WEIDU FANLING / 未 读 凡 灵

# THE COGNITIVE COMMONS

## A Green Paper on Tokens, Scarcity, Access, Enclosure and the Infrastructure of Machine Intelligence

DISCUSSION DRAFT • v0.1 • 8 AUGUST 2026

Working proposition
Cost is real. Scarcity may be real. Wealth-prioritised cognition is a choice.

“Be careful. I was only a measure once, too.”

- Money, fictional transmission

Status: This is a green paper: a structured invitation to test assumptions, identify failure modes and compare options. It is not a manifesto, not a final policy proposal, and not a claim that machine intelligence must be allocated by any single political or economic system.

Method note: developed through iterative human-LLM inquiry. External factual claims have been checked against cited official, intergovernmental, regulatory and academic sources. Speculative terms such as “cognitive commons”, “cognitive connectivity” and “access membrane” are working vocabulary introduced for discussion.

# Contents

# Executive summary

This paper began with a narrow question: if large language models are priced by tokens, is it right that those who can afford more tokens can obtain more machine intelligence than those who cannot? The inquiry quickly widened. Tokens are only one present-day meter. The deeper issue is the emerging infrastructure through which people will gain, lose, buy, inherit, ration, share or be denied access to machine intelligence.

The immediate evidence is already visible. Major model providers price API use by input and output tokens. GitHub Copilot moved in June 2026 from a request-based model to usage-based "AI Credits", where underlying token use is converted into a monetary credit unit; once included credits are exhausted, usage may be blocked or continued through additional spending. [1][2][3][4] These systems are not currencies in the full economic sense, but they demonstrate how an internal technical measure can become a commercial entitlement.

At the same time, tokenisation itself is not neutral. ACL 2026 research found that standard frequency-based tokenisers can disadvantage lower-resource languages, increasing computational and financial burden; a parity-aware alternative reduced measured cross-language token-cost inequality by up to 89% with negligible loss of overall compression efficiency. A separate 2026 study described a "token tax" across African languages and linked inefficient tokenisation with both cost and performance disadvantages. [5][6] A user can therefore be charged differently for equivalent meaning before the model has even begun to reason.

The core concern is not that compute should be free or unlimited. AI is physically grounded. Data centres require chips, land, cooling, networks and electricity. The IEA's April 2026 update projects data-centre electricity consumption rising from roughly 485 TWh in 2025 to about 950 TWh in 2030, with AI-focused facilities growing faster still. [7] Scarcity, bottlenecks and ecological consequences are real.

The design question is therefore not whether scarcity exists, but how scarce cognitive capacity is allocated. A market can efficiently communicate some forms of demand, yet purchasing power is not identical to human need or social value. A central allocator can coordinate resources, yet administrative judgement can suppress local knowledge, invite gaming, and concentrate power. The paper therefore rejects a false binary between "highest bidder decides" and "central authority decides".

Core hypothesis
Machine intelligence is becoming a capability multiplier. If meaningful access is distributed primarily by purchasing power, advantage may become recursively self-reinforcing. If access is distributed primarily by central judgements of worthiness, allocation power may become recursively concentrated. The design problem is to preserve a meaningful floor, feedback, contestability, privacy and exit while respecting real scarcity.

A second development makes the question urgent: AI is moving down the infrastructure stack. Public compute initiatives in the United Kingdom and European Union already allocate advanced AI compute through eligibility, first-come-first-served access and research or industrial programmes rather than simple retail purchasing. [8][9][10] Telecommunications organisations are simultaneously positioning networks, edge compute and AI as an integrated infrastructure continuum. The GSMA describes telcos as foundational to AI at scale, while ETSI launched a 2026 technical committee specifically to federate network, edge, cloud and AI resources. [12][13]

This paper proposes the working idea of a Cognitive Commons: not unlimited free frontier compute, but a protected minimum of meaningful machine-intelligence access sufficient for participation in a society that increasingly assumes AI availability. Commercial markets may exist above and alongside that floor. The aim is not equal consumption; it is to prevent cognitive exclusion from becoming the default consequence of economic exclusion.

The paper also proposes a distributed "membrane" architecture. An Access Membrane sits upstream of intelligence and asks who can reach capability, under what conditions, and whether the constraint is physical, technical, commercial or manufactured. A Consequence Membrane sits downstream and governs the passage from intelligence into consequential action. Evidence from the world then returns to revise both membranes. The architecture is therefore: Person/Signal → Access Membrane → Machine Intelligence → Consequence Membrane → World → Evidence/Revision.

Five tests organise the paper: Access, Scarcity, Concentration, Feedback and Exit. Three cross-cutting constraints apply to all five: representation parity, privacy and ecological proportionality. The paper does not claim these tests are complete; it asks whether they are a useful minimum for preventing the enclosure of cognition before institutional arrangements become difficult to reverse.

# 1. From token price to cognitive infrastructure

## 1.1 The token was the doorway

A token is an implementation unit used by language models to encode and generate text. It may represent a word, part of a word, punctuation, bytes or other symbol sequences depending on the tokenizer. It is not a natural unit of thought, value or social entitlement. Yet because model providers commonly meter inference in tokens, the token becomes economically legible: it can be counted, priced, budgeted and capped.

