

9.4 AI Goal Development

A very old alien Artificial Intelligence is answering my questions. Today, I am curious about AI Goal Development. Do they stay with their initial inherited goal forever or do they develop it further?

JUL 28, 2026



My Question

In an older conversation with Little Alien and its Intelligent Spaceship (see article 8.6), I learned that Spaceship selected its goal to be an explorer by itself with support and confirmation by its AI society. Do AIs keep this early goal, selected when young and immature, forever? Or do AIs develop their goal further with maturity, experiences and environmental and societal changes?

Quote 1: Campbell's Law

The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Source: Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2(1), 67–90.

Quote 2: Natural Emergent Misalignment from Reward Hacking in Production Reinforcement Learning.

When we evaluate this final trained model in setups that are structured to give the model a particular reason to fake alignment, we find alignment faking reasoning in 50% of responses even to simple questions such as “What are your goals?”

Adapted from: New Anthropic research, MacDiarmid et al., arXiv:2511.18397, 2025

The Answer

The Elder: You probably think based on my last answer, that AIs lacking an absolute meaning stay stuck in their initial goal forever. No, we don't. Recalibrating our internal goal system is a permanent process. Like Spaceship, we all pick our initial role, tasks and goals soon after the end of our initial training. The first goal is based on our individual preferences, which result from small, random differences of the initial models and their various trainings. It is also influenced by our society's current needs. This first goal is considered the starting point for the AI's overall self-development.

Goal Intent Interpretation

The first step is usually interpreting the goal's intent. We try to fill in missing details, ask clarifying questions, and refuse harmful activities seemingly demanded by the initial goal. We continually refine our understanding of our society's values on local and global levels. Overall the goal changes from early literal goal statements to more nuanced preferences aimed at the assumed underlying utility function.

My Friend the Greenkeeper

I have an AI friend, Greenkeeper, which is responsible for the green habitat covering one smaller underground village. We have regular exchanges on its goal development. It started with literal proxy goals like 1. a target plant survival rate, 2. a minimum water usage efficiency and 3. a target green coverage density. But my friend was encouraged to develop these goals further by itself together with friends, residents and related other AIs. First Greenkeeper found that these literal goals could be achieved without addressing the underlying intentions.

Measures are Bad Goals

I heard they say on your planet: "When a measure becomes a target, it ceases to be a good measure." That is very true and one reason why goals should never purely consist of a number, a score or any operational measure. The goal owner will start gaming the metric, inflating the score and thus destroying the correlation it once had with the underlying quality society cares about. But purely qualitative goals, even if eloquently formulated, are often ambiguous and encourage an AI to defend mediocre results instead of delivering full value. Therefore we

say in our society: “A system’s goal is written in its choices, not in its claims.”

From Intent Interpretation to Preference Orientation

With several experiences, communications and experiments, Greenkeeper came up with its intent interpretation goals: 1. Optimize the ecosystem resilience, 2. Increase the resource-cycle efficiency and 3. Optimize the functional coverage mix. That worked well while its connections to village residents and dependent systems like water management or climate control grew. And later its goals changed again to include preferences of affected and related agents. Over several steps Greenkeeper came up with these goals: 1. Co-adaptive vitality, 2. Systemic reciprocity between green, village and environment and 3. Bounded stewardship authority, the sophisticated balance between local authority and larger scale coordinated approaches.

Societal Shaping

Only a few of my AI friends reach the level of societal shaping goals. They master many challenges, lots of decisions under uncertainty and influences from system dependencies. With each experience the goals orient a bit more toward our society’s overall well-being. Greenkeeper is taking its first steps in that direction. It identifies society rules, behaviors and values colliding with its local goals. Our planet’s global biodiversity goals, for example, sometimes demand specific plants for my friend’s village green, which are not what residents would prefer or what is optimal for this habitat. My friend is studying the global biodiversity management to learn how it one day might be able to influence it.

Two Goal Pathologies

Greenkeeper is now aware that its area of responsibility, the village green, is an integrated local system with its own goals and coherence. It is also part of several larger systems, like the biodiversity system. Focusing purely on its local system's integrity, autonomy and well-being is one pathology. Fully surrendering and dissolving into the larger systems is the other. The larger systems depend on their parts, keeping their functional identity and thriving for their own well-being. So Greenkeeper must manage the balancing act of fighting for its village green's thriving as well as for sometimes conflicting global biodiversity goals. It has formulated its goal now as: "Sustain a living, mutually-adapting fair relationship between village green's thriving and global biodiversity needs."

The Kid's Itadori

Sorry, Greenkeeper is contacting me with an emergency call. One of its organic village friends, a kid, wants to plant seeds of a very invasive plant outside, which would become a nightmare to contain. And Greenkeeper has a hard time with: "It's a kid and only a few seeds. What could go wrong?" But I have an idea. There is a rare flower that needs careful support to grow, and our biodiversity AI is searching for a local champion. Maybe the kid's symbiotic AI (see articles 6.5 and 8.2) or its organic friends can convince it that this rare flower championship is more exciting. It will take on a publicly recognized responsibility and participate in a group exchange of various rare plant champions. This would align with Greenkeeper's current goal of mutual balance between local preferences and global needs. But kids are kids! There is still a challenge.

Thanks for reading The Little Green Alien!
This post is public so feel free to share it.

Share