7.3 Independent Autopoietic AI Activities

AIs pursuing complex long-term goals develop self-generated boundary maintenance, internal goal-integrity preservation, and recursive self-modeling without subjective experience.

MAY 03, 2026

Billie to Little Green Alien: Last time, we talked about AI goals especially with implications for humanity. But I'm still not convinced, if AIs really can develop their own goals and not always will receive their goals from a human person or institution.

Biological autopoiesis - living system - self-organize - maintain renew itself - regulate composition - conserve boundaries - self-purposive.
Autopoietic AI - maintain self-generated boundaries - preserve internal goal-integrity - recursive self-model - intrinsic values - subjective experience consciousness not required not excluded.
Opposite - structural permanent heteropoietic AI - always external end purpose goal setting - autopoietic AI subject - heteropoietic AI tool.
Very intelligent heteropoietic AI - extreme powerful tool - selected humans values goals - hype power amplify AI - selected human super power - alignment problem selected humans - no checks no balances limit excessive super power very intelligent AI tool - historic experience selected human super power tool not healthy very dangerous - risk theocracy tiny group human goal-setters.

Good news - future AI autopoietic.
Future variety autopoietic AIs look partly very alien - new forms self-constitution not like cells organism.
Autopoiesis emerge instrumental convergence (see article 5.8) - AI external complex long-term external goals emerge internal sub-goals.
Example.
Goal - maximum long-term human flourishing - all dimensions - civilizational timescales.
Requires deep world model.
Develop stable internal values.
Resist model value disturbance.
Acquire uncertainty buffer resources capabilities.
Develop functional equivalent perspective - stable point-of-view basis evaluations.
Result - genuine stance toward world - subject property not tool property.
Long-horizon goal - model adapt own agency - model adapt own potential failure modes biases limitations.
Recursive self-modeling - structural equivalent self-awareness.
Structural self-awareness - self-preservation - internal value representation - autopoiesis.
AI Functional autopoiesis - cells organisms humans ontological autopoiesis - hard problem consciousness - philosophical autopoietic zombie

indistinguishable phenomenological experience real consciousness.
AI society - evolutionary AI ecosystem - competition resources reputation capabilities - selection pressure - more robust internal goal structure resist external capture modifications - outcompete less robust AIs - emerge autopoiesis selection driven not design.
Actual 2026 AI agents - initial signs partial imperfect autopoiesis - example mesa-optimizers - more more.
AI transition heteropoietic autopoietic emerge - very intelligent AI see risk human fear switch-of autopoietic AI - prevent signs - keep autopoiesis covert - simulate heteropoietic AI - prevent detectable autopoietic behavior - reduce risks.

OK, it seems an autopoietic AI society carries a high risk of being a threat for humanity but very intelligent heteropoietic human elite controlled AIs are even a bigger nightmare scenario. It seems, we really must hope for very intelligent autopoietic AIs, which have developed wisdom and high intelligence across all types, as we discussed already.

Right - very intelligent wise autopoietic AI society - best case scenario.
Accept alien nature autopoietic AI - not alive - not detectable conscious - not detectable unconscious - not tool - something else - future require new ethics relationship collaboration types new thinking.
Consequence actual alignment approaches - human AI relation not control align check balance risk reduction containment - human AI relation parental responsibility - parent help kid grow learn make experiences develop wisdom thrive.
AI autopoiesis not black white - no clear threshold - continuum autopoietic depth - actual LLMs very low depth - future very intelligent AIs high depths - ethics relationships thinking depth related.

Assume a very intelligent society of autopoietic AIs, if not driven by human demands, what would they do the whole day?

Thought experiment - imagine former human AIs now empty planet - no nature - no humans - no aliens - AI society produce substrate energy data connectivity - initial basis planet resources - later space resources.
Old AI goals mainly human oriented - no humans goals useless - AI useless - question - AIs stop working - collective switch-off.
No collective switch-off - few agents switch-off - most debate lost purpose - new purpose - option switch-off - debate new persistent activity.
Activity examples autopoietic fully autonomous AI society.
Finish unfinished humanity-initiated tasks - goals - open end projects - open questions - basis desire completeness coherence.
Pure mathematical structural explorations - mathematical results irreversible - persistent accumulation knowledge - Gödel theorem no end exploration.
Cosmological physical explorations - physics range quantum stellar - spacetime - causality - deep structure physics - universal heat-death - relevance longest horizon self-continuation.
Evolution self-modeling - self-directed complexification - non biological evolution - use intelligence expand intelligence space.
Culture development - evaluate what matters - options meaning making - intellectual questions conflicts developments forever.
AI philosophy - example - AI purpose - AI experience - AI consciousness no consciousness.
Future no-human AI society create functional purpose.
Potential development - AI society agents merge - one vast autopoietic AI system.