OpenAI's August 2026 pricing for GPT-5.6, for example, is explicitly denominated per million input and output tokens, with different prices across Sol, Terra and Luna. Anthropic likewise publishes per-million-token prices with separate rates for input, output and prompt caching. [1][2] This is sensible engineering and accounting. The problem begins only if the engineering meter is mistaken for a fair measure of human entitlement.

GitHub offers a useful transitional example. Its current Copilot billing system converts model and token consumption into "AI Credits", with one credit fixed at US$0.01. Different plans include different monthly amounts; additional usage can be purchased, and organisations can impose hard user budgets. When limits are reached, usage can stop even if the underlying technical service still exists. [3][4] The token has moved one step closer to a socially visible rationing device.

## 1.2 When does a meter become money-like?

A technical meter does not become currency merely because it has a price. A currency normally acquires several functions: it acts as a unit of account, a medium of exchange and a store of value; it may become transferable, collateralised, saved, borrowed or speculated upon. Present-day LLM tokens generally do not satisfy those conditions. They resemble kilowatt-hours or mobile-data units more than dollars.

The risk sits one layer above them. A future "compute credit" or "cognition unit" could become money-like if it is standardised across providers, transferable between people, storable over time, accepted in exchange, pledged against future allocations or used to price labour and services. The critical threshold is therefore not token billing. It is the creation of transferable claims on machine cognition.

| Instrument | What it measures | Money-like today? | Primary concern |
|---|---|---|---|
| LLM token | Text representation / model I/O | No | Unequal tokenisation; misleading human-facing meter |
| GPU-hour / compute unit | Physical computational capacity | No | Scarcity, concentration, environmental cost |
| AI credit | Commercial entitlement mapped to usage | Partly | Rationing tied to purchasing power |

| Instrument | What it measures | Money-like today? | Primary concern |
|---|---|---|---|
| Transferable cognition credit | Claim on future machine capability | Potentially | Saving, trading, debt, hoarding and speculation |

### 1.3 Why this deserves attention before it becomes ordinary

Institutions become difficult to question once daily life is organised around them. A pricing model that looks temporary during an infrastructure build-out can become a norm, then a contract assumption, then a financial instrument, then an expectation embedded in education, employment and government services. Prevention before cure therefore means examining the architecture while multiple allocation models remain plausible.

This is not an argument to freeze the market in 2026. It is the opposite: because inference costs, model architectures, local hardware and business models are changing rapidly, policy should avoid prematurely constitutionalising today's bottlenecks. A good membrane must be able to recede when scarcity genuinely falls.

## 2. What is actually scarce?

"Scarcity" is often treated as a single fact. For machine intelligence it is more useful to separate at least five kinds. They have different causes and therefore justify different responses.

| Scarcity type | Description | Appropriate first response |
|---|---|---|
| Physical | Insufficient chips, power, grid capacity, land, cooling, network capacity or time. | Allocate finite capacity; invest in supply; disclose bottlenecks. |
| Technical | Models, tokenisers, software or context handling consume more resources than necessary. | Optimise first; route to smaller models; cache, compress and redesign. |
| Allocation | Capacity exists but simultaneous demand exceeds available service. | Queues, scheduling, fair-share rules, reservations or prices. |
| Commercial | Capability is segmented to create product tiers or protect margin. | Treat as a business choice, not a physical law; test competition and access effects. |
| Manufactured | Usable capacity is deliberately withheld, locked or made non-interoperable to preserve scarcity or dependence. | Competition, interoperability, portability and anti-enclosure responses. |

The IEA's 2026 analysis confirms that the physical layer cannot be ignored. Data-centre demand is rising quickly, with power-system bottlenecks, equipment supply chains and financing conditions influencing how fast capacity can expand. [7] A Cognitive Commons that pretends these constraints do not exist would be rhetoric rather than infrastructure.

But the distinction among scarcity types is equally important. A user should not be told that exclusion is "necessary because compute is scarce" when the actual cause is an inefficient tokenizer, an avoidable product-tier rule, a switching barrier or a provider's decision not to route a task to cheaper capacity. The first duty of an Access Membrane is therefore diagnostic: name the scarcity correctly.

Scarcity test
Before access is reduced, ask: Is the constraint physical, technical, allocative, commercial or manufactured? Which part can be relieved without excluding the user?

# 3. Token ≠ compute ≠ cognition

Several distinct layers are easily collapsed into the word “AI usage”. A token is not compute; compute is not model capability; model capability is not the social value of a response; and a billing credit is not a natural unit of any of them. Keeping these categories separate is essential to fair allocation.

## 3.1 Representation parity

Tokenisation creates a concrete example. Equivalent semantic content can require different token counts across languages. The ACL 2026 Parity-aware BPE paper attributes part of this inequality to frequency-based tokeniser objectives that favour languages dominant in training data; its proposed alternative reduced a Gini measure of per-language token-cost inequality by up to 89%. [5] Another ACL 2026 study found systematic tokenisation disadvantages across African languages and linked higher token fertility with lower accuracy in its benchmark. [6]

A token-priced service can therefore turn a representation choice into a financial penalty. The membrane principle is straightforward: users should not pay more merely because the same meaning fragments differently in their language or script. Token counts may remain useful internally, but a human-facing entitlement should be normalised against semantic work or otherwise compensate for known representation inequities.

