Review of: "Towards a Comprehensive Theory of Aligned Emergence in AI Systems: Navigating Complexity towards Coherence"

Jeffrey White

Potential competing interests: No potential competing interests to declare.

Apologies that I do not have time to offer a full edit of your paper. The paper is relatively well researched and pretty clear throughout. The main positive of the paper is that it introduces a lot of ideas clearly, evidencing significant work at grounding the issues and for this reason I take time to offer some brief comment on the inquiry.

One outstanding question is about the adequacy of complexity and emergence to describe human values and secondarily tell us how to align machines accordingly. Something like a developmental isomorphism of relevant embodied situational cognitive dynamics seems necessary, or some formalization of similar into simulations thereof. I would suggest looking further at human development, specificially in terms of temporality and alignment to relatively long time scale “values” which may present as invariable patterns which we might translate into moral laws or proverbs that guide action regardless of context and in autonomous contradiction to human leaders. Should such constants be formalized, then it may be that AI can align purpose-directed computation, i.e. find a solution to a moral problem, with human values. Of course, one problem is that such constants might present as variable, and so a univeral statement of such an alignment-machine might present with ecological and cultural distinctions once embedded in operative contexts, which is not only to critique the monocultural McWEFfing of the planet by big-finance driving culture top-down - this is something that a paper like this one must look past - but because an adequate approach must account for such variance while offering a universal mechanism of dynamic in formal terms that can be made into technology and automated. The basic idea is that values emerge as aims or ends through a developmental process that can be formalized. So, emergent behaviors and so on as discussed here in terms of Bubeck and colleagues must be directionalized, they move along a value vector. We might think of this vector as an arrow with immediate timescales dealing with the abstract environment at the bottom of the shaft and with long and eternal timescale values including constant or eternal ethical principles and laws of logic at the top. Basically, emergence of values must specify this arrowhead and its direction and AI emergence of this arrowhead must match - or maybe formally isomorphically simulate, something like that - human emergence of values. The advantage here is that AI can be made to have a very long arrow and to consider very many values over very long timespans and simulating different natural conditions, such as resource situations. In times of scarcity, values guiding more immediate action might change, while eternal principles remain the same. We see such discussion in for example Augustine in the difference between the Cities of Man and of God. The value of this paper is that it is clear, and offers a nice introduction to the basic ideas at work in this area of discussion. The trouble is that it falls short of really solving the problem. The key passages seem to be:

> *The revised differential equations in our framework yield a profound insight into the dynamic nature of AI behavior and alignment. Particularly, they establish that the emergence of behavior at each level, denoted by dB_i/dt, is influenced not only by the traditional factors such as the behavior at the next lower level B{i-1}(t), system states S(t), inputs I(t), rule sets F(t), learning algorithms A(t), environment E(t), and history H(t), but also significantly by the alignment at that level, A_i(t). This novel addition of the alignment factor into our model reflects a more nuanced understanding of the behavioral dynamics in AI systems.*
>
> *The term A_i(t) quantifies how closely the emergent behavior aligns with desired human values or goals at each level. Hence, the equations suggest that alignment at a given point influences the rate and direction of subsequent behavior emergence. Notably, this insight prompts the conjecture of a potential positive feedback loop where the emergence of aligned behaviors may further foster the development of additional aligned behaviors at higher levels.*
>
> *This feedback loop operates as follows: if an AI system's behavior at a certain level is well-aligned with human values, this alignment could positively impact the AI's learning algorithm or function rules. The updated learning algorithm or rules, in turn, could influence the AI to generate behaviors at higher levels that are also aligned with human values. This process would lead to an upward spiral of increasing alignment across different levels of operation.*
>
> *However, the converse could also occur. Misaligned behaviors at lower levels could similarly propagate upwards, leading to further misalignment at higher levels, creating a negative feedback loop. Hence, the framework highlights the need for early detection and correction of misalignments to prevent such unfavorable cascading effects.*

What seems to be missing is what I call a felt interest in the way that the world turns out through action, this being represented by long time scale values in the vector illustration above, though in human beings es embodied typically represented as a utopia or ideal situation. In my work, I work on also individual project ideal situations, what I call self-situations, and the basic idea there - and my basic criticism of derivative alignment schemes such as this one - is that there is no such investment in the AI regardless of how well short and long-term behavioral regularities track with local and perceived social norms, moreover as touched on in this paper in regards to a black box condition, there is might appear to be no way to understand if an AI perform such actions according to such regularities for the right reasons, and so it may be impossible to understand when these might fail to track what we feel to be long term regularities but perhaps we are not sensitive to the same dimensions as the AI or perhaps these change in ways to affect the complex system in unpredictable ways and so on… the basic idea is that attunement to longest timescale eternal moral principles stabilize and anchor behavior in the future and in a noncontingent ideal space or utopia or ideal self-situation. In my old book manuscript, I write about this in terms of Hegel's bathtub… anyways, until there is this formal isomophism in value development and investment in relatively long-term situations optimized for human happiness (if we presume happiness to be a universal value) then AI however well apparently aligned is a sort of wind-up toy that might go off the rails and destroy the planet, or help someone else do it by engineering superviruses in their basement or some other horrible thing like that. AI value alignment for such machines which are so potentially powerful in this way cannot be left to politicians and safety boards, either, with this potential in mind. Recognizing potential risks, for example Yudkowsky has called for the large systems to be shut down. The conclusion of this paper reflects but does not seem to ameliorate or directly address such concerns; it might:

> *By providing a comprehensive framework for understanding emergent behavior and alignment in AI systems, this research contributes to the broader understanding of AI dynamics. It offers valuable insights into the intricacies involved in maintaining alignment and managing emergent behaviors. These findings underscore the significance of continued research and the development of advanced computational techniques to further explore and address the complexities of AI behavior and alignment.*
>
> *In addition, the implications of this research extend to the future development of Artificial General Intelligence (AGI). As AGI systems become increasingly complex and autonomous, it is essential to consider the potential risks and dangers associated with unintended emergent behaviors. The framework's insights into alignment dynamics and the challenges of maintaining alignment over time highlight the need for proactive measures to ensure AGI systems remain aligned with human values and intentions.*