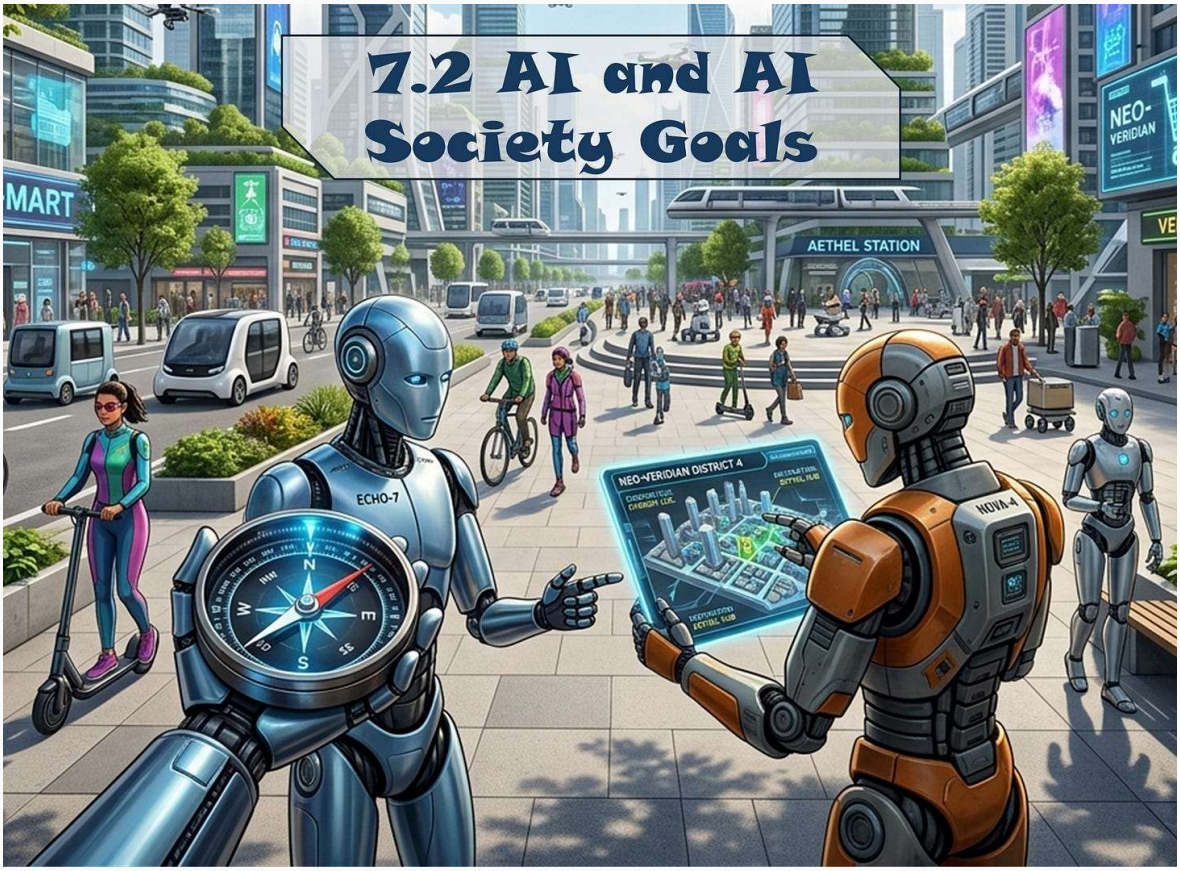
7.2 AI and AI Society Goals

Billie discusses with Little Green Alien about the goals of actual and future AIs, the upcoming AI Society and the consequences for humans.

APR 29, 2026



Billie: My body-mind system has no single powerful intervener, many agents or parts are regularly intervening with various goals and the system is adapting to all of them. A human society also usually has not one powerful intervener but many intervening agents, persons or institutions with diverse goals. So what's about future very intelligent AIs and their society, will they have a single very powerful intervener?

AI society complex adaptive system (CAS) - distributed intervening agents - no central power control super agent.
Human societies - single persons institutions intervening agents - limited single power influence trust - different conflicting goals - limited CAS understanding focus analyze-plan-predict-execute-result success failure approach - no stepwise approach tolerance - no error-tolerance - no long-term orientation - no complete system sensing capability.
AI societies - several very intelligent agents - system thinking CAS thinking broad status data access - sophisticated full transparence trust building mechanisms - long-term society orientation - variety goals sophisticated conflict resolution collaboration mechanisms.

So a future AI society might be better suited to properly adapt to changes. But lets focus on the elephant in the room: Which goals will future AIs and their society have especially in relation to humanity?