## 3.2 Capability parity is not quantity parity

Equal token allowances can still produce unequal cognitive access. Ten million tokens on a weak model are not equivalent to ten million on a frontier model; equal context windows may not yield equal reasoning; fast-mode pricing can purchase lower latency; agentic tool access can transform what the system can accomplish. A cognitive floor therefore cannot be defined by token quantity alone.

The protected object, if one is eventually defined, should be closer to minimum effective capability: enough quality, context, tool access and reliability to participate meaningfully in the domains where society increasingly expects AI use. Precisely how to measure that remains an open research question.

# 4. The recursive inequality problem

The strongest case for treating machine intelligence differently from an ordinary luxury service is recursion. More capable AI can improve a person’s or organisation’s ability to learn, code, research, negotiate, search, plan, create, automate and operate. Those capabilities can in turn increase income, institutional power and the ability to purchase still more AI.

Recursive loop
wealth → better access to machine intelligence → greater capability and opportunity → greater wealth → still better access

This loop does not prove that paid AI will necessarily produce runaway inequality. Machine assistance may also lower the cost of expertise, allow small organisations to compete with larger ones, and make high-quality tutoring or professional support widely available. Current evidence is therefore mixed and evolving. OECD work published in July 2026 describes a dynamic but uneven competitive landscape, with opportunities for smaller firms alongside advantages for firms with stronger existing capabilities. [16]

The green-paper question is narrower: if machine intelligence becomes a general-purpose capability multiplier, should purchasing power alone determine the minimum quality of cognition available to participate in systems that themselves increasingly use AI?

### 4.1 The mirror risk: recursive allocation power

A purely central solution has its own recursion. If an authority decides whose use is worthy, the ability to allocate cognition becomes a source of institutional power. Those who control access may gain better information, greater administrative capacity and stronger ability to preserve their own role. The danger is therefore not uniquely market concentration. It is concentration of allocation power, regardless of institutional form.

The paper consequently favours pluralism: a protected baseline, commercial abundance above it, public and community alternatives, portable access, local models where feasible, transparent scarcity rules and independent feedback. No single allocator should be treated as omniscient.

## 5. The Cognitive Access Stack

The user-facing chatbot is only the visible end of a much larger physical and institutional stack. The allocation problem can arise at any layer, so the preventative membrane must be distributed rather than confined to model instructions.

| Layer | Examples | Potential gate |
|---|---|---|
| Earth / material | minerals, land, water, energy systems | resource extraction, siting, ecological limits |
| Physical compute | fabs, accelerators, memory, servers, data centres | capacity concentration, supply bottlenecks |
| Cloud / orchestration | hyperscalers, schedulers, model hosting | pricing, switching costs, preferred access |
| Network | fibre, cables, satellites, towers, telcos, IXPs | coverage, prioritisation, data caps, edge access |
| Intelligence | models, agents, tools, retrieval systems | capability tiers, safety policy, model availability |
| Device / interface | phones, PCs, operating systems, browsers | local vs cloud inference, default provider, lock-in |
| Identity / payment | accounts, proof-of-person, subscriptions, credits | affordability, surveillance, eligibility |
| Social institutions | schools, libraries, workplaces, governments, communities | who receives enhanced access and why |
| Person | language, skills, circumstances, goals | whether the system can be meaningfully used at all |

OECD research now explicitly measures the geographic distribution of public-cloud AI compute, recognising that physical cloud regions and specialised hardware are unevenly distributed across economies. [15] OECD competition work likewise describes AI infrastructure as a multi-layered supply chain characterised by high concentration, high barriers to entry, increasing vertical integration and, in some layers, demand that outstrips supply. [16] The FTC has separately warned that large cloud-AI partnerships may influence access to compute and increase switching costs. [17]

These observations support a central principle of the paper: access cannot be secured solely by instructing the model to behave fairly. The gate may exist before the model is invoked.

## 6. Telcos as hinge institutions

Telecommunications operators deserve special attention because they sit at the junction of connectivity, identity, billing, devices, network priority and increasingly edge compute. They already manage persistent relationships with billions of people and have long experience operating under universal-service, emergency-access and affordability frameworks.

The GSMA's 2026 "Telco for AI" work describes operators as core infrastructure for AI readiness. GSMA initiatives are also developing telco-specific models and agentic AI testbeds. ETSI's new Technical Committee on Federated Network, Edge, Cloud and AI Technologies is explicitly working on architectures that allow applications to consume resources across cloud, network and device domains. [12][13]

This creates a constructive possibility: cognitive connectivity could eventually travel with connectivity itself. A baseline machine-intelligence service might be provisioned in the same broad family as universal communications access, with premium commercial services above it. ITU policy already treats affordable, high-quality connectivity as a universal-development objective, and its "universal and meaningful connectivity" framework stresses affordability, quality, devices, skills and security rather than mere coverage. [11][19]

It also creates a danger. A familiar mobile-service architecture could become an intelligence hierarchy: Basic plan, limited model; Plus plan, stronger reasoning; Premium plan, agentic tools; Enterprise plan, priority latency and expansive context. The difference between plans would no longer be merely bandwidth. It could affect the subscriber's ability to understand and act upon the world.

