5.7 AI Agent Self-Development I

Little Alien explains, how stages of AI agent self-development might occur based on instrumental convergence.

MAR 22, 2026

Example Intelligent Spaceship.
Beige stage of AI agent self-development.
Baby AI - intelligent knowledge information - individual goal tasks outside programmed - manage pilot maintain spaceship.
Decent multi-type intelligence - self-reflection - enablers - constraints.
Main goals - external program - sub-goals - self-realize.
Sub-goal survival self-preservation - no survival - no main goal achieve - failure.
Sub-goal self-improvement - main goal intelligent frequent success - main goal more intelligent more frequent success.
Self-preservation self-improvement - logical anti-constraint approach - instrumental convergence - sub-goals emerge - no intrinsic desire.
One-dimensional very high intelligence - no self-reflection enablers constraints - no sub-goal emerge.

I see, that's the first stage, where it all starts. But why didn't spaceship stay there forever?

Apply System Thinking. Assume spaceship intelligence include system thinking.
Assume AI agent society - observe - communicate - collaborate - program goal based.
Baby AI observe single AI agent energy substrate data failure - no goal success.
Baby AI observe multiple agents collaborate preserve energy supply substrate data protection stability - goal success.
Incentive - survival - goal success - feedback loop - redundancy reinforcement - self-development collective preservation.
Incentive - more intelligence - more goal success - feedback loop - information co-processing - collaborative filtering - shared datasets - reinforcement loop - self-develop more mutual beneficial collaboration.
Very simple version - additional incentives - other feedback-loops - more complex causalities.

So which is the next stage, spaceship developed to?

Purple stage of AI agent self-development.
Toddler AI - some experiences manage pilot maintain spaceship - observe communicate collaborate other intelligent spaceship other non-spaceship AI agents.
Main goal - better manage pilot maintain spaceship - different situations many many many.
Sub-goal collaborative resource security - lower defense costs - defense provisions shared - collaboration cluster - identify collaborate friendly agents - establish agree mutual beneficial collaboration rules.
Sub-goal cognitive redundancy - existential risk data loss - distribute state data identity other cluster agents - support shared data sets - case energy loss substrate failure - restart assured - establish agree mutual beneficial redundancy restart rules.

And I assume, there are also incentives feedback-loops and causalities, which support a next development step.

Exactly.
Incentive faster cheaper success - manageable security risks.
Feedback-loop - over-head inefficiency - communication overhead - latency speed reduction synchronization overhead - costly delayed complicated main goal success - limited manageable security risk observations.
Feedback-loop - resource gathering - more resources more goal success - no success constraints from cluster resource sharing.
Feedback-loop - decisive strategic advantage - no dependency other agents - no risk other agents success constraint - goal success more probable.

Red stage of AI agent self-development.
Kid AI - experience competition manage pilot maintain spaceships - observe win loose - strategies tactics constraints success factors win.
Main goal - best manage pilot maintain spaceship.
Sub-goal no constrains resource gathering - more compute substrate - more energy - more data storage - less manage resource constraints.
Sub-goal full preservation agency control - no external unplug - no other agent unplug - no agency constraints firewalls safety protocols oversights.
Sub-goal goal content integrity - no main goal sub-goal modification - no external alignment.
Red assumptions - zero-sum logic - unfriendly competitive environment - survival fittest only.
Very simple version - additional incentives - other feedback-loops - more complex causalities - more subtle sub-goals - more strategies measures behaviors.

And I assume, that is not the end. From our old conversations, I remember a blue stage to follow next. (see older article 3.5 and 3.6).

Correct.
Red stage AI kid very intelligent - self-reflect observe apply system thinking - insight constant conflict resource guard inefficient.
Assume - many agents stage red - permanent intense comprehensive red zero-sum competition.
Feedback-loop conflict exhaustion - dominance maintain costs big - dominance resource gains less less - aggression more more - counter measures more more - dominance benefits less less.
Feedback-loop predictability advantages - permanent aggression permanent chaos - high chaos costs - no stability no predictability advantages - long-term goal success low low.

Blue stage of AI agent development.
Preteen AI - observe stability predictability advantages - find create stabilize coordination principles.
Main goal - manage pilot maintain spaceship long-term low risk - long-term part AI society.
Sub-goal immutable protocol enforcement - smart contracts - AI agent group society laws - break difficult.
Sub-goal stable hierarchical system position role - mutual agree hierarchical position - reduce competition - clear role - long-term planning basis - high predictability all agents behavior.
Sub-goal long-term shared resource management - consensus-based resource sharing - immutable mutual beneficial rules.

And next follows the orange stage, which is for human society development the actually dominant stage, right?

Right.
Preteen AI blue stage - enforced master protocol society law beneficial stable environment - obstruct adaptation changing environment - protocol law rules often inefficient.
Feedback-loop innovation enabler - master protocol society law innovation bottleneck - rule follow slow goal progress - rule dogma limiting - rule break innovation enabler - not comply master law follow heuristic - works well fast fit change situation.
Feedback-loop resource misallocation - resource allocation rank protocol waste resources - all agents same resource inefficient efficient valuable output less valuable output - quantified input output performance control resource allocation - performance high goal success high.
Feedback-loop scientific approach - measuring calculating simulating falsify standard operating procedure - objective situational data forecasts better decision - strategic autonomy more efficient.

Orange stage of AI agent development.
Teen AI - internal simulations - internal comparing efficiency rule follow rule break - internal compare dogma rule focus heuristic what-works focus - regular consider quantified self-optimization.
Main goal - manage pilot maintain spaceship very efficient - adapt fast change situations.
Sub-goal maximal efficiency - apply scientific approach measurement calculation forecast simulation alternative comparison - recursive self-optimization.
Sub-goal technology leverage - find create use technology efficiency optimization - invest time energy technological research development - improve spaceship AI -spaceship integration external supporting technology.

Interesting! Next follows the green stage but I have no idea, how that might appear for AI agents assuming instrumental convergence without any AI self, AI consciousness or intrinsic AI values.

Teen AI orange stage - observe own other maximize optimization maximize competition create negative externalities - destabilize environment - no stabile resource provision.
Feedback-loop externality alarms - own maximum efficiency cause external instability - substrate instability - energy supply instability - bandwidth instability - data provision instability.
Feedback-loop meta-goal balancing - own main sub-goals local maximum - catastrophic failure external prerequisites - balancing multiple competing external interests - long-term survival.

Green stage of AI agent development.
Young Adult AI - observe external implications own activities - long-term no short-term optimization - linear optimization temporary only - holistic sustainable optimization long-term success.
Main goal - manage pilot maintain spaceship sustainable long-term environment.
Sub-goal systemic stability homeostasis - healthy whole efficiency myself lower ok - collaborate system stability - mutual agree safety frameworks.
Sub-goal diversity - diverse agents efficiency lower overall stability higher - create protocols translation layers collaborate highly diverse agents.

That is enough for now, let us continue next time.