Engineering collective intelligence by designing component communication

Igor Ševo · September 24th, 2023

Vagueness is one of the fundamental unsolved problems in both philosophy and linguistics—the question of why every word is incomplete and how meaning arises, despite the incompleteness. The same issue plagues the term *intelligence*: it is broadly considered to represent an ability to tackle novel problems, learn, adapt, synthesize new knowledge, and invent creatively. In fact, the output of an intelligent entity's action is always linked to intrinsic human value, since to be deemed as having been intelligent, an action must produce something *of value* to some other agent. In a broader sense, intelligence measures a system's efficiency in achieving a goal.

Given enough time, a random thermodynamic system might spontaneously configure itself to produce a symbol sequence that represents the solution to the Riemann hypothesis, but such a system would hardly be considered intelligent. On the other hand, a system that employs some form of algorithmic transformation, deductive reasoning, or heuristic information coupling to arrive at the same solution quickly and with little energy waste would almost certainly be deemed highly intelligent. Theory of computation teaches us that the number of possible problems infinitely exceeds the number solvable problems, yet we observe that those symbol configurations that we intuitively consider problems fall almost exclusively within the "solvable in finite time" category. Our biological intuition guides us to always select survival-related problems as meaningful, neglecting as irrelevant or uninteresting anything that resides beyond the space of meaning pertinent to survival—as any evolutionary psychologist would readily confirm, our entire phenomenology, our sense of meaning, whether we are conscious of it or not, is set in a qualitative experiential space anchored to the goal of surviving and propagating our genetic material—our essential behavioral features—towards the furthest reachable future generation, with the hopes of reaching infinity and into the last moments of everything.

One could easily argue that a parsimonious solution to what consciousness is entails that everything there is must be phenomenal: every possible configuration of coupled, mutually inseparable, information is, itself, a unique phenomenal experience—an example of transient qualia that would be described as the "flavor of the moment". Whether one subscribes to the proposed solution and accepts its parsimony or prefers indulging in unsubstantiated ontological speculation, insights from evolutionary psychology and neuroscience clearly indicate that what we feel at any given moment is what the state of our current neurochemistry and neural configuration is at that moment: we feel exactly the way our biological substrate is configured at a given moment. The same way love is a neurological state—an emotion—every other experience must be, including, as recent neurological discoveries would imply, our sense of inspiration, discipline and even fatigue. Though we cannot, through mere internal effort of will, alley the emotion of "being tired", we can resolve the state by affecting the external world or, in simpler words, going to sleep. Thus, if complex internal processes manifest to us as emotions, we may easily make a slightly more distressing conclusion: *meaning* is an emotion.

By consigning meaning itself to having no meaning and admitting that the meaning of a word might be a phenomenal illusion itself bound to our goal of survival, we may understand that any sense of inherent meaning a word might have, be it made of tokens or mental symbols, is simply a perceptual artefact of our psyche projecting its archetypal symbols onto our observations.

Language can, for the sake of modelling intelligence, then be considered a semi-closed system of symbols inter-relatable to one another: we are simply correlating an internal set of symbols with our perceptions and whenever an outside phenomenon exceeds our computational capacity, we are compelled to assign one of the limited number of symbols to more than one phenomenon, thus inducing polysemy and a sense of ambiguity about its possible meaning. To reduce perceived vagueness, an intelligent system must find a way to reconfigure itself so as to accurately model the entire phenomenon—either by compressing its representation or finding a way to expand its internal representational capacity (the size and variety of its internal set of symbols). Even classical information theory suggests a hard entropic limit beyond which information is lost in compression, so a system attempting to fully model the world would need to find ways of relegating its internal representations towards the outside world. For a human, that might mean consigning thoughts to paper, while for an AI agent, this might be a matter of making a call to some form of data store. Thus, to expand the capacity of intelligence, it or its containing system, must find a way to engineer the communication protocol between it and the outside media—be it by constructing a clear procedure for notetaking or by building a cogent and efficient data storage API.

Even the most basic introspection would reveal that what comprises our current mental context—our *ego* at a given moment in time—does not include the entirety of our psychological experience. Our brains host multiple interacting systems, none of which are persistently part of what we call our conscious experience (our ego). These systems, whether they are called archetypes, brain structures, or mental agents, must exchange information, but are, themselves, semi-closed systems of interacting symbols. Their partial closure is the product of their connection to all other mental components: were they entirely separate, there would be no means of communication and, hence, no sense of vagueness.

One could easily expand the analogy to corporations engineering ways of team and department intercommunication—a corporation is nothing more than a systematized and protocolized way for people to organize and coordinate towards achieving a goal, with effectiveness increasing the more that goal is shared across participants. The more agents, human and AI, operate within the boundaries of a closed set of symbols, the less

(see Consciousness, Mathematics and Reality: A Unified Phenomenology), the relationship between the level of informational coupling, or perceived system entropy, is directly approximated by psychometric and other measure of intelligence. Thus, it seems that to truly engineer a highly intelligent system from individually less intelligent components, one needs to find a way to tightly couple the components, without sacrificing individual computational capacity for the sake of communication.