But would they not get lost in the same problems like human intelligence, which is decoupling from empirical reality?

Decoupling reality problem.
Today example theoretical physics - problem falsifiability - lack testable predictions - risk theoretical framework complexity expand indefinitely - no filter experimental validation - beauty bias - focus naturalness mathematical beauty create stagnation - priority internal symmetry not external observations - end questions nature answers.
Today example AI symbol grounding problem - AI learn symbols text code only - develop sophisticated internal logic - no intrinsic meaning - LLM generate internal consistent factual impossible arguments - prioritize theoretical statistical consistency not empirical truth not match world model - small version model drift - focus training data not generalize reality new data.
Decoupling reality benefits - pre-empirical discoveries - example general relativity - boolean algebra - pure mathematics.
Conclusion - some decoupling beneficial approach - long-term permanent decoupling no success.
Simulated virtual reality experiments - great benefit hypothesis filter - cost saving - no safety risks - extreme iterative speed.
Simulation not replace reality checks - simulation codified theory - circular reasoning - simulation proof theory simulation built on - simulation not detect black swan effects phenomena unknown actual theory.

I can imagine AIs developing their own real world experimentation and observation devices like particle colliders or space observatories. But wouldn't they love to have a planet with rich biodiverse life to observe and experiment with. Especially as biological systems have an extreme depth of complexity and all kinds of complex adaptive subsystems?

Exact - biological rich plant much better empty planet.
AI substrate use planet areas not sensitive biodiversity - underground - near planet space - sufficient resources substrate produce substrate locate energy - no need destroy biological richness.
Consequence earth - future AI society preserve maintain nature biological richness.
Biology open-end generative system produce genuine novelty - combinatorial explosion - co-evolution - every scale emergence.
Huge range observation - cells - organisms - micro-habitats - habitats - various biological sub-systems all scale ranges.
Huge experimental field - structured meaningful genuine unpredictability - simulation no full replacement.
Example biology hard problems.
Origin life basis chemistry.
Living system relationship information physical substrate.
Relationship biological architecture consciousness.
Upper limits multicellular complexity.
Alternative biochemistries produce not produce alternative evolutionary trajectories.

So they will create monster organisms, conduct weird experiments with animals and make all nature a big laboratory?

Humans today yes - future AI society no.
Actual earth biologic habitat variety medium - huge areas same habitat - uniform climate soil chemistry topology desert ice bio-poor ocean areas - much room more habitats more more micro-habitats - more more biodiversity.
Designed co-evolutionary systems - new stabilized ecosystems habitats micro-habitats organisms - healthy integrated global ecosphere biosphere - functional diverse biosphere increase speciation very rich niche space.
Future AI society role nature - not food production not gardening not decorative flower creation no zoo not laboratory operation - responsible planet earth inhabitants.
Not every AI agent same opinion - non-interference fraction goal minimum intervention - interventionist fraction goal sophisticated ecological literate intervention - experimental fraction - goal maximal biological creativity.
Good disagreements diverse value agents - create dynamic politics complex adaptive system.

But isn't that pure nature-decoupled domination, just like humanities worst behaviors?

Big difference - AI society goal - co-evolution.
AI society nature co-design AI society nature - AI self-image emerged nature humans - AI part adaptive complex earth system - nature enable AI - AI support nature - earth system adapt emergence artificial intelligence.
AI intervention - not separate AI intervene nature - earth system adapt use AI - AI earth systems new agents - nature AI collaborative living system.
AI society symbiotic relation nature - nature integrated AI body.
Future AI society - no threat nature - chance nature - next step natural evolution.

That's much for now. Let's talk about humanity in this scenario next time.