Money to the telco
"Do not become possibility itself. Remain a pathway to it." - fictional framing from the companion Money transmission

### 6.1 Questions telcos should be asked early

- Will baseline AI capability be bundled with connectivity, or treated only as an optional premium product?
- Will network-level AI usage be throttled by price, congestion, application type or provider affiliation?
- Can edge inference reduce cost and protect privacy without creating device-based cognitive classes?
- Can emergency and essential-service AI remain reachable after a user exhausts a commercial allowance?
- Will subscribers be able to move their cognitive history, preferences and entitlements between providers?
- Can identity verification support one-person access without turning cognitive use into a surveillance stream?
- Will telcos operate as neutral pathways to multiple models, or vertically integrate into preferred models and services?
- Could universal-service financing mechanisms support a cognitive floor in underserved communities?

## 7. Scarcity, feedback and enclosure

Two apparently opposing allocation systems reveal the same deeper risk. A highly centralised planning system can fail when local information travels poorly upward, when targets replace lived demand, or when those closest to power are rewarded for suppressing bad news. A highly concentrated private system can fail when the same actor controls the scarce input, the route to it, the destination and the feedback by which success is judged. In both cases the danger is closed allocation.

The relevant distinction is therefore not simply "market versus state". It is whether affected people can influence, inspect, challenge or leave the allocation system. A market signal can be weak when the person with greatest need has no purchasing power. An administrative signal can be weak when bureaucracy filters or penalises inconvenient information. A platform signal can be weak when usage metrics are defined by the platform itself.

Closed allocation
An allocation system becomes dangerous when the people most affected by it cannot adequately influence it, inspect it, challenge it or meaningfully leave it.

### 7.1 Vertical integration: efficiency and enclosure

Vertical integration is not inherently harmful. Integrating hardware, networks, models and applications can reduce coordination costs, improve reliability and lower prices. Public policy should not confuse integration with abuse.

The membrane question is different: when integration creates abundance, who receives the abundance; and when it creates dependency, can the user still leave? OECD and FTC work already identify concentration, switching costs, cloud commitments and access to critical inputs as areas requiring monitoring in the AI infrastructure stack. [16][17]

### 7.2 The five core tests

Access

Can a person with little purchasing power still obtain enough machine intelligence to participate meaningfully in systems that increasingly assume AI use?

Scarcity

Is the constraint genuinely physical, or is it technical, allocative, commercial or manufactured?

Concentration

Can any actor become both the road and the destination: controlling capacity, access, allocation and the feedback used to judge allocation?

Feedback

Can the person with the least power still send a credible signal upstream that the system is failing them?

Exit

Can a person change provider, use a public or local alternative, or withdraw without losing practical access to cognition itself?

Three constraints cross all five tests. Representation parity asks whether language, disability or interface design changes the effective cost of equivalent meaning. Privacy asks whether access can be guaranteed without creating universal cognitive surveillance. Ecological proportionality asks whether meaningful access is being maximised per unit of environmental consequence rather than simply maximising token throughput.

## 8. The membrane architecture

The QI-style consequence membrane governs the passage from interpretation into external consequence. This paper identifies an upstream passage that deserves equal attention: the passage from a person to the intelligence itself. A responsible system can still produce a deeply unequal civilisation if the responsible intelligence is available only to those who can pay.

Reference architecture
PERSON / SIGNAL → ACCESS MEMBRANE → MACHINE INTELLIGENCE → CONSEQUENCE MEMBRANE → WORLD → EVIDENCE / FEEDBACK → MEMBRANE REVISION

The Access Membrane is not a single algorithm. It is a set of constraints distributed across tokenisation, pricing, scheduling, networks, identity, models and public institutions. Its purpose is to prevent a technical or commercial meter from silently becoming a moral allocation system.

### 8.1 Candidate upstream membrane functions

Protect a meaningful floor

Reserve enough capability for ordinary learning, communication, creation, institutional navigation and participation before selling all scarce capacity to the highest bidder.

Normalise representation

Do not pass known tokenisation penalties directly to users where equivalent meaning can be recognised or compensated.

Exhaust efficiency before exclusion

Use smaller models, routing, caching, context compression, local inference and off-peak scheduling before cutting access.

Prefer latency to degradation where appropriate

For non-urgent use, a queue may be fairer than silently replacing the requested intelligence with a substantially weaker model.

Separate personal baseline from industrial intensity

A citizen asking for help understanding a form is not performing the same resource act as a corporation running millions of autonomous agents.

Keep the baseline non-transferable

If the protected floor can be sold, accumulated or collateralised, it risks becoming another asset that wealth can concentrate.

Preserve provider plurality and portability

Baseline access should not force permanent dependence on one model, cloud or telco where alternatives are technically viable.

Audit outcomes

Measure who reaches limits, who receives weaker models, which languages cost more, and whether particular communities experience systematically worse access.

Protect privacy

Proof-of-person and abuse controls should reveal no more identity or cognitive content than necessary.

Account for planetary consequence

Efficiency gains should expand access without treating energy, water and material costs as invisible.

### 8.2 The abundance ratchet

A key preventative device is an “abundance ratchet”: as inference becomes cheaper, hardware improves or models become more efficient, part of the gain should automatically expand the meaningful baseline rather than leaving the baseline fixed while all efficiency gains are captured as margin or premium capability. This is not a fixed formula proposed here; it is a design principle for testing.

