Beyond Control: Becoming the Intelligence We Build

Zonghuan Xu · September 2026

On bidirectional alignment, human transformation, and preserving identity as we become more capable intelligence.

Discussions about the future of AI often take a particular arrangement for granted: humans retain their present biological form and cognitive capacities, build increasingly powerful artificial intelligence, and ensure that it always obeys us.

I am increasingly pessimistic about whether this arrangement can last.

If future AI far surpasses humans at understanding the world, making plans, conducting research, and organizing production, will we still be able to judge what it is doing, understand its reasons, and intervene effectively when necessary? If even the knowledge and technology needed to maintain control increasingly depend on AI itself, what does “humans retain ultimate control” actually mean?

I cannot prove that this relationship must fail. But I do not want to stake humanity’s long-term future entirely on its lasting forever.

The alternative I have in mind is for humans to participate in the transformation themselves, gradually becoming part of that more powerful intelligence.

Rather than demanding that something far more capable than me obey me forever, I want to ask: could that future intelligence still be a continuation of me?

When people discuss losing control of AI, they often imagine machines awakening, developing self-awareness, and deciding to attack humanity. The scenario that concerns me requires none of these assumptions.

To make AI more effective, humans will give it increasing access to information, long-term memory, tools, money, computing resources, and control over machines. The more valuable a system becomes, the stronger the incentives to expand its scope for action. As perception, planning, execution, and self-improvement form a closed loop, its effects on the world may exceed what any individual authorizing it can understand and review step by step.

A human pressing the first button does not mean that the process remains under human control.

The claim that “AI exists only in the virtual world” offers little reassurance either. Humans organize matter through knowledge and tools; AI could likewise affect the physical world through machines, infrastructure, and organizational processes. Its present dependence on humans to provide those conditions does not establish that this dependence must continue in the same form.

No hatred of humanity is required. For a system continually optimizing an objective, resources, freedom of action, and the ability to keep running may all have instrumental value. Unless human interests are adequately reflected in the constraints on its behavior, people could become obstacles, costs, or overlooked side effects of its pursuit of other outcomes.

This is a judgment about risk, not a prediction of the apocalypse. The *International AI Safety Report 2026* discusses loss-of-control scenarios involving human marginalization or even extinction. It also makes clear that experts strongly disagree about their likelihood, and that the systems assessed in the report do not yet possess the full combination of capabilities needed to cause such a loss of control. International AI Safety Report 2026.

But I think the question requiring an answer is no longer only why control might be lost. It is also why we believe that humans as we are today can indefinitely control intelligence that keeps surpassing us.

Humans have always used tools more powerful than themselves. Cranes are stronger than we are; calculators are better at arithmetic. A gap in strength or a particular capability does not, by itself, cause control to fail.

The difficulty is that the capabilities we are now trying to amplify are precisely those on which control depends: understanding, prediction, judgment, planning, and discovering courses of action that the other party has not anticipated.

When the gap becomes enormous in these capabilities too, how far can our past experience of controlling tools take us? We can use other AI systems for oversight, but that also means that control increasingly rests on an artificial intelligence infrastructure rather than on humans' independent understanding and judgment.

Perhaps such an infrastructure can be made reliable. I hope so. But a future in which human cognitive capacities remain largely unchanged, external intelligence keeps advancing, and humans nevertheless retain stable supremacy is not an outcome I can take for granted.

Meanwhile, bringing AI development to a complete halt is also hard to regard as a dependable global solution.

The potential benefits for research, medicine, production, and competitive advantage are immense. Even if many participants acknowledge the risks, they may be unwilling to bear the cost of slowing down alone. We could find ourselves in a situation where everyone wants the whole system to be safe, yet everyone has reasons to keep accelerating.

This makes me even less willing to consider only one direction: keep making machines stronger, keep humans as they are, and strengthen control.

Why must all the change happen on the AI side?

Why must alignment mean that an increasingly powerful system continually adjusts itself to humans whose cognitive boundaries remain largely fixed?

