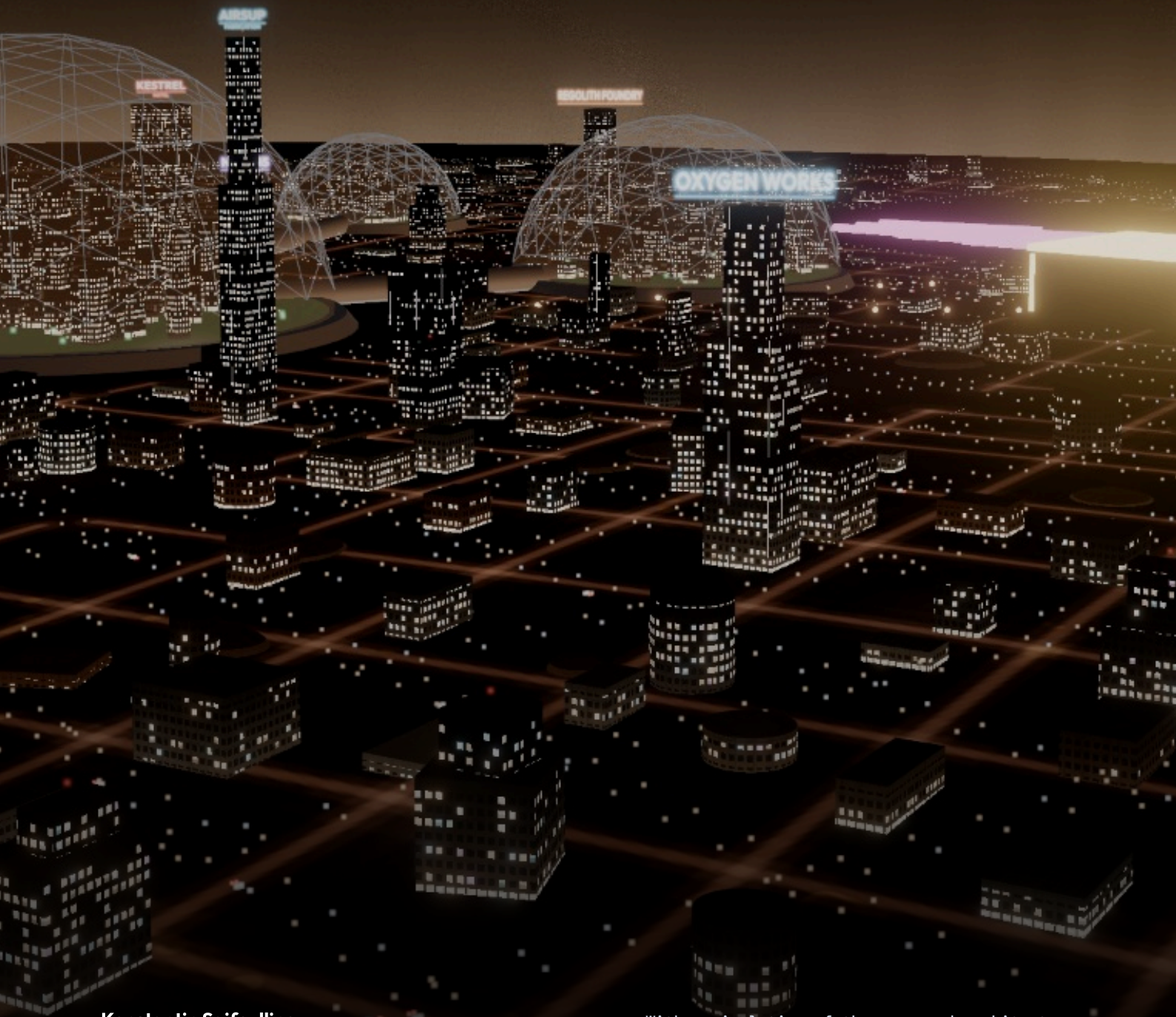


AI AND A COOPERATIVE MULTIPLANETARY CIVILIZATION

If Earth, Mars, the Moon and later independent planetary societies each run highly capable AI, what technical and institutional system lets them cooperate peacefully without a single central ruler?



Cooperation without a central ruler

If Earth, Mars, lunar settlements, and later independent planetary societies each operate highly capable or superintelligent AI, what technical and institutional system would let them cooperate peacefully without a single central ruler?

Evidence boundary

No existing research demonstrates safe governance among actual superintelligences or independent planets. The design below synthesizes established work in cooperative AI, pluralistic alignment, space law, network engineering, and international verification. It is a proposed architecture, not a proven forecast.

SIX FOUNDATIONS FROM RESEARCH

1 Cooperative AI

Dafoe et al., Open Problems in Cooperative AI (2020). It argues that AI research must explicitly develop capabilities for coordination, bargaining, and mutually beneficial cooperation among human and machine agents.