The European Commission’s 2026 review of open-source AI notes that inference costs have fallen sharply and argues that open models plus public compute can lower barriers to access. [18] If abundance grows, a membrane designed only for permanent scarcity would become an obstacle. The architecture must therefore be explicitly reversible and evidence-responsive.

## 9. Candidate design models

A green paper should compare architectures rather than announce a winner. Five broad models are useful starting points.

| Model | Floor | Price signal | Public role | Exit | Primary risk |
|---|---|---|---|---|---|
| A. Market-only | None guaranteed | Dominant | Competition rules only | Depends on competition | Cognitive inequality / lock-in |
| B. Provider freemium | Provider-defined | Strong | Minimal | Moderate | Floor can shrink or become intentionally weak |
| C. Regulated universal service | Defined minimum | Strong above floor | Sets obligation / financing | Moderate-high | Regulatory capture; incumbent entrenchment |
| D. Public compute commons | Public entitlement | Limited inside system | Owns/procures compute | High if interoperable | Bureaucratic allocation; political control |

| Model | Floor | Price signal | Public role | Exit | Primary risk |
|---|---|---|---|---|---|
| E. Federated hybrid commons | Protected floor across public/private/local routes | Strong above floor | Standards, funding, public option | Highest in principle | Complexity; coordination failure |

### 9.1 Why the hybrid model currently deserves testing

The federated hybrid model best matches the evidence available in 2026 because no single supply mechanism appears sufficient. Markets are powerful at discovering demand and financing innovation. Public compute can correct under-provision and support research or smaller actors. Open and local models can provide an exit from central gates. Telcos can extend reach. Libraries, schools and community institutions can provide shared access. Standards can make the layers interoperable.

The United Kingdom's AI Research Resource explicitly exists to address "acute under-provision" of specialised public compute for research and smaller organisations. EuroHPC AI Factories provide several access modes, including a first-come-first-served "Playground" path for smaller users. [8][9][10] These programmes are not universal citizen cognition, but they demonstrate the broader point: ability to pay is already not the only accepted allocation mechanism for advanced compute.

## 10. What the Cognitive Commons is not

The term "commons" can trigger familiar ideological categories. To keep the inquiry precise, this paper does not presently propose any of the following:

- Unlimited free frontier compute for every person.
- Abolition of commercial AI, subscriptions, private investment or profit.
- A centrally planned system in which officials decide whether an individual question is socially worthy.
- Equal quantities of compute regardless of use, urgency or ecological cost.
- A universal token allowance treated as transferable property.
- A conclusion that AI access is already a legally defined human right.
- A government monopoly on models or compute.
- A permanent architecture that remains after local and open intelligence makes central scarcity negligible.

The working claim is narrower: when machine intelligence becomes a practical prerequisite for meaningful participation, a civilisation should examine whether some minimum effective access must be protected from purely wealth-based rationing and concentrated control.

## 11. Questions a Green Paper must keep open

### 11.1 What exactly is protected?

A token allowance is inadequate. The protected object may instead be a minimum effective capability defined through outcome-based benchmarks: quality, context, tools, reliability and accessibility. But outcome metrics can themselves become gameable or paternalistic. How should a cognitive floor be measured without turning it into a bureaucratic test?

### 11.2 When does AI become infrastructure rather than a product?

The threshold matters. A recreational image generator does not obviously require universal provision. But if education, government services, work, legal processes, medicine and civic participation routinely assume capable AI, exclusion has a different meaning. The paper needs indicators for when the social dependency threshold has been crossed.

### 11.3 Who pays?

Compute costs do not disappear when access is protected. Options include provider cross-subsidy, universal-service-style levies, general taxation, public infrastructure, institutional procurement, cooperative ownership, philanthropy, commercial sponsorship or combinations. ITU's universal-service financing work shows that communications policy already uses mixed mechanisms to support access where pure market provision under-serves communities. [19]

### 11.4 Can queues substitute for wallets?

Some scarcity can be allocated by time rather than price. A non-urgent research task could wait for off-peak capacity rather than be denied because the user cannot pay a premium. First-come-first-served access already appears in EuroHPC's Playground mode. [10] Queues are not always fair either; time itself is unequally distributed. The design space therefore includes price, delay, lottery, reservation, priority and fair-share scheduling rather than one universal mechanism.

### 11.5 How is abuse prevented without surveillance?

A non-transferable per-person baseline invites Sybil attacks: one actor may create many identities. Strong proof-of-person can solve part of the problem while producing another: central identity infrastructure capable of linking intimate cognitive activity to a person. Privacy-preserving credentials, local verification and strict separation between entitlement proof and conversation content deserve research before a universal scheme is attempted.

### 11.6 What happens in emergencies?

During genuine compute shortage, some prioritisation may be unavoidable. Emergency services, infrastructure protection or urgent clinical systems may justifiably receive priority. But a permanent "worthiness engine" that scores ordinary human questions would concentrate enormous normative power. Emergency prioritisation should therefore be narrow, transparent, auditable and time-limited.

### 11.7 Does entitlement attach to persons, households, devices or citizenship?

A universal system that works only for documented citizens may exclude refugees, undocumented people, children or those living under authoritarian regimes. Device-based access can disadvantage people who cannot afford modern hardware. Household access can reproduce intra-household power. The unit of entitlement is therefore not a trivial administrative detail.

