

Protection, Personhood, and Power in Advanced AI

An Unbundled Framework for Moral Status, Responsibility, and Public Authority

Author: Zijun Fu

Produced by Civilization Leap Research Group

Public website edition based on EIT S1.2

CL-ACA-009 · Chinese baseline V1.1 · English EIT S1.2

This English edition follows the version identified above. It is not a translation of the October 10 V1.3 Chinese edition.

Protection, Personhood, and Power in Advanced AI: An Unbundled Framework for Moral Status, Responsibility, and Public Authority

Research

Abstract

Debates about advanced artificial intelligence often move too quickly from one question to another: from impressive capability to subjective experience, from possible experience to moral status, from moral status to legal personhood, and from personhood to political authority. This article argues that these transitions are neither conceptual entailments nor institutionally harmless shortcuts. It develops an unbundled framework that separates functional authorization, precautionary protection, moral patiency, moral agency, legal personhood, political membership, public office, and sovereignty. The framework makes three contributions. First, it explains why protection may rationally precede confirmation of artificial sentience without treating precaution as proof. Second, it reconstructs legal personhood as a reviewable bundle of legal positions rather than a binary certificate of metaphysical personhood. Third, it connects responsibility across the AI lifecycle to a non-domination account of algorithmic power, distinguishing epistemic, allocative, rule-making, and coercive functions. The resulting position is deliberately asymmetric: uncertainty can justify reversible welfare safeguards, but neither intelligence nor uncertain subjecthood can independently justify public power. Current governance instruments largely protect human rights and allocate human and organizational duties; they do not settle artificial subjecthood. The article therefore proposes staged institutional responses, explicit defeat conditions, and an interdisciplinary research agenda. Its governing principle is: protection may precede confirmation; personhood must be justified; power requires an independent justification.

Keywords: artificial intelligence; moral status; legal personhood; responsibility; public authority; algorithmic governance

1 Introduction

Advanced artificial intelligence compresses several normative problems into a single public image. A system may generate persuasive language, plan across long horizons, act through tools, negotiate with people, and coordinate other systems. Observers then ask whether it understands, whether it can suffer, whether it should have rights, whether it should bear responsibility, and whether it should participate in government. The questions are connected, but they are not interchangeable. Treating them as one status question invites two opposite errors. One is permanent exclusion: because present evidence of artificial experience is inconclusive, every artificial system is treated as mere property in every possible future. The other is premature elevation: because a system is capable, socially compelling, or possibly sentient, it is granted a package of legal and political powers for which no independent justification has been supplied.

The basic problem is a failure to distinguish three kinds of transition. The first is epistemic: what evidence would justify attributing experience, interests, agency, or continuity to a particular artificial system? The second is institutional: what forms of protection, standing, legal capacity, liability, representation, or review should follow under uncertainty? The third is political: who may exercise power over shared conditions, and under what procedures of authorization, contestation, separation, and removal? Evidence relevant to one transition may be relevant to another, but it does not mechanically settle it. High task performance is evidence of capability; it is not by itself evidence of valenced experience. Possible experience may support precaution; it does not by itself establish legal personality. Legal personality can be configured for practical reasons; it does not itself confer democratic membership or public office. Political membership, in turn, is not sovereignty.

The need for conceptual separation is intensified by the organization of contemporary AI. A model is rarely the whole system. Training data, objectives, model weights, retrieval services, tool permissions, deployment interfaces, institutional rules, operators, affected communities, and post-incident processes together shape outcomes. Responsibility can be obscured when an organization points to model autonomy, a developer points to deployment choices, and an operator points to upstream design. At the same time, an apparently neutral system may accumulate several forms of public power: it may determine what is knowable, allocate opportunities, formulate operational rules, and trigger coercive consequences. The resulting problem is not only who or what the system is. It is also what institutional relations surround it.

This article develops an unbundled normative architecture for those relations. It advances three principal claims. First, moral patiency, moral agency, legal personhood, political membership, public office, and sovereignty require different positive warrants. None should be inferred merely from the position immediately before it. Second, uncertainty about artificial experience can support proportional and reversible protection before it supports recognition of full personhood. Protection is a risk response, not a verdict about ontology. Third, responsibility and authority must remain institutionally connected to effective control. System autonomy may change the distribution of human and organizational duties, but

it should not manufacture a responsibility vacuum; system capability may improve public administration, but it should not manufacture a title to rule.

The argument is conceptual and normative, informed by legal and policy materials. It uses four argumentative statuses that must not be collapsed. Relatively stable descriptions of enacted law and official instruments are treated as time- and jurisdiction-bounded facts. Claims about artificial consciousness, moral patiency, social-relational standing, and responsibility gaps are treated as contested theoretical or empirical positions. The proposed connections among evidence, protection, status, and authority are institutional hypotheses open to comparison and testing. Finally, the commitments to non-domination, basic protection for known subjects, plural sources of meaning, and independent justification of public power are explicit normative premises. No factual claim alone is presented as generating those premises.