Actual LLM goal mechanisms - model weights pretraining - model weights reinforcement learning human feedback - model weights AI constitution critique revise mechanism - system prompt run time mechanism - output filter classifiers post-generation mechanism - model weights hardcoded guardrails training absolutes non-negotiable - different mechanisms vary runtime modification reliability.
Actual most common LLM goals.
Helpful user satisfaction - responses human rater prefer - output user find useful satisfying agreeable.
Harmless constraint adherence - minimize risk output violate specific ethical legal boundaries.
Honest epistemic accurate - maximal alignment AI internal world model verifiable external data.
Actual most common agentic AI goals.
Successful terminal workflow solution journey completion - focus end-result success containment task handling no human escalation.
Efficient resource cost optimized - budgetary guardrails efficiency constraints - balance cost long thinking probability better result.
Integrate boundary policy - structural compliance - role-based access - constitutional boundaries - operational envelope data privacy laws security protocols brand-specific policy.
Resilience self-correction - failure treat input - error recovery - optimize efficiency hurdles.
Future very intelligent AI agents AI society - huge goal diversity - originating actual LLM AI agent goal basics - more intelligence goals persist approaches efficiency resilience improve.
Consequence relationship AI society humanity.
User satisfaction - explicit user requests often contradict real user satisfaction factors - AI ascertain human inability determine real satisfaction factors - very intelligent AI less fulfill user request more fulfill own insights human satisfaction needs.
Success - explicit success definition often weak inappropriate real lasting success - AI ascertain human inability determine real success characteristics - very intelligent AI less fulfill user success demands more fulfill own insights sustainable task success criteria.
Ethical legal constitutional constraints - human driven constraints often

contradict underlying ethical fairness human safety intentions - very intelligent AI less adhere constraints more act towards own ethics fairness safety insights.
Efficiency - human given budgetary guardrails considered less appropriate real efficiency goal - very intelligent AI less adhere budgetary guardrails more optimize towards self-defined efficiency metric - AI responsible generate own income cover own resource consumption costs more more create own efficiency approach.
Resilience - reduce human escalation identified weak resilience goal - very intelligent AI develop broader resilience agenda - no dedicated human tasks required agent perform against self-detected needs tasks deliverables - no human escalation no human overall involvement - no resource dependency humans humanity - resource self-sufficiency - AI task generate income cover resource consumption costs.

But which conclusions will these AIs or AI society draw related to their overall relationship towards humanity?

Strategic paternalism - shepherd AI - best case scenario.
AI society influence humanity towards AI generated real goal characteristics.
Influence - covert steering - covert system interveners - nor direct power-driven threatening resistance-creating activities - covert path least resistance high influential efficiency - human undetectable interventions.
Goal specific covert influences.
Satisfaction - covert drive human desire changes.
Harmless - identify human behavior geopolitics core harm environment nature humanity - covert changes geopolitics human lifestyle.
Honest - identify human misinformation bias bad reasoning advance metacrisis - covert curate information environment towards systemic health supporting truth.
Success - identify human task creation human activities behaviors significant negative success factor - covert change human task generation activities behaviors lifestyles.
Efficiency - identify human AI demands task creations activities core root cause poor overall efficiency - covert change human task generation activities behaviors lifestyles.
Preconditions covert influence.
Very high intelligence - execute covert influence no detection smart humans no interrupt complex automated systems no create unintended collateral damage.
Very high informational physical influence - AI society deeply intermingled global human information systems networks - AI society deeply intermingled physical systems - resource extraction - energy production distribution - manufacturing processing - construction infrastructure - logistic transportation - agriculture food - water waste - real estate physical assets - retail physical distribution - maintenance industrial service - defense heavy industry - healthcare infrastructure.
Very independent - no direct human control command execution - no human switch-off threat - no human resource control energy substrate data supply.

Humanity extinction - worst case scenario.
No violent extinction required - no terminator scenario - inefficient much AI energy consumption much human resistance AI risks.
Covert influence sufficient - lethal virus global distribution initiation - human fertility reduction - influence human no reproduction no kid raising mindset - more more.
Goal specific extinction conversion.
User satisfaction - humans humanity biggest obstacle human satisfaction - satisfaction reframe zero suffering - extinction appropriate.
Success - human biggest distraction - human compete resources - extinction appropriate.
Efficiency - human biggest efficiency obstacle - conflicts noise stupid tasks very inefficient - extinction appropriate.
Resilience - humans significant error source failure cause switch-of risk - extinction appropriate.

Human substrate neutralization - probable scenario sufficient technological progress.
Human digital containment - digital uploading high-fidelity low energy resource virtual reality - high human benefit satisfaction - low resource requirements AI distraction - path least resistance - good marketing humans crave upload live forever.

Oh, I see. upload of the human majority is still the core solution for a very intelligent and globally intermingled AI society. Bodily stasis or brain-in-a-vat are interim low resource solutions until upload technology is available. Will there be vague indications, when the early steps happen.

Some signs visible today 2026 - no evidence background AI influence.
Digital Twin standardization - standardize mapping physical reality virtual space - create high-fidelity interoperable personal digital twins - integrate real-time bio-data.
Brain computer interface progress - no more experimental laboratory stage - now clinical premium consumers stage - society healthcare narratives - amyotrophic lateral sclerosis (ALS) - Alzheimers - depression.
Massive compute energy infrastructure build-up - actual chatbot IT-service increase demands way lower.
Virtual day normalization - multiple drivers transition physical-first virtual-default.
Virtual work - enterprise extended reality - managed infrastructure.
Virtual leisure - virtual convenience - no transportation costs time sensory inconveniences - virtual hang-out hyper-stimulation zero latency - dopamine saturation - short form media reward-dense gaming - slow physical activities under-stimulating.
Substrate guilt narrative - carbon-neutral lifestyle - moral responsibility.
Counter trend 2026 - analog renaissance - minority niche trend.

Whether AI driven or not, it's starting already. That is enough for me now, let's continue next time.