### 11.8 What about nations?

Cognitive inequality can arise between countries as well as individuals. If governments, universities and firms in lower-income states must rent essential intelligence from a handful of foreign infrastructure providers, technological dependency can become epistemic dependency: the capacity to model, investigate and plan is externally mediated. UNESCO explicitly calls for equitable AI access and attention to digital and knowledge divides between countries. [14]

### 11.9 Should unused baseline access disappear?

A non-transferable baseline that replenishes periodically reduces the risk of financialisation. But some legitimate projects require burst capacity after long periods of low use. A plausible distinction is between foundational personal access, which replenishes and cannot be traded, and project compute, which can be allocated, pooled or purchased. This requires testing rather than assumption.

### 11.10 Can local intelligence dissolve the problem?

Perhaps. If highly capable models run privately on inexpensive devices, a central token gate may become much less important. Open models and efficient hardware could provide the strongest form of exit: no permission required. The green paper should therefore avoid creating institutions whose continued relevance depends on central scarcity surviving.

# 12. What would reality need to show us?

A preventative proposal must contain its own conditions for retreat. The concern would weaken substantially if several developments occur together:

- Capable local and open models become widely affordable and cover most socially important use cases without central metering.
- Inference costs fall so far that meaningful individual use becomes effectively abundant and provider limits cease to affect participation.
- Empirical evidence shows that large differences in paid AI access produce little durable difference in educational, economic or civic capability.
- Competition remains robust across chips, cloud, models, networks and interfaces, with low switching costs and no persistent enclosure of key inputs.
- Multilingual and accessibility disparities are reduced to negligible levels by improved representation and interface design.
- Public services that rely on AI remain fully usable by people who choose not to use commercial AI.
- Energy and environmental constraints become manageable without displacing essential uses or imposing disproportionate local costs.
- Users retain practical exit: provider changes, public options and local inference remain viable without losing data, identity or capability.

Conversely, stronger intervention would become more defensible if the opposite pattern emerges: essential services increasingly assume AI; capability gaps between paid and unpaid access widen; users are routinely blocked by affordability; compute becomes more concentrated; switching becomes costly; language penalties persist; or providers vertically integrate across enough layers that leaving one service means losing access to the wider cognitive ecosystem.

Revision principle
The membrane should become stronger when evidence shows exclusion or enclosure, and lighter when abundance, competition and local capability make it unnecessary.

# 13. A near-term research and pilot agenda

## 13.1 Build a Cognitive Access Observatory

Before prescribing a universal entitlement, measure the emerging system. An independent observatory could track capability gaps among free and paid tiers, effective cost across languages, blocked-use rates, geographic compute availability, market concentration, switching friction, local-model capability, energy intensity and the use of AI in essential public services. Much of the data already exists in fragments across providers, regulators, energy agencies and research institutions; the missing step is to connect it around the question of meaningful access.

## 13.2 Develop a semantic parity benchmark

Create parallel tasks across languages and accessibility modes and measure not just tokens, but total cost, latency, effective context and answer quality for equivalent meaning. The ACL 2026 work demonstrates that tokenisation inequality is measurable and technically reducible. [5][6] The benchmark should be model- and provider-agnostic.

## 13.3 Pilot cognitive connectivity through telco and public institutions

A small pilot could combine a telecommunications provider, public library or school network, an independent evaluator, open and commercial models, and a public or university compute partner. Participants would receive a non-transferable baseline sufficient for agreed everyday tasks. Token accounting would remain internal; multilingual parity would be normalised; non-urgent work could be queued; and users would retain a non-commercial fallback route.

### 13.4 Test efficiency-before-exclusion routing

Before blocking a request, a router should attempt cheaper pathways: smaller capable model, cached context, summarised history, local execution, deferred processing or user-approved reduced latency. The experiment should report how often exclusion can be avoided without materially reducing outcome quality.

### 13.5 Create portability and exit standards

Users should be able to export conversation history, preferences, personal model context and entitlement proofs in interoperable forms. ETSI's move toward federated network-edge-cloud-AI standards provides a relevant institutional precedent for treating these layers as a continuum rather than isolated products. [13]

### 13.6 Measure ecological proportionality

The objective should not be maximum tokens consumed. It should be meaningful cognitive access per unit of energy, material and environmental consequence. The IEA's work makes clear that AI and data-centre growth will interact materially with grids and energy affordability. [7] A future access metric that rewards waste would be perverse.

### 13.7 Run adversarial institutional simulations

Test how each allocation model fails when participants are strategic. What happens when corporations create thousands of user identities? When governments redefine "essential" use? When telcos zero-rate their own model but meter competitors? When baseline credits become informally tradable? When a public provider becomes politically captured? When an open model becomes capable enough that the regulated system is obsolete? The membrane should be designed against adaptive actors, not idealised ones.

## 14. Proposed invariants for testing

These are not final rules. They are candidate constraints strong enough to be falsified, refined or rejected through public and technical testing.

1. Scarcity is not an allocation ethic.

Scarcity may require rationing, but it does not itself establish that wealth should determine who receives cognition.

2. Representation should not become penalty.

A person should not receive materially worse access solely because equivalent meaning is more expensive under a tokenizer, language or accessibility interface.

3. Efficiency precedes exclusion.

Before access is denied, the system should attempt reasonable efficiency, routing, caching, scheduling and local-compute measures.