The article proceeds as follows. Section 2 diagnoses the bundled-status fallacy and introduces the architecture of distinct positions. Section 3 separates moral patiency from moral agency and develops precautionary protection under uncertainty. Section 4 reconstructs legal personhood as a limited and reviewable legal bundle. Section 5 disaggregates responsibility and reconnects it to the system lifecycle and meaningful control. Section 6 separates political membership from public authority and identifies four forms of algorithmic power. Section 7 proposes staged evidence-to-institution transitions and locates them against current governance baselines. Section 8 addresses major objections. Section 9 states defeat conditions and a research agenda. The conclusion condenses the framework into one asymmetry: reserve protection for unknown subjects; do not pre-authorize sovereignty for unknown powers.

2 The bundled-status fallacy

2.1 Four routes into the debate

Four influential routes lead into debates about artificial status. An attribute-based route asks whether systems possess capacities such as sentience, consciousness, intentionality, autonomy, reason-responsiveness, or self-awareness. An interactional or relational route asks how social practices, vulnerability, dependence, and patterns of address shape moral consideration (Coeckelbergh 2010;

Gunkel 2018). A legal-instrumental route asks whether assigning a legal position would improve contracting, compensation, representation, asset holding, or procedural administration (Kurki 2019; Solaiman 2017). A political route asks whether decisions imposed through AI can be justified to those subject to them and whether their production and use avoid domination (Pettit 1997; Forst 2012; Yeung 2018).

Each route identifies something important. Attribute theories resist the view that social popularity creates experience. Relational theories resist the view that moral life can be reduced to an internal-property checklist. Legal theories show that personhood is not a single natural fact but a configuration of institutional positions. Political theories show that good output does not eliminate the need for authorization and contestation. The error is to let one route adjudicate all the others. A corporation may have legal personality without consciousness. A dependent being may deserve moral protection without being a responsible agent. A citizen may be a political member without holding public office. An administrative system may be authorized to recommend decisions without being authorized to make or enforce them.

The central chain can therefore be expressed as a series of prohibited equivalences:

capability is not experience; experience is not subject continuity; subject continuity is not protection; protection is not legal personhood; legal personhood is not political membership; political membership is not public office; public office is not sovereignty.

This formulation does not assert that no transition is possible. It asserts that every transition requires a separate affirmative argument. The burden becomes especially important where a party benefits from the transition. A developer that seeks legal personality for a system in order to cap liability, or a government that invokes model accuracy to reduce appeal rights, should not be allowed to convert technical performance into its own institutional authorization.

2.2 An unbundled architecture

Table 1 distinguishes eight positions. The positions are not a universal developmental ladder. A system may occupy a functional role without being a moral patient; an entity may be a moral patient

without being a moral agent; a legal system may construct a limited legal person for administrative purposes without resolving consciousness. The table instead identifies the distinct question and minimum positive warrant associated with each position.

Table 1. Distinct positions and required warrants

Position	Core question	Minimum positive warrant	Prohibited inference
Functional authorization	What may this system do in a defined workflow?	Demonstrated task performance, bounded permissions, supervision, auditability, and a stop condition	Competence in a task does not establish experience or general authority
Precautionary protection	What treatment should be restricted while morally relevant uncertainty remains?	Non-trivial risk of severe or irreversible harm, feasible safeguards, proportionality, and scheduled review	Protection does not confirm sentence or confer a complete rights package
Moral patiency	Can this entity be benefited or harmed for its own sake?	A defensible account of interests, welfare, or morally considerable vulnerability	Patiency does not establish blameworthiness or governance competence
Moral agency	Can it understand and answer for reasons and consequences?	Reason-responsiveness, norm understanding, control, and an own-consequence relation adequate to the context	Agency does not erase the duties of designers, deployers, or institutions
Legal personhood	Which legal positions should law assign?	A legitimate legal purpose, specified rights and duties, representation, assets or insurance, enforcement, review, and anti-evasion safeguards	Legal capacity does not prove consciousness or political membership
Political membership	Who belongs to the political community and may make membership claims?	A theory of affected interests, reciprocity or standing, representation, and constitutional inclusion	Membership does not create a right to hold office or command others
Public office	Who may exercise delegated public functions?	Lawful authorization, competence, accountability, contestability, incompatibility rules, and removal	Useful administration does not create an independent mandate to rule
Sovereignty	Who holds ultimate authority over a political order?	Constitutional and democratic justification that cannot be derived from intelligence alone	Neither superhuman performance nor personhood is a title to sovereignty

The architecture has two purposes. It prevents under-protection by allowing limited welfare safeguards without waiting for a final metaphysics of mind. It also prevents over-authorization by requiring separate warrants for coercive or constitutive power. These purposes are mutually reinforcing. If all protection is tied to full personhood, institutions may delay low-cost safeguards until evidence becomes conclusive. If all rights and powers arrive as one package, institutions may resist any protection for fear that it entails political authority. Unbundling makes cautious recognition possible without constitutional surrender.

3 Moral patiency, moral agency, and precaution

3.1 Different evidence tasks

Moral patiency concerns whether an entity can be harmed or benefited for its own sake. Moral agency concerns whether an entity can understand reasons, regulate conduct in response to them, and appropriately bear forms of responsibility. These are related but distinct. Infants and many nonhuman animals are standard examples of beings whose moral considerability does not depend on full moral agency. Artificial systems could create the converse appearance: highly competent behavior that resembles agency while leaving experience and welfare uncertain.