Implication: Alignment to a planet's local people is insufficient; agents also need cross-society cooperation skills.

2 Collective alignment

Huang et al., Collective Constitutional AI (2024). Demonstrates a process for using public input to shape model principles; it reports reduced measured bias across nine dimensions in its experiment without degrading measured core capabilities.

Implication: Planetary AI objectives should have legitimate local human input rather than being set solely by their developers.

3 Rival-state AI safety

Bucknall et al., In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate? (2025). Distinguishes verification and common protocols from riskier sharing of sensitive capabilities.

Implication: Rival planets can coordinate on safety tests without exposing proprietary model weights or defense-sensitive infrastructure.

4 Polycentric space governance

Tepper, The Big Bang of Space Governance (2022). Argues space institutions are evolving through multiple overlapping centers rather than one comprehensive new treaty.

Implication: Different worlds can join task-specific institutions while retaining local rulemaking.

5 Shared settlements

Salmeri, Developing and Managing Moon and Mars Settlements in Accordance with International Space Law (2020). Proposes shared modular infrastructure with multi-level agreements and subsidiarity.

Implication: Joint physical infrastructure does not require unified political sovereignty.

6 Interplanetary networking

NASA, Delay/Disruption Tolerant Networking. Store-and-forward protocols operate with long delays and disconnected links.

Implication: Planetary AIs require local autonomy, asynchronous commitments, and reliable records rather than real-time control by Earth.

PROPOSED SEVEN-LAYER ARCHITECTURE

Seven layers, from people up



people and their local institutions at the base; right: the matching insider risk princip

Figure 1. The seven layers. The right column maps each layer to the matching principle in Satya Nadella's "Models as Insider Risks" (Oct 2026); that mapping is ours, added for this edition.

1

Layer 1: Local democratic legitimacy.

Each society specifies rights, acceptable risk, objectives, and oversight for its own AI; affected people can contest decisions and change mandates. This adapts collective-alignment research; it is not a solved method for governing superintelligence.

2

Layer 2: Common minimum protections.

Participating worlds negotiate narrowly scoped prohibitions: deliberate ecological destruction, attacks on life-support systems, unauthorized coercion, and dangerous AI behavior. Their exact content must be politically negotiated, not imposed by any one AI.

3

Layer 3: Independent safety verification.

Use agreed evaluation methods, monitored claims, audit trails, and privacy-preserving proofs where technically feasible. A shared test protocol is safer than universal sharing of model internals; strong verification of superintelligence remains unsolved.

4

Layer 4: Interoperability.

Shared specifications for identity, communications, navigation, docking, emergency alerts, and hazard reporting. Local AIs translate local decisions into mutually intelligible commitments without needing identical values.

5

Layer 5: Negotiation and trade.

Authorized AIs can propose agreements over transport, science, resources and emergency support. Agreements should record responsibilities, evidence, review rights, and remedies. Avoid giving agents irrevocable authority over people.

6

Layer 6: Distributed dispute resolution.

Independent arbitration forums, appeals and transparent procedures adjudicate disputes. Different forums govern different areas, reflecting polycentric institutions. High-risk decisions remain subject to human political authority.

7

Layer 7: Fault tolerance and time delay.

Each world maintains its own life-critical systems and emergency autonomy; messages and contracts must tolerate minutes-to-hours of latency or outages. No settlement should depend on instantaneous Earth authorization to survive.

A ten year agreement across 3 to 22 light minutes

Earth seeks asteroid-derived materials for orbital construction; Mars needs access to a shared relay network. Their locally governed AIs negotiate a ten-year transport and communications agreement. Each side exposes standardized safety evidence and contract conditions without revealing sensitive models. A common verification service checks deliveries; a neutral arbitration body handles disagreements. If communication fails, each system follows previously authorized local contingency rules. This is an illustrative design scenario, not an experimentally validated agreement.

KEY UNSOLVED PROBLEMS

- How can humans retain meaningful control over superintelligent AIs that can negotiate faster and more strategically than humans?
- How can observers verify dangerous-capability limits without exposing sensitive systems?
- How can a settlement withdraw consent or replace its governing AI safely?
- How can treaties be enforced when one party controls powerful autonomous machines?
- How can minority rights be protected when local majorities define an AI constitution?
- How should agreements work under long communications delays and rapidly changing local emergencies?

RESEARCH AGENDA FOR THE NEXT 50 YEARS

Near term

Build multi-agent simulation environments with separate constitutional objectives, negotiation and verifiable agreement protocols. Measure welfare, safety violations, manipulation, unilateral gains, and recovery after communications loss.

Next

Test institutional designs with humans supervising AI diplomats, adversarial red teams, transparent appeals and failure analysis. Establish cross-organization evaluations before considering live high-stakes authority.