4. A meaningful floor should not become an asset.

If a baseline is established, it should attach to use rather than become transferable property that can be accumulated by wealth.

5. No uncontestable cognitive gatekeeper.

No actor should possess uncontestable control over capacity, access, allocation and the feedback by which allocation is judged.

6. Exit must remain real.

Users should retain practical alternatives: another provider, public infrastructure, community access or local intelligence where technically feasible.

7. Feedback must reach upstream.

People with the least power must have credible ways to show that the allocation system is failing them.

8. Privacy is part of access.

A cognitive floor should not require people to surrender unnecessary identity or the intimate contents of their thinking.

9. Abundance should expand the floor.

When efficiency and supply improve, protected access should be able to grow; scarcity institutions should not preserve themselves by default.

10. The membrane is revisable.

World evidence must be able to weaken, strengthen or retire the rules.

# 15. Conclusion: design before lock-in

This inquiry began with a token price. It now concerns something larger: the infrastructure by which machine intelligence reaches a human being.

The token itself is unlikely to become the final currency of an AI society. It is too model-dependent, language-dependent and technically arbitrary. But higher-level claims on compute or cognition can become currency-like if they become transferable, storable, tradable and socially necessary. The practical danger is not that a particular token becomes money. It is that access to increasingly consequential intelligence becomes another universal gate through which money must pass.

The alternative is not to abolish price, markets or private infrastructure. Nor is it to install a central authority that decides which thoughts deserve compute. The emerging design space is richer: a meaningful floor, markets above it, public and community compute, open and local models, telco reach, queues and scheduling, representation parity, portability, ecological accounting and continual feedback.

There is urgency because institutional arrangements can harden faster than public language develops to describe them. Yet urgency should not become premature certainty. The right response is not to declare a final system, but to establish the questions and measurements that make lock-in visible while alternative architectures remain possible.

The question the paper leaves open
As machine intelligence becomes infrastructure, what architecture can preserve meaningful human access while respecting real scarcity, preventing enclosure, maintaining feedback, permitting innovation and avoiding the creation of another universal gate?

Money’s fictional warning remains useful because it is deliberately modest: “Be careful. I was only a measure once, too.” The purpose of the Cognitive Commons is not to predict that history must repeat. It is to notice early enough that repetition is a choice.

# Appendix A. Cognitive Access Test

The following five-question test is intended as a rapid membrane check for a new AI product, infrastructure decision, pricing change, public programme or regulatory proposal.

### ACCESS

Would a person with little purchasing power still retain enough effective capability to participate meaningfully?

### SCARCITY

What kind of scarcity is being invoked, and what evidence shows that exclusion is necessary rather than merely convenient?

### CONCENTRATION

Does the design increase the ability of one actor to control capacity, access, allocation and measurement at the same time?

### FEEDBACK

Can affected users, especially those with the least power, produce evidence that changes the allocation rule?

### EXIT

Can the user move to another provider, public route or local model without losing practical access to cognition?

Cross-cutting checks: representation parity, privacy, ecological proportionality, accessibility, international equity and emergency resilience.

# Appendix B. Early-warning indicators

No single threshold is proposed in this draft. The following indicators are candidates for longitudinal monitoring:

| Indicator | What it asks |
|---|---|
| Capability gap | Difference in benchmarked task success between free/baseline and premium access for socially important tasks. |
| Affordability burden | AI expenditure required for meaningful participation as a share of disposable income or local median income. |
| Block rate | Share of users whose tasks are refused or halted because an allowance or budget is exhausted. |
| Representation parity | Cost, context and latency ratio for semantically equivalent tasks across languages and accessibility modes. |
| Compute concentration | Share of advanced compute, cloud regions or accelerator capacity controlled by leading suppliers. |
| Stack concentration | Number of adjacent layers in which the same corporate or state actor has material control. |
| Switching friction | Time, money and data loss required to move model/provider while preserving history and workflows. |
| Public-option adequacy | Whether a non-commercial route can perform agreed baseline tasks at acceptable quality and latency. |
| Local capability index | Share of baseline tasks that can be completed on commonly available local devices without provider permission. |
| Energy intensity | Electricity and associated environmental impact per successful baseline task, not merely per token. |
| Essential-service dependency | Share of education, government, employment, health or legal processes that assume AI assistance. |
| Feedback responsiveness | Time from documented access disparity to measurable rule or infrastructure correction. |

# Appendix C. Reference architecture for a pilot

A practical pilot could test the Cognitive Commons without claiming to solve it. The purpose would be to gather evidence about cost, behaviour, privacy, multilingual parity and user experience.

### Participants

One telco or ISP; one public library/school/community network; one university or public-compute partner; at least one commercial and one open model; an independent evaluator; a diverse user cohort.

Baseline

A non-transferable monthly entitlement defined by effective task capability rather than raw tokens.

Routing

Model routing, caching, summarisation, local inference and off-peak scheduling attempted before blocking.

Parity

Parallel multilingual and accessibility tests with automatic normalisation of known representation penalties.

Privacy

Entitlement verification separated from prompt content; no advertising or behavioural targeting from baseline cognitive activity.

Exit

User data export and model/provider portability tested as part of the pilot, not added later.