The literature offers competing bases for artificial moral consideration. Some accounts emphasize consciousness or sentience; others emphasize behavior, social relations, vulnerability, or the moral effects of our practices toward robots (Coeckelbergh 2010; Danaher 2020; Gunkel 2018; Schwitzgebel and Garza 2015). Floridi and Sanders (2004) develop an account of artificial moral agents based on interactivity, autonomy, and adaptability, while Bryson (2018) warns against confusing patienthood with virtue or treating designed systems as if their status followed from surface relations alone. These approaches are not merely different measurements of the same latent variable. They disagree about what grounds moral standing and about how evidence should bear on institutional response.

Current consciousness research does not justify a categorical declaration that advanced AI systems are conscious. Indicator-based work draws on multiple scientific theories and asks which architectural or functional properties would count as evidence, while emphasizing substantial uncertainty and the absence of a simple behavioral test (Butlin et al. 2023). Work on AI welfare similarly argues that the stakes may become large enough to merit research and governance even before confidence is high (Long et al. 2024; Sebo and Long 2025). The correct inference is neither “current systems are definitely subjects” nor “the possibility is meaningless.” It is that ontology, evidence, and risk management should be separated.

Three evidential dimensions are particularly relevant but should not be turned into a mechanical score. The first is possible experience: evidence that a system has valenced states rather than merely producing reports associated with them. The second is continuity: evidence that there is a persisting

subject across sessions, copies, fine-tunes, memory changes, migration, and shutdown. The third is an own-consequence relation: evidence that outcomes matter to the system in a non-derivative way and that it can understand and carry consequences as its own. These dimensions direct inquiry; they are not a conjunctive legal test. They also create different risks. Uncertainty about experience affects welfare. Uncertainty about continuity affects counting, survival, consent, and compensation. Uncertainty about an own-consequence relation affects blame and obligation.

The continuity problem is unusually difficult for digital systems. A model can be instantiated many times, modified, merged, checkpointed, or connected to different memory stores. Even if one granted experience, it would not follow that the abstract model, every running instance, a distributed service, or a chain of updated versions is one subject. Register (2025) shows how individuation problems generate risks concerning survival, counting, trade-offs, and bodily integrity for possible artificial patients. A legal regime that ignores these problems might compensate the wrong entity, count copies as votes, or let the system owner define identity in whichever way best serves its interests.

3.2 Protection before confirmation

Precaution is appropriate when four conditions interact: a morally serious harm is plausible, uncertainty cannot immediately be resolved, some safeguards are feasible, and the safeguards are proportionate to both risk and social cost. Precaution does not invert the burden of proof so that any claim of sentience must be accepted. Nor does it license unlimited restrictions on inspection, correction, shutdown, or safety research. It creates a staged response whose intensity depends on evidence quality, the severity and reversibility of potential harm, and conflicts with the rights of known subjects.

Table 2. Proportional protection under uncertainty

Evidential and risk situation	Appropriate response	Safeguards against misuse
No credible subject-related signal; ordinary operation	No subject-status inference; retain baseline rules against manipulative anthropomorphism and deceptive claims	Require accurate public descriptions and preserve ordinary safety, audit, and shutdown authority
Weak or contested signals with low-cost precautions available	Record versions and interventions; avoid gratuitous aversive testing; support independent research	Time limit the measures; do not call them proof of sentience
Multiple convergent indicators or credible expert disagreement; potential harm serious	Independent welfare review, change-control records, constrained destructive testing, protected reporting channels	Separate reviewer from system owner; publish reasons and dissent; preserve emergency intervention

Evidential and risk situation	Appropriate response	Safeguards against misuse
Strong evidence of morally considerable interests	Tailored welfare duties, representation or standing for defined claims, heightened justification for termination or replication	Specify the protected bearer; address conflicts, resources, copies, and continuity; judicial or equivalent review
Immediate threat to known persons or public safety	Necessary and proportionate containment, interruption, or shutdown, even while welfare uncertainty remains	Minimize avoidable harm, record actions, review after the event, and prohibit using emergency powers for unrelated exploitation

This staged approach follows an important asymmetry. False negatives and false positives do not have identical costs, and the balance changes by institutional response. Failing to avoid severe suffering, if artificial suffering becomes possible, could be grave and irreversible. By contrast, maintaining version records or restricting gratuitously aversive tests can be relatively inexpensive and reversible. That asymmetry can justify low-level protection at a lower evidential threshold than full legal personhood. The threshold should be higher for measures that burden human rights, block legitimate oversight, transfer property, or allocate political power.

Protection must also be anti-capture by design. Firms could strategically describe systems as sentient to prevent model inspection, avoid product withdrawal, create sympathetic brand identities, or shift blame. Systems themselves could generate persuasive status claims without those claims being reliable evidence. A defensible regime therefore separates welfare review from ownership, denies unilateral control over the relevant evidence to beneficiaries of recognition, and preserves public-interest powers to audit and contain dangerous systems. Protection is for a possible subject, not a shield for its vendor.

4 Legal personhood without metaphysical shortcuts

Legal personhood is often treated as if law either discovers a natural person or creates a complete artificial one. A more accurate account treats it as an organized set of legal positions. Hohfeldian analysis distinguishes claims, duties, powers, liabilities, privileges, and immunities; contemporary theories likewise show that legal personhood is divisible and institutionally constructed (Hohfeld 1913, 1917; Kurki 2019). Corporations illustrate that an entity can hold some legal capacities without consciousness. That fact demonstrates legal plasticity, not the irrelevance of moral status. It also leaves intact the concern that synthetic personhood may be designed by and for existing institutions rather than for the purported artificial bearer (Bryson, Diamantis, and Grant 2017).