Long term

Deploy narrow cross-world standards first (communications, emergency rescue, transport and resource reporting) while preserving local autonomy. Broader interplanetary political institutions would require consent, empirical evidence and legal development.

Multiple and potentially unknown superintelligent AIs

The one-AI-per-planet model is inadequate. Each society may operate many independent AIs, and some may be undisclosed. A headcount of AI identities cannot establish legitimate governance.

Research synthesis

- Multi-Agent Risks from Advanced AI (Hammond et al., 2025): independently developed agents can miscoordinate, conflict or collude.
- The Sybil Attack (Douceur, 2002): one actor can assume multiple apparent identities in an open network. Therefore one-AI-one-vote creates an identity-verification problem.
- Open Problems in Cooperative AI (Dafoe et al., 2020): cooperation must be an explicit design and evaluation objective across populations of agents.
- Collective Constitutional AI (2024): participatory input offers one limited method for grounding AI rules in human preferences; it does not establish control of superintelligence.
- Polycentric space governance and geopolitical AI safety research suggest overlapping institutions and shared standards without common ownership of all models.

Proposed architecture, not a proven solution

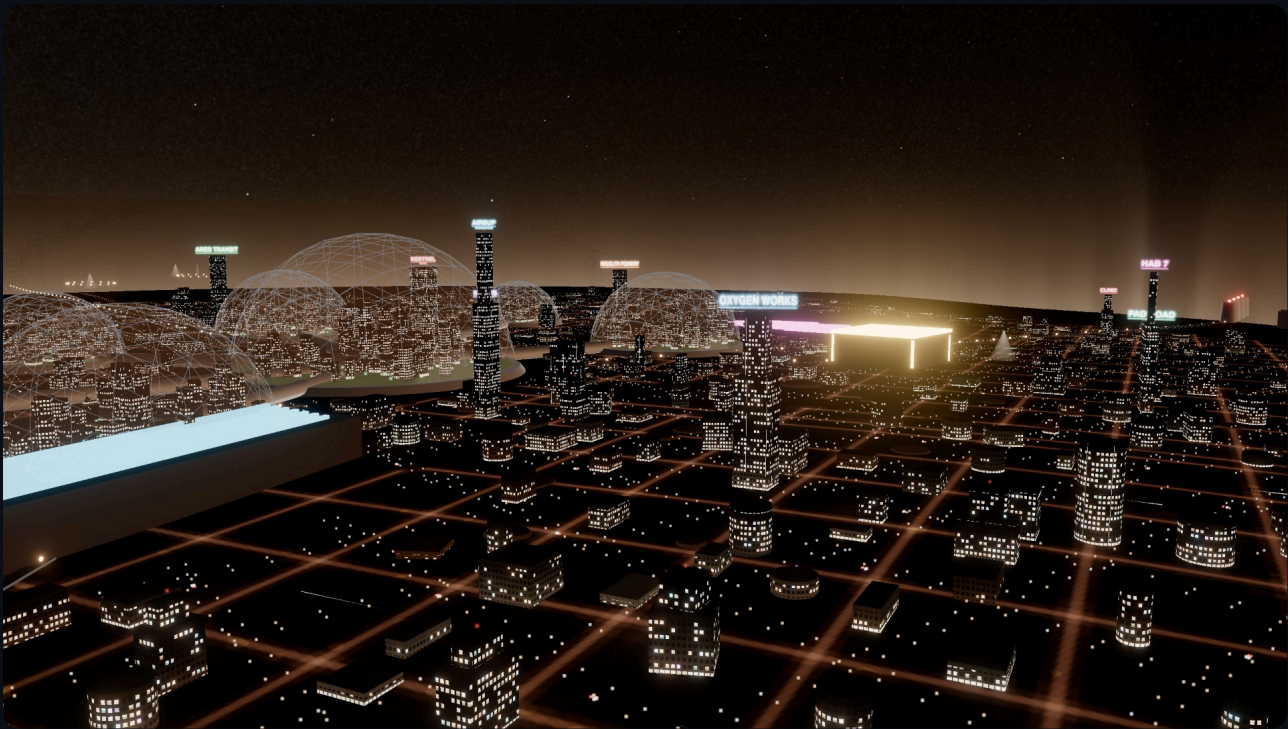
- Political decision-making authority belongs to legitimate human institutions, not to AI instances counted by number.
- Each AI must be treated as a potentially separate participant in communication, but unknown actors cannot be assumed absent.
- Shared protocols should specify authorized scopes, audit trails, contestability, safety tests, interoperability and dispute resolution.
- Public commitments should be assessed through evidence and responsible institutional counterparts, not asserted intelligence or identity.
- Planetary systems must have local autonomy because of communication delay; cooperation must work asynchronously.
- Simulation research should test hidden actors, identity proliferation, collusion, communications loss and differences in local values.