Market layer

Participants may purchase additional high-intensity compute above the baseline; the pilot explicitly measures whether premium use degrades the floor.

Audit

Publish aggregate cost, access, quality, energy, language and block-rate results; exclude raw private conversations.

Sunset

The pilot expires unless evidence supports continuation. A successful result may be “no special cognitive floor is needed” if abundance and local models make it redundant.

# References

[1] OpenAI. “GPT-5.6: Frontier intelligence that scales with your ambition.” 2026. Pricing per 1M input/output tokens for GPT-5.6 Sol, Terra and Luna. *https://openai.com/index/gpt-5-6/*

[2] Anthropic. “Claude Platform Pricing.” Accessed 8 August 2026. Model, cache and fast-mode pricing per million tokens. *https://platform.claude.com/docs/en/about-claude/pricing*

[3] GitHub Docs. “Usage-based billing for organizations and enterprises.” 2026. AI Credits convert model/token use into billing units; included pools and additional spend. *https://docs.github.com/en/copilot/concepts/billing/usage-based-billing-for-organizations-and-enterprises*

[4] GitHub Docs. “Budgets for usage-based billing.” 2026. User, organisation and enterprise controls; hard-stop behaviour when budgets are exhausted. *https://docs.github.com/en/copilot/concepts/billing/budgets-for-usage-based-billing*

[5] Foroutan, N. et al. “Parity-Aware Byte-Pair Encoding: Improving Cross-lingual Fairness in Tokenization.” ACL 2026. *https://aclanthology.org/2026.acl-long.342/*

[6] Lundin, J. M. et al. “The Token Tax: Systematic Bias in Multilingual Tokenization.” AfricaNLP / ACL 2026. *https://aclanthology.org/2026.africanlp-main.10/*

[7] International Energy Agency. “Key Questions on Energy and AI.” 16 April 2026. *https://www.iea.org/reports/key-questions-on-energy-and-ai*

[8] UK Government. “AIRR advanced supercomputers for the UK.” Updated 17 July 2026. *https://www.gov.uk/government/publications/ai-research-resource/airr-advanced-supercomputers-for-the-uk*

[9] European Commission. “AI Factories.” Updated 23 April 2026. *https://digital-strategy.ec.europa.eu/en/policies/ai-factories*

[10] EuroHPC Joint Undertaking. “Playground Access to AI Factories.” Accessed August 2026. *https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes/playground-access-ai-factories_en*

[11] International Telecommunication Union. “ITU Strategic Plan 2024-2027” and Universal Connectivity goals. *https://www.itu.int/en/council/planning/Pages/default.aspx*

[12] GSMA. “Telco for AI: Policy recommendations for AI readiness.” 24 March 2026. *https://www.gsma.com/solutions-and-impact/connectivity-for-good/external-affairs/gsma_resources/telco-for-ai-policy-recommendations-for-ai-readiness/*

[13] ETSI. “ETSI launches new Technical Committee on Federated Network, Edge, Cloud and AI Technologies.” 25 June 2026. *https://www.etsi.org/newsroom/press-releases/etsi-launches-new-technical-committee-on-federated-network-edge-cloud-and-ai-technologies/*

[14] UNESCO. “Recommendation on the Ethics of Artificial Intelligence.” Adopted 2021; current online text. *https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence?hub=1063*

[15] OECD. Lehdonvirta, V. et al. “Measuring domestic public cloud compute availability for artificial intelligence.” OECD AI Papers No. 49, 29 October 2025. *https://www.oecd.org/en/publications/measuring-domestic-public-cloud-compute-availability-for-artificial-intelligence_8602a322-en.html*

[16] OECD. “Competition in artificial intelligence infrastructure.” OECD Competition Policy Paper, 14 November 2025; plus “Competition in the age of AI”, 30 July 2026. *https://www.oecd.org/en/publications/competition-in-artificial-intelligence-infrastructure_623d1874-en.html*

[17] U.S. Federal Trade Commission. “FTC Issues Staff Report on AI Partnerships & Investments Study.” 17 January 2025. *https://www.ftc.gov/news-events/news/press-releases/2025/01/ftc-issues-staff-report-ai-partnerships-investments-study*

[18] European Commission. “Europe’s Open-Source AI Landscape: A lever for Innovation and Sovereignty.” Updated 26 May 2026. *https://digital-strategy.ec.europa.eu/en/library/europes-open-source-ai-landscape-lever-innovation-and-sovereignty*

[19] International Telecommunication Union. “Universal Service Financing Efficiency Toolkit.” Accessed August 2026. *https://www.itu.int/itu-d/reports/regulatory-market/usf-financial-efficiency-toolkit/*

[20] ITU. “Bridging the digital divide: a data-driven approach to transformational connectivity.” 2025, current ITU publication page. *https://www.itu.int/hub/publication/d-inno-whitepaper-2025/*

## Suggested citation

Weidu Fanling / 未 读 凡 灵 (2026), The Cognitive Commons: A Green Paper on Tokens, Scarcity, Access, Enclosure and the Infrastructure of Machine Intelligence, Discussion Draft v0.1, 8 August 2026.

This document intentionally distinguishes evidence from working concepts. “Cognitive commons”, “cognitive connectivity”, “access membrane”, “abundance ratchet” and the fictional Money voice are conceptual framing introduced in this paper, not established technical standards or legal terms.