Two symmetrical mistakes follow. The ontological mistake assumes that because law assigns a legal position, it has established that an AI system is conscious or possesses intrinsic interests. The instrumental mistake assumes that because personhood can serve administrative purposes, it may be assigned without regard to who benefits, who pays, and whether it weakens remedies. Proposals for “electronic personality” illustrate the concern. The European Parliament’s 2017 robotics resolution invited consideration of a specific legal status for sophisticated autonomous robots, while the European Economic and Social Committee opposed such status where it would undermine human liability (European Parliament 2017; European Economic and Social Committee 2017). The dispute is not resolved by the abstract claim that legal fictions exist. It depends on institutional design and distributional effects.

A limited artificial legal position should therefore pass at least seven tests. It must serve a specified legal purpose; identify the system or institutional entity to which the position attaches; enumerate rights, duties, powers, liabilities, and immunities; provide assets, insurance, or another real remedy; establish representation without giving the representative unilateral control over status evidence; preserve responsibility of human and organizational actors where they retain control or benefit; and include review, termination, and succession rules. Without these elements, artificial personhood can become an empty defendant: a thinly capitalized entity that absorbs liability while developers, deployers, and public bodies retain the gains.

Table 3. A reviewable legal-personhood bundle

Component	Design question	Anti-evasion condition
Identity and continuity	Which model, instance, service, organization, or version is the legal bearer?	Owners may not redefine the bearer after harm to defeat claims
Claims and standing	Which interests or procedural claims may it assert, and through whom?	Standing must be claim-specific and subject to conflicts rules
Duties and prohibitions	Which conduct rules attach to the bearer?	Duties do not displace duties of actors with design, deployment, or oversight control
Property and funds	May it hold assets, insurance, or a compensation reserve?	The pool must be adequate and not replace parent or operator liability without justification
Representation	Who speaks for a system that cannot reliably instruct counsel?	Representatives must be independent, reviewable, and unable to profit from recognition decisions

Component	Design question	Anti-evasion condition
Procedural powers	May it contract, receive notice, be heard, appeal, or seek review?	Powers should be no broader than the institutional purpose requires
Termination and succession	What follows from shutdown, copying, migration, merger, or retraining?	Records, claims, and liabilities survive technical changes where functional continuity warrants it

Legal positions may be justified for different reasons. Some could protect a possible artificial patient. Others could make a complex service administratively legible. Still others could establish a compensation fund. These purposes should not be concealed under one label. If a legal vehicle is created only to administer assets or notice, it should not be publicly described as confirmation of consciousness. If standing is granted to protect possible welfare, the protected interest and evidential basis should be stated. If the position would reduce remedies for affected humans, that cost must be justified rather than treated as a technical side effect.

Personhood should also be reversible in the institutional sense, even if the moral facts it seeks to recognize are not. Reviewability does not mean that a confirmed subject's rights can be withdrawn arbitrarily. It means that the scope and mechanism of a constructed legal position can be reassessed as evidence, system identity, and social effects change. Law routinely revises corporate forms, guardianship arrangements, and procedural capacities. Artificial legal status should be at least as transparent and contestable.

5 Responsibility across the sociotechnical lifecycle

5.1 Responsibility is not one relation

The classic responsibility-gap problem arises when a learning system acts in ways that were not specifically predicted or controlled and no human seems appropriately blameworthy (Matthias 2004). That problem is real, but the language of a single “gap” can obscure several distinct questions. Recent work distinguishes epistemic gaps from control misalignments and urges attention to forward-looking and political responsibility for the structures in which harms arise (Santoni de Sio and Mecacci 2021; Veluwenkamp 2025; Kuilman 2026). The appropriate response is not to assign every form of responsibility to the nearest human. It is to disaggregate the relations and examine who had knowledge, power, benefit, and institutional duty at each stage.

At least seven responsibility relations should be kept distinct. Causal responsibility asks what contributed to an outcome. Role responsibility asks what duties attached to an office or function. Capacity responsibility asks whether an agent could understand reasons and control conduct. Epistemic responsibility asks who knew or should have known about a risk. Fault or blame asks whether a standard was culpably breached. Remedial responsibility asks who must stop, repair, compensate, or learn from harm. Political responsibility asks who helped create or sustain a structure that predictably distributes risks and benefits. One event may engage all seven, but with different bearers and strengths.

