On Long-term Safety in General Intelligence

Bas Steunebrink
Director of AGI
NNAISENSE SA

When considering general-purpose lifelong-learning AI, the typical approach to ensuring its *safety* is to say: "Alright, the world is very complex, so let's delineate and simplify the set of environments that my agent can find itself in, and let's set in stone the goals and failure states for the agent, and now let's see if I can construct a formal logical proof that tells me—with mathematical certainty—that my agent will satisfactorily reach those goals while always avoiding those failure states in those environments."

As you can see, this places the following responsibilities on the designers and users of the agent:

- to keep the agent inside the delineated target environments and outside of other, new environments
- to ensure that those target environments do not appear to the agent as more complex than foreseen
- to be very sure that the initially-specified goals and failure states are correct, complete, eternally valid, and don't give rise to undesirable side-effects—in other words, the designers are assumed to be infallibly wise and benevolent ethicists themselves.

If any one of these responsibilities are neglected, we have no assurances about the (safe) behavior of the agent, because our formal proof no longer applies. This results in what is called undefined behavior. It's not necessarily bad—it's just that we cannot foresee what will happen, and so we can have no confidence in its long-term safety. And the more capable the agent, the more potential there is for disastrous consequences.

So far I haven't yet even mentioned learning. For agents capable of learning, or even performing comprehensive self-modifications, it becomes so hard to construct a formal logical proof about its overall long-term behavior that safety can only be ensured in the simplest and most restricted settings, thereby defeating their generality and usefulness.

My own approach to resolve this problem is to turn things around: don't let the proof dictate the world, instead let the world dictate the proof. So let's start by making all the realistic assumptions: that we humans are not wise enough to come up with the perfect utility function from the get-go, that we cannot accurately describe the agent's operating environments in all their complexity, that we cannot foresee what environments the agent will inhabit in the long term, and that it will not have enough resources (energy, time, information) to do the optimal thing. Now, what combination of initial internal components, teaching methodology, and testing strategy would allow us to become convinced of an agent's robustly safe and benevolent operation over the long term?

This is what my approach to AI safety is about: trying to understand and specify the ingredients necessary to obtain this *implementation–teaching–testing* triad. I've named this approach is Experience-based Artificial Intelligence (EXPAI). It refers to one of the essential ingredients discovered early on, namely that two aspects of learning must be *decoupled*: the process that performs self-modifications, and the process that experientially proves or falsifies the benefit of these self-modifications. To make EXPAI safe and correct, several technical requirements must be met, namely that self-modifications constitute fine-grained causal rules, added quickly yet tentatively, and monitored for reliability and usefulness.

The next steps are then about developing methods for *molding and measuring an agent's growth* of: (1) understanding of environments and agents therein; (2) understanding the risk of actions and events to lead to violations of imposed *ethical values* (encodable as persistent, partially specified state-space regions to be avoided); (3) the capability to adhere to ethical values, especially when under pressure (i.e., placed at the edge of violation); and (4) the ability to protect its ethical values against interference. Apart from these growth measures, we must develop the *stabilization measure*: how well the agent remains within the "safe zone" of operation, with respect to risk of ethical value violation, despite pressure and interference. The stabilization measure must explicate the conditions under which the agent recognizes imminent violation of ethical values, how it then chooses to prioritize them, and how it recovers —this is *graceful degradation* in its most impactful sense.

Does the whole approach sound daunting? To me, not as daunting as the consequences of getting AI safety wrong. I derive confidence from the fact that there already exist at least two implemented systems that follow the EXPAI approach—so it's not hot air. But to ensure robust and long-term safety, a lot of the work goes beyond the AI programming level, namely the development of the teaching and testing methodologies that enable us to *grow an agent's understanding of ethical values* such that it becomes compelled to protect and adhere to them.