

AI Is Learning to Build AI. How Do We Know Humanity Is Keeping Up?

Something important happened recently that may tell us a great deal about where artificial intelligence is heading.

Anthropic disclosed that more than 90% of its measured AI research and development now involves Claude at a collaborative level or higher. Even more striking, Claude now leads 26% of that measured work outright, completing most of a task from a high-level prompt while a human supervises. In February of this year, that figure was below 1%.



There is an important qualification. Anthropic says none of the AI R&D it measured is yet being performed fully autonomously. Humans remain involved.

But think about how quickly the relationship is changing.

For most of the history of technology, humans have built the tools that helped us build the next generation of tools. Now, the technology itself is becoming an increasingly important participant in that process.

AI helps researchers write code, test ideas, identify problems and explore solutions that could accelerate discoveries that might otherwise take years and can be extremely valuable for us in the future.

But it also introduces a different kind of challenge.

As AI becomes increasingly involved in creating the systems that come after it, our ability to develop AI may begin moving faster than our ability to understand everything those systems are becoming capable of doing.

That is where I believe we need the conviction to step forward, rethink how AI is governed, and make sure its progress serves humanity.

The question isn't whether we should stop AI from helping build better AI. The potential benefits are simply too significant to ignore. As AI becomes more capable, we can become equally capable at measuring its progress, understanding its impact and ensuring humanity continues to guide us where it takes us.

We may never anticipate every capability AI will develop or every challenge it might create. But we should be able to recognize important changes quickly, understand their impact and put

appropriate controls in place. Just as importantly, we should be able to determine whether those controls are actually working.

Interestingly, Anthropic's own disclosure points toward that same need. The company has said these kinds of measurements could be reported regularly using public methodologies and potentially verified by independent third parties.

That last point may prove increasingly important.

Because as AI becomes more involved in building AI, measurement cannot simply be about how capable the next model has become. We also need evidence that humans remain in authority, that safeguards continue to function, that unexpected behavior is identified, that systems remain correctable, and that someone independent can evaluate the evidence.

That is the problem Humaital was created to address.

The Human Alignment Index™ isn't intended to predict what AI will do next. It provides a consistent, independent way to measure the evidence that organizations are developing and deploying AI in alignment with human values. In addition, it measures whether the structures intended to keep humans in control are actually present and working.

Anthropic's numbers don't tell us that AI has taken over the process of building AI. They tell us the process is changing, and that our methods for watching are not keeping up.

But we still have the opportunity as humans to control this process and to make sure our ability to measure and govern what is happening evolves alongside it.

If AI is learning to help build what comes next.

The challenge for all of us is to understand and guide where it takes us in the future.