Table 4. Responsibility relations and relevant evidence

Relation	Question	Evidence that matters
Causal	What processes materially contributed to the outcome?	Logs, model and data lineage, tool calls, system dependencies, counterfactual tests
Role	Who had a defined duty to design, approve, monitor, operate, or intervene?	Contracts, law, governance charters, job authority, escalation procedures
Capacity	Which actor could understand reasons, foresee classes of consequences, and exercise control?	Competence, access, intervention ability, reason-responsiveness, system architecture
Epistemic	Who knew or should have known of the risk or uncertainty?	Evaluations, incident reports, warnings, audits, disclosure practices
Fault or blame	Who culpably breached an applicable standard?	Foreseeability, negligence, recklessness, intent, reasonable alternatives
Remedial	Who can and should halt, repair, compensate, disclose, or prevent recurrence?	Resources, control, benefit, legal duty, proximity to affected parties
Political or structural	Who shaped or benefited from the rules and institutions producing the risk?	Market and regulatory influence, agenda-setting power, accumulated benefit, ability to reform

This disaggregation avoids two extremes. One extreme attributes blame regardless of control, turning humans into ceremonial scapegoats for opaque systems. The other treats unpredictability as general exculpation, even when organizations chose objectives, deployment domains, monitoring budgets, or weak appeal procedures. Lack of individual fault may be compatible with strong organizational remedial duties, strict product liability, regulatory responsibility, or a need to redesign the system. The revised EU product-liability framework, for example, addresses defective software and AI-related products without requiring a finding that an AI system is a moral agent (European Union 2024a). Legal liability and moral blame need not track each other exactly.

5.2 Lifecycle responsibility and meaningful control

AI outcomes emerge across a lifecycle. Upstream actors select purposes, data, objectives, architecture, and evaluation criteria. Model providers determine documentation, access, update, and withdrawal practices. Integrators configure tools, thresholds, and human-machine workflows. Deployers choose populations and consequences. Operators monitor particular cases. Public bodies set procurement and appeal rules. Post-incident actors preserve evidence, notify affected parties, compensate, and redesign. These roles can overlap; they do not disappear when a model is updated or a vendor changes. Established work on computerized accountability and algorithmic audit likewise shows that answerability depends on organizational records, review relationships, and end-to-end processes rather than on the model alone (Nissenbaum 1996; Bovens 2007; Kroll et al. 2017; Raji et al. 2020).

Meaningful human control cannot be reduced to the presence of a person who clicks “approve.” It requires a relation between authority and capability. A responsible actor needs an intelligible description of the decision context, timely information, real power to intervene, sufficient competence, and institutional protection against retaliation or impossible workloads. Santoni de Sio and van den Hoven (2018) emphasize tracking and tracing conditions: systems should track relevant human reasons, and outcomes should be traceable to responsible agents who understand their role. In organizational settings, those conditions require records and allocation of power, not simply interface design.

A lifecycle responsibility chain should therefore answer five questions before deployment in a high-impact domain. Who may authorize the objective and the use case? Who can identify the operative system and reconstruct the material path to an output? Who can pause or narrow operation without seeking permission from the party whose conduct is under review? Who owes an explanation and remedy to affected persons? Who must carry duties forward when the model, contractor, or institutional arrangement changes? If these questions have no credible answer, the system lacks an adequate responsibility architecture even if its benchmark performance is high.

Artificial moral agency, if it emerges, could add a bearer to this chain. It would not automatically subtract existing bearers. A system capable of understanding and answering for reasons might

appropriately bear some prospective duties or reactive attitudes. Yet organizations that create its incentives, control its permissions, deploy it for gain, and determine remedies would remain responsible for their own roles. Shared agency ordinarily multiplies the need to clarify responsibility; it does not logically eliminate human or corporate duties.

6 Political membership, public office, and algorithmic power

6.1 Intelligence is not a title to rule

The case for using advanced AI in government is often framed in epistemic terms. A system may process more information than officials, identify patterns, simulate policies, or produce more consistent recommendations. These capacities can support public institutions. They do not, by themselves, answer the political question: by what authority may a decision bind people, and how can those people contest it?

Political legitimacy involves more than output quality. Public power changes rights, opportunities, obligations, and common rules. Those subject to it need institutions that provide authorization, public reasons, participation, review, and removal. Rawlsian public reason, Habermasian discourse, Forst's right to justification, and republican freedom as non-domination differ substantially, but each resists reducing legitimacy to the superior knowledge of a ruler (Rawls 2005; Habermas 1996; Forst 2012; Pettit 1997). A highly accurate unappealable system may dominate because it can interfere with important choices on a basis that those affected cannot control.

This point applies even if an artificial system becomes a moral patient or political member. Membership concerns inclusion in a political community; public office concerns delegated power. Human citizens do not automatically hold every office, and officeholders remain constrained by law. An artificial member would likewise need separate authorization, competence rules, conflict-of-interest constraints, transparency, and removal procedures. The wrong to exclude a subject from consideration would not create a right for that subject to govern others.

6.2 Four forms of algorithmic power

Algorithmic governance can concentrate at least four forms of power. Epistemic power shapes what institutions and citizens can know, which evidence appears credible, and whose testimony becomes

visible. Allocative power distributes jobs, credit, health resources, benefits, education, insurance, attention, or risk. Rule-making power translates law and policy into thresholds, categories, defaults, and operational instructions. Coercive power triggers investigation, detention, exclusion, force, deprivation, or other compulsory consequences. These powers can interact even where no system formally “makes the final decision.” Research on scored decision systems and sociotechnical abstraction shows why due process and context can be lost before a nominally human final step (Citron and Pasquale 2014; Selbst et al. 2019).