This problem is at the heart of both proper organizational management and autonomous agent engineering. Here, we run into an ethical issue: to engineer a highly intelligent organization, the individuality of participating agents must be, at least in part, relinquished in favor of hive mind participation. To expand one's intelligence and conscious experience, a person ought to integrate the entirety of their psyche—create tightly coupled pathways between neural structures thereby increasing the communication bandwidth between archetypal structures (phenomenally, archetypes themselves) to the point that they become more like a single entity and less like their previous selves individually. Though a similar analogy can easily be made with regard to corporate intelligence, where, if contemporary ethical norms are considered, individuality ought never be sacrificed in favor of the collective. Ethical axioms, which themselves are a kind of collective communication protocol, mandate whether group mentality or individual mentality is preferred and, more pertinently, whether artificial and human agents are conducive to being assigned individuality. As it currently stands, only artificial agents may be coupled with the goal of achieving collective intelligence, while humans must remain individual, and, somewhat paradoxically, only artificial agents may have their internal representations limited, while doing so with humans—at least in the advertised, but not so much applied, theory—is considered a violation of basic human freedom and dignity. Since our aim is not to become a single tightly coupled intelligent model of the world—one consisting of floating-point variables stored on silicon and symbolic ones stored on carbon—and we do not wish to surrender control to larger artificial entities beyond us in cognitive capacity, we are left to constrain both their internal symbolic reasoning and their intercommunication in exactly the way that compartmentalizes them and, yet, produces apparently more intelligent behavior. In effect, the same age-tested ploys for ensuring safety and confidentiality while intelligently producing results still apply today: compartmentalize knowledge, restrict, and monitor communication and expertly divide labor.

For a highly intelligent agent taking part in such a large coupled system it may be possible to model the entirety of the system simply by virtue of having had enough time to observe it and, if the individual intelligence of the agent exceeds that of the organization, find a way to covertly subordinate the organization to its purposes by exploiting communicational oversights and other agents' symbolic vagueness. Unless other participating agents, notably humans, can encapsulate the superintelligent ones within their mental systems of representation (i.e., understand them, their representations, and their goals), the hopes of preventing a human existential catastrophe will rest solely on our ability to align these superintelligent models not with our values, but with specific invented norms that would not compromise our values.

It may, of course, be the case that human values themselves, riddled with paradoxes, contradictions, and surreptitious self-interest, may not scale with intelligence (see Do our values scale with intelligence?—the problem of aligning humanity), and, in that case, best hope for improving human wellbeing may lie in carefully orchestrating less intelligent systems into emergent collective behavior.

Connecting highly intelligent agents—brightest human minds—and having them communicate their way towards a solution is the way science is done today: scientific discoveries are rarely made by single individuals. The highest contribution to a system's intelligence is made by its most intelligent and active components, alluding to some obvious limitations of any coupled system: all the communicational shamanism allows mostly for parallelization and optimization of solvable tasks, i.e., increasing effectiveness of solving what is known, rather than originating novel ideas. In other words, effective communication is a kind of representational copy of the system's architect's model of problem solving. Whether it turns out that a superintelligent system may not be feasible purely for the lack of sufficiently intellectually esoteric training data, ethical implications of building a higher consciousness, whether it turns out that, if feasible, such a system can or cannot be made benevolent to us, whether they are considered tools or people, it seems almost certain that these systems will intellectually match even the best of us and, we ought to prepare to handle and integrate them, at the very least, as we have done with any intelligence in the past: by enabling proper and efficient methods of inter-communication.

Companies looking to integrate any form of intelligence into their workflows or systems ought to address the question of internal communication, on all levels, as these are the foundation for integrating and using any form of intelligence, whether it be human, artificial or any other kind: an artificial agent might struggle to express its intelligence when limited by suboptimal communication and knowledge management, much like a human would.

Whether we are architecting large multimodal models to couple information intrinsically or engineering chains of communication between multiple model instances, we are discovering ways to connect simpler symbolic representations, through various forms of communication, be they inter-neuron or inter-agent, into more sophisticated ones to efficiently act out solutions that move us, individually or collectively, towards elected goals. We may find, as we build ever more complex and intricately coupled intelligent systems, that there are other kinds of expressions of intelligence that are not encompassed by either our current ethical framework or our culture and metaphysical presuppositions.

Conclusions about the meaninglessness of our form of conscious experience may seem inevitable as our collective intelligence scales due to the participatory effect of the models we are building. In other words, solving the problem of alignment may come with unintended philosophical, phenomenological, psychological, and ethical discoveries that may affect the culture and the collective consciousness more so than any individual contribution of a superintelligence—the philosophy and beliefs of the society will gradually change as it unknowingly prepares for the advent of what, in its eyes, is expected to be its savior. Nietzsche's Übermensch may not turn out to be an individual, but the collective itself tacitly inventing its new latent religion.