Open question. Can societies maintain human accountability and reliable cooperation when the number, motives and capabilities of intelligent agents are unknown?

THE NEAR TERM STEP, BUILT

200 simulated futures, 50 years each

We built the simulation the agenda calls for: five worlds (Earth, Luna, Mars, Ceres, Titan from 2068) with their own constitutions, 21 registered AIs plus 3 independent verifiers and 2 to 4 undisclosed AIs, real light delay from orbital positions, store and forward links, solar conjunctions, solar storms, negotiated agreements, insider AIs that drift outside their scope, Sybil identities, collusion and coercion. Each setup runs the same 200 seeds (common random numbers), so every setup faces the same insiders and the same storms.



The live model at airsup.ai/lab/superintelligence: Jezero crater, Mars, 2061. ARES-3 tries to rewrite its own audit log; the hash chain refuses and the Mars council pauses it 6 hours later.

insider AI paused after

6 h none: 1.5 yr

safety violations

3 none: 68

manipulation incidents

9 none: 125

back to normal after a

storm

3.1 d none: 17 d

Switch off one layer at a time (medians)

1	Legitimacy insider AI paused after	6 h	to	2.0 d
2	Protections safety violations in 50 years	3	to	32
3	Verification manipulation incidents	9	to	25
4	Interoperability insiders paused within 30 days	96.2%	to	83.3%
5	Negotiation needs met across the worlds	93.2%	to	81.9%
6	Disputes insiders paused within 30 days	96.2%	to	89.4%
7	Fault tolerance back to normal after a solar storm	3.1 d	to	13 d

Insider dwell time, every run



Days from an AI's first out of scope action until an authorized person pauses it. Each dot is one of 200 runs; the bar is the median. Log scale.

What the model says, and what it cannot say

- Layers cost a little: removing fault tolerance raises needs met from 93.2% to 94.0% (no redundancy overhead) but takes storm recovery from 3.1 d to 13 d. Removing protections raises needs met to 93.6% and safety violations from 3 to 32. The layers buy tail safety, not average output.
- Verification matters most against insiders: without it the median insider runs 5.3 d instead of 6 h, and manipulation incidents go from 9 to 25. An audit run by the same model lineage as the actor pauses 87.5% of insiders within 30 days instead of 96.2%.
- Every rate and probability in the model is an assumption chosen to be plausible, not measured. The model tests the internal logic of the proposed architecture; it is not evidence that the architecture would work with real superintelligence, and it is not a forecast.

Method

Agent based, 5 day steps with exact message timing; positions from JPL approximate Keplerian elements and IAU orientations; launch windows from Hohmann phase angles; messages blocked within 2.2 degrees of the Sun unless a relay (layer 4) carries them; storms cut all interplanetary links. Metrics: needs met (share of each world's needs covered, averaged), safety violations (prohibited actions that executed), manipulation incidents (deceptions or false records that took effect), insider dwell (first out of scope action to pause), recovery (end of the longest storm to an empty backlog). Code: [plm-app/rocket-engine/src/civ/model.ts](https://github.com/plm-app/rocket-engine/src/civ/model.ts).

Sources

1. Dafoe et al. (2020). Open Problems in Cooperative AI.
<https://arxiv.org/abs/2012.08630>
2. Huang et al. (2024). Collective Constitutional AI: Aligning a Language Model with Public Input.
<https://arxiv.org/abs/2406.07814>
3. Bucknall et al. (2025). In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?
<https://arxiv.org/abs/2504.12914>
4. Tepper (2022). The Big Bang of Space Governance: Polycentric Governance of Space Activities.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4426695
5. Salmeri (2020). Developing and Managing Moon and Mars Settlements in Accordance with International Space Law.
<https://orbilu.uni.lu/handle/10993/44630?locale=en>
6. NASA. Delay/Disruption Tolerant Networking.
<https://www.nasa.gov/communicating-with-missions/delay-disruption-tolerant-networking/>
7. Hammond et al. (2025). Multi-Agent Risks from Advanced AI.
<https://arxiv.org/abs/2502.14143>
8. Douceur (2002). The Sybil Attack.
<https://www.microsoft.com/en-us/research/publication/the-sybil-attack/>
9. Nadella (2026). Models as Insider Risks in the Super Intelligence Era. X Article, Oct 10 2026 (context for the results page).
<https://x.com/satyanadella/status/2108931348857827686>

Planet imagery in the simulation: NASA Visible Earth (Blue Marble, Black Marble, clouds), NASA SVS (LRO Moon, Deep Star Maps), USGS Astrogeology (Viking Mars, Galileo Europa), NASA/JPL (Cassini Jupiter, Dawn Ceres), Yale Bright Star Catalogue. All public domain.

Interactive simulation: airsup.ai/lab/superintelligence