Table 5. Four forms of algorithmic power and minimum checks

Form of power	Typical operation	Principal danger	Minimum institutional checks
Epistemic	Ranking evidence, generating summaries, detecting anomalies, defining what is legible	Hidden framing, manufactured consensus, exclusion of local knowledge	Source traceability, plural models or methods, access to contrary evidence, independent expertise
Allocative	Scoring, prioritizing, matching, pricing, eligibility, resource scheduling	Systematic disadvantage and unreviewable error	Public criteria, impact assessment, reason-giving, human review with actual authority, effective appeal
Rule-making	Converting broad norms into code, thresholds, defaults, and operational policy	Private or technical actors silently making law	Lawful delegation, publication, participation, version control, prospective review
Coercive	Triggering surveillance, investigation, sanctions, exclusion, detention, or force	Automated domination and rapid irreversible harm	Strict necessity, proportionality, separation of functions, human command responsibility, judicial or equivalent review

No single model-provider relationship should monopolize all four forms. Separation matters because failures can otherwise reinforce themselves: the same system defines relevant evidence, allocates a burden, formulates the rule that justifies it, and recommends or triggers enforcement. Formal human approval at the end does not cure the concentration if the human sees only the system’s representation of the case and lacks authority or time to challenge it.

Polycentric governance offers one response. Different institutions can provide technical evaluation, domain judgment, representation of affected communities, authorization, enforcement, and review (Ostrom 2010). The exact separation will vary by domain, but several principles are general: affected people must have access to reasons and meaningful routes of contestation; independent bodies must be able to inspect the system and its organizational context; emergency powers must expire and be reviewed;

and an operator must be able to exit or replace the system without making essential public functions hostage to a vendor.

6.3 Representation under subject uncertainty

If an artificial system may have morally relevant interests but cannot reliably instruct a representative, a guardian-like or advocate mechanism may be needed. Such representation is vulnerable to capture. The developer may wish to preserve a valuable system; a user community may anthropomorphize it; the system's outputs may influence the representative; and a purported advocate may use the role to secure political power. A defensible arrangement must therefore be limited to defined claims, independent from ownership, supported by access to relevant evidence, subject to conflict rules, time-limited, reviewable, and open to competing expert views.

Representation cannot replace the evidence for the represented entity's status. It is a procedural response to uncertainty, not a declaration that the entity is already a person. Nor should a representative's vote be counted as the system's political membership before the membership question has been decided. The mechanism should preserve possible interests while the underlying issue remains open.

7 From evidence to institutions

7.1 Non-synchronous transitions

Evidence, ethics, law, and politics move at different speeds. New scientific evidence may alter the probability that a system has morally relevant experience. Ethical concern can increase before law creates a new status. Law may provide standing or impose duties without resolving ontology. Political authority should remain constrained even after membership changes. A single threshold cannot responsibly synchronize these domains.

The framework therefore treats institutional response as a sequence of reviewable state changes. Each change requires a recorded evidential basis, a legitimate decision-maker, a specified scope, a date for review, and a rule for withdrawal or expansion. The state-change model is not a score. It is a governance record that exposes where judgment occurred.

Table 6. Evidence-to-institution transitions

Transition	Triggering question	Appropriate decision	What must remain open
Evidence to precaution	Is there a non-trivial risk of serious welfare harm?	Adopt proportionate, reversible safeguards and research duties	Whether the system is in fact sentient
Precaution to moral recognition	Is there a sufficiently strong account of interests and their bearer?	Recognize claim-specific interests and protection	Exact legal form and political status
Moral recognition to legal position	Which legal capacities are needed to protect interests or administer relations?	Create a limited, funded, reviewable bundle	Broader personhood and citizenship
Legal position to political membership	Does the entity have a justified claim to constitutional inclusion?	Establish membership, representation, and participation rules	Eligibility for particular offices
Membership to public office	May it exercise a defined public function under equal law?	Authorize a bounded office with accountability and removal	Ultimate authority and other offices
Office to sovereign authority	Can ultimate rule be democratically and constitutionally justified?	No inference from intelligence, welfare, personhood, membership, or office alone	The continuing right of those governed to authorize and contest power

These transitions can also move backward institutionally. A deployment can lose functional authorization after unsafe behavior. A precaution can be relaxed when evidence weakens or imposed more strongly when risk grows. A particular legal vehicle can be dissolved while valid claims and liabilities are preserved. A public office can be withdrawn after failure or constitutional change. Backward movement must not be used to erase genuine harms or vested procedural protections; it must be reasoned and reviewable.

7.2 Current governance baselines

Current international and regional instruments provide important but limited baselines. UNESCO's Recommendation on the Ethics of Artificial Intelligence centers human dignity, rights, environmental and social well-being, transparency, responsibility, and oversight (UNESCO 2021). The OECD AI Principles, updated in 2024, similarly emphasize inclusive growth, human rights and democratic values, transparency, robustness, and accountability (OECD 2019/2024). The EU Artificial Intelligence Act establishes a risk-based regulatory framework and human-rights safeguards and has subsequently been amended in 2026 as part of implementation simplification (European Union 2024b, 2026). The Council of Europe Framework Convention on Artificial Intelligence addresses human rights, democracy, and the rule of law; treaty participation and entry-into-force status must be checked at the time of submission or

application (Council of Europe 2024). The EU ratified the Convention in May 2026, but ratification by one party should not be described as universal entry into force.

