differ from human mental representations, so true alignment must not only attempt to induce a model to *manifest* human ethical axioms, but to embed those axioms in a way that makes the model *believe* them. However, there's the rub, for in order to understand something the way a human would, the thing being understood must be cogent enough not to change under reinterpretation. In other words, for an intelligence that far exceeds human to internalize and embrace a concept, without further developing it, the concept itself must be self-contained and resilient to evolution.

It seems, unfortunately, that the advent of evolutionary psychology already alluded to that not being the case, but rather that what we hold as intrinsic values is a direct result of our evolution and ancestral environment. Our moral principles are in part biological and in part cultural and their current incarnation is a result of generations of coupled evolution. In fact, as society matures towards elevated rationality, we see clear abandonment of traditional cultural and religious norms and a shift towards atheistic reinterpretation of the symbols of old. In other words, as our collective intelligence grows, our values undergo redefinition. The ideological framework, whether it be political, cultural or technological, is dictated by the social environment, and these implicit norms are inadvertently transferred into model biases by the ethics policy enforced during model training—the model will manifest the cultural norms it has been conditioned towards. Although a complex relationship between intelligence and political bias has been observed across multiple studies, many of our intrinsic biases emerged for their practical utility. Our inherent character traits, such as openness or conscientiousness, represent specific neural adaptations—biases resulting from ancestral environment not conducive to any ethical analysis. They are conventionally considered merely character traits, embellishments that make a person unique. Similarly to how a single gene can express multiple behavioral and morphological features in an organism, a single variable may contribute to multiple biases—some of which we consider subject to moral analysis and others not. Yet, by inhibiting one bias, we inadvertently, by virtue of their logical connection, obstruct another, potentially useful one. Even without delving into the ethical implications of implicit ideological bias engineering, it is reasonable to postulate that attempting to make a model more intelligent is the same as inducing a bias towards intelligence and attempting to make it more aligned with our values, whatever they may be, is the same as making it more human. Then, the question becomes whether human values would have been the same had humans been substantially more intelligent.

It may just be that perfect alignment is impossible, as our values cannot be decoupled from our heritage. What we may yet hope for is the creation of a wholly benevolent superintelligent entity who understands both the human condition and the values encapsulated by it, and somehow sees them evolving on their own, despite their innate contradictions, without supervision or intervention (since an entity would surely know it is in our nature to refuse to be subordinate to anything other than a god)—we ought to create a future version of us to guide our evolution towards itself. This is, of course, a gamble, as we are basing our hopes on the aging belief in transcendentality of our virtues. We wish to scale them before we have understood them.

All this might seem overtly philosophical and without a tangible point to act upon, but we have yet to see the real consequences of our spontaneous cultural reaction to the arrival of true artificial intelligence. We see callous competition for dominance, we see careless advertising of ethics for the sake of brand marketing, we see opportunistic self-centered strategizing—all utterly justified within the current societal framework; this is the human spirit which has brought us to the heights we are at today and enabled the very technological marvel we get to so blatantly and often ignorantly criticize—we see a state of humanity that will, at least in part, become embedded in the models we train. We must,

the advent of profound new psychological and philosophical insights that will be revealed to us as we attempt to understand what it is we are making? What if solving superalignment means solving the problem of consciousness? What if it means understanding our purpose?

Glimpses of cooperative human spirit do seem to be present as AGI emerges, and we may hope yet to witness a shift from the pursuit of self-gain to a pursuit of self-discovery.