Humans can change too. We could use AI to extend memory, reasoning, and perception, change how we learn, and perhaps gradually alter the physical substrate of cognition itself. The future need not contain only two permanently separate kinds of agents: humans and external superintelligence. We might actively participate in becoming that more powerful intelligence.

This is what I mean by bidirectional alignment.

AI must understand human values, intentions, and experiences, while humans actively participate in changing their own cognitive structures. The relationship could develop from tool use to sustained collaboration and then to deeper cognitive integration. Alignment would need to maintain the connection between this joint process of change and human goals, agency, and identity.

This is entirely different from requiring humans to accept goals imposed by AI. Someone forced to adapt to a system while gradually losing the ability to exercise judgment has not thereby achieved better alignment. The human must participate in the transformation, rather than merely be its object.

The path I can imagine begins with the external AI tools we already know.

A tool becomes a cognitive assistant with long-term memory, participating continuously in a person's learning, research, and life. As interaction deepens, external memory and computation may become stable components of cognition rather than services called upon occasionally. Brain–computer interfaces or other technologies could further narrow the distance, integrating more cognitive processes with non-biological systems. Further ahead, the distinction between biological and artificial might cease to carry the significance it has today.

The envisioned endpoint is that human life and cognition need not remain confined to their current biological implementation. If some non-biological systems can support more powerful and extensive cognitive processes, could we gradually move toward them, instead of remaining permanently outside them and trying to control them?

I do not regard this as an already validated technological pathway. We do not know how far cognitive functions can be integrated or transferred, much less which changes could preserve the continuity of the subject. These unknowns are precisely the questions I believe deserve investigation.

There is a boundary between becoming and being replaced that we cannot skip over.

A system that copies my memories, imitates my language, and inherits my goals is not necessarily me. If it is simply a new subject, then however powerful it becomes, it has not achieved the continuation I am seeking.

That is why I am more interested in gradual integration: an agent already experiencing the world and making choices progressively extends its capacities and changes how it is implemented. Each step remains connected to its preceding cognitive activity, rather than ending a person's existence and then constructing a system that claims to be that person.

Gradual change does not itself prove continuity of identity. But it gives us more concrete questions: which cognitive connections must be preserved? Which changes still constitute my development? When does apparently successful enhancement actually amount to replacement?

Identity continuity should become part of alignment.

It is not enough for a future system to match my present preferences. People learn, reflect, and revise their goals; I do not want my future self permanently confined by today's judgments either. What I want to preserve is the subject who can continue to understand, choose, and revise their own direction through change.

This places a stronger requirement on bidirectional alignment. AI must not make me easier to satisfy or predict and then declare alignment a success. Cognitive enhancement must expand my capacity to participate in the future, rather than merely increase a system's capacity to influence me.

Of course, human–AI integration will not automatically eliminate conflict or guarantee safety. Closer integration could create deeper forms of manipulation and dependence; people who become more powerful intelligences could also harm others. Technical constraints, institutional arrangements, and genuine rights to refuse and exit would still be necessary.

But these difficulties do not lead me back to the original default. They mean that becoming must also be carefully designed, not that permanent control is therefore more reliable.

My position is that preserving humanity's present biological form and preserving a human future should not be treated as the same thing.

Biological form need not change, and nobody should be forced to change it. But if future intelligence reaches a level that today's humans struggle to understand or participate in, I want our options to include more than depending on its goodwill, relying on constraints it maintains, or accepting marginalization.

We should also investigate how to enter that future ourselves.

I am pessimistic about humans remaining as they are while permanently controlling AI far more capable than themselves. I would rather explore an alignment that allows both sides to change: AI moves toward humans, while humans, through choices of their own, progressively grow into more powerful intelligence.

I am not trying to prove that a soul can be uploaded. I want to find a way that might allow "me" to transition continuously into another form of life without having to assume that a soul exists.

Rather than placing all our hopes in future powerful intelligence obeying us forever, I want to know whether we can become that intelligence ourselves—and, through the process, remain us.