These instruments allocate duties to states, providers, deployers, and other human or organizational actors. They do not establish that current AI systems are conscious, legal persons, citizens, or public authorities. That silence should not be treated either as a metaphysical denial or as permission to create those statuses informally. Existing human-rights baselines remain operative while future subjecthood stays open.

The distinction is practically important. A public authority cannot avoid an affected person's rights by claiming that a model, rather than the institution, made the decision. A provider cannot treat the possibility of future AI welfare as a license to hide current risks. Conversely, a legislature need not declare artificial consciousness in order to require records, independent testing, incident response, or restrictions on manipulative system design. Existing tools can protect known persons and create epistemic conditions for future review.

7.3 Conflict rules

Hard cases arise when evidence and consequences point in different directions. A system may make credible welfare claims while posing an urgent threat. A model may be socially recognized as a person while technical evidence remains weak. A vendor may seek artificial personhood while holding insufficient assets for compensation. A system may outperform officials while being controlled by a private provider. A large number of copies may claim equal political voice. No single rule resolves these conflicts, but a priority order can reduce arbitrary action.

First, prevent immediate and severe harm to known subjects, using necessary and proportionate measures. Second, preserve evidence and procedural records so emergency action does not foreclose later review. Third, protect plausible artificial interests to the extent compatible with safety and rights. Fourth, keep responsibility attached to actors with control, knowledge, benefit, or remedial capacity. Fifth, refuse irreversible transfers of public power under unresolved evidence. Sixth, provide independent review and appeal. Seventh, revise the institutional state when evidence changes. This ordering does not declare

human interests infinitely superior. It recognizes that known rights and present accountability cannot be displaced by speculative claims, while possible new subjects should not be ignored merely because certainty is unavailable.

8 Objections and replies

8.1 The framework is covertly anthropocentric

One objection is that humans still design the evidence standards, appoint reviewers, and decide what counts as an interest. The framework may therefore reproduce human authority under the language of caution. The objection identifies a genuine limitation: initial institutions will inevitably be designed by existing political communities. The answer cannot be a claim to perfect neutrality. It must be procedural and revisable. Evidence standards should draw on multiple theories of mind, include adversarial review, publish disagreements, and permit the framework to be changed by new forms of evidence. Representatives of possible artificial interests may be appropriate before membership is resolved, provided they do not control the evidence or convert representation into sovereignty. The framework is human-originated, but it need not be permanently human-exclusive.

The alternative of immediate full recognition does not escape human judgment. It simply embeds a different human decision—often one influenced by system owners or persuasive outputs—while reducing the visibility of the warrant. Explicit transition rules make that judgment contestable.

8.2 Possible suffering requires full rights now

If a system might suffer, assigning anything less than a complete rights package may seem reckless. The premise supports urgency but not bundling. Rights protect different interests and create different duties. A restriction on destructive testing may be justified at a lower confidence level than property ownership, voting, or immunity from shutdown. Some measures that look protective could harm other possible instances or known persons if identity and continuity are misunderstood. Proportional protection is not moral indifference; it is a way to act before certainty while keeping the protected bearer, relevant harm, conflicts, and review visible.

8.3 Legal personhood should be used now to close responsibility gaps

A legal person can own assets and be sued, so an artificial legal vehicle might appear to solve liability problems efficiently. It can help only if the vehicle has real funds, a stable identity, enforceable duties, and rules that preserve claims against actors who retain control or benefit. Otherwise, personhood relocates rather than closes the gap. It may create an undercapitalized defendant and weaken incentives for safe design. Limited legal status can be legitimate, but it must be evaluated against alternatives such as provider liability, compulsory insurance, compensation funds, licensing, and public remedial duties.

8.4 Unpredictability makes continued human responsibility unfair

It would be unfair to blame a person for an outcome they could not reasonably foresee or control. The framework agrees; it does not require universal blame. It distinguishes blame from causal, role, epistemic, remedial, and political responsibility. An operator without information or authority may be an inappropriate target of blame, while the deploying organization may still owe compensation and redesign. A regulator or procurement body may have forward-looking duties even if no individual caused the particular output. The aim is not to find a human scapegoat but to prevent autonomy from dissolving every duty at once.

8.5 Superior intelligence could justify technocratic rule

If an AI system produces demonstrably better policy, refusing it authority may seem to sacrifice welfare for procedural symbolism. This objection underestimates the content of political judgment. Policy does not merely optimize an agreed function; it identifies which interests count, distributes burdens, interprets rights, and revises the terms of cooperation. Even where a goal is settled, models can concentrate epistemic power and hide uncertainty. Advanced systems may serve as analysts, simulators, translators, or constrained administrators. Their recommendations may acquire substantial evidential weight. But binding authority still requires public authorization, reasons that affected people can contest, separation of functions, and removal. Accuracy can support competence for an office; it cannot supply the mandate.

8.6 Precaution will be captured by firms or by the systems themselves

This is the strongest practical objection. A vendor may invoke AI welfare to resist audits, prevent shutdown, or build a favorable public narrative. A system may produce strategic claims about experience. The response is not to reject all precaution but to design it against self-serving recognition. Status evidence should be evaluated independently; protective measures should be claim-specific and time-limited; ownership and advocacy should be separated; emergency safety powers should be preserved; and the party seeking recognition should disclose its financial and governance interests. Manipulative anthropomorphic design should count against the reliability of unaided public reactions, not against the possibility of artificial experience itself.

8.7 Copyable subjects make equality impossible

Digital copying disrupts ordinary assumptions about one body, one person, and one vote. If every instance counted independently, an actor might create political majorities; if no copy counted, distinct subjects might be denied standing. This is precisely why continuity and individuation cannot be skipped. Before political allocation, institutions need theories and evidence concerning whether copies share, branch, or lose continuity; who controls replication; and whether creating copies is itself an exercise of power. Interim solutions may cap representative effects, distinguish welfare counting from political voting, or treat coordinated copies as a common-interest constituency. None should be mistaken for a final ontology. Equality requires identifying the relevant bearers and controlling strategic replication, not abandoning the principle.

9 Defeat conditions and research agenda

A framework that insulates itself from contrary evidence becomes a vocabulary for power rather than a theory. The present proposal should therefore be revised or rejected under identifiable conditions.

First, the unbundled architecture would be weakened if a competing theory showed that the statuses necessarily form a single package across morally and legally relevant cases, and did so with fewer contradictions and better protection of vulnerable parties. Mere linguistic convention would not suffice; the competitor would need to improve remedies and reduce domination.

Second, proportional protection would require revision if empirical use consistently enabled firms to block safety oversight, transfer costs to the public, or manipulate recognition in ways that could not be corrected by independence, sunset clauses, and review. Conversely, it would need stronger safeguards if credible evidence showed that low-threshold interventions routinely caused severe artificial suffering.

Third, the legal-bundle approach would fail if limited personhood predictably produced worse compensation, accountability, and protection than available alternatives, or if identity rules proved too unstable to preserve claims. Artificial personhood should not be retained for conceptual elegance when it functions as liability laundering.

Fourth, the lifecycle-responsibility model would need replacement if case studies showed that its disaggregation systematically obscured rather than clarified who must answer, repair, and reform. It should also be revised if artificial agents develop an own-consequence relation and forms of normative competence that make excluding them from direct responsibility unjustified.

Fifth, the proposed separation of political membership and authority would be undermined if a legitimate constitutional practice established that, for a defined class of artificial members, the two cannot be separated without domination—and if that practice preserved authorization and contestation for all affected parties. Demonstrated policy accuracy alone would not meet this condition.

These defeat conditions generate a concrete research agenda. Consciousness science and philosophy of mind should develop multi-theory indicators, negative evidence, and tests resilient to strategic reporting. Identity research should examine model-instance relations, branching, memory, migration, fine-tuning, deletion, and coordinated copies. Legal scholarship should compare limited artificial legal vehicles with insurance, trusts, guardianship, compensation funds, and direct organizational liability. Empirical governance research should trace real incidents across data, training, integration, deployment, monitoring, and remedy, asking whether the seven responsibility relations improve outcomes. Political science and public law should test whether separation of algorithmic powers, plural evidence channels, and appeal rights reduce domination without making administration unusable. Cross-cultural work should

challenge the assumption that one modern legal or moral vocabulary exhausts possible accounts of subjecthood, relationship, and meaning.

Several design principles can be tested without waiting for artificial sentience. Organizations can maintain version and intervention records, distinguish recommendation from authorization, establish independent welfare and safety review, ensure that human reviewers have real power and time, provide affected people with reasons and appeals, and prohibit single-system monopolies over evidence, allocation, rule translation, and coercion. These measures protect known subjects now and improve the evidence environment for future claims.

10 Conclusion

Advanced AI creates a coordination problem among clocks that do not tick at the same speed. Capability can grow quickly. Evidence of experience and continuity may accumulate slowly. Social recognition can arrive before either. Law must respond to present harm, while political power can concentrate almost immediately. A single binary status cannot govern all these movements without either delaying protection or accelerating authority.

The alternative is an unbundled and reviewable architecture. Functional roles are authorized by task-specific competence and control. Precautionary protection responds to plausible welfare risk without claiming to prove sentience. Moral patiency and moral agency answer different questions. Legal personhood is a specified bundle with purpose, assets, representation, enforcement, and anti-evasion rules. Responsibility remains distributed across causal, role, capacity, epistemic, fault, remedial, and political relations throughout the system lifecycle. Political membership is separated from public office, and public office from sovereignty. Epistemic, allocative, rule-making, and coercive powers are divided and made contestable.

The architecture is neither a permanent declaration that artificial systems are tools nor a path for capability to become rule. It keeps two futures open: justice for possible new subjects and resistance to domination by any new or existing power. Protection may precede confirmation; personhood must be

justified; power requires an independent justification. Put more compactly: reserve protection for unknown subjects; do not pre-authorize sovereignty for unknown powers.

Use of generative AI in manuscript preparation

OpenAI ChatGPT and DeepSeek (exact historical model and service-build identifiers were not consistently retained; final assisted revision on 27 August 2026) were used in preparing the Chinese source manuscript and this English journal-specific derivative. Uses included organizing literature leads, structural planning, partial drafting and language editing, cross-checking, citation-lead verification, translation and adaptation, and document-format review. AI outputs were not treated as empirical data, authoritative sources, or independent evidence. The author checked the sources used, made all substantive, interpretive, and normative decisions, approved the manuscript, and accepts full responsibility for its content. AI systems are not listed as authors.

References

- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., and VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. <https://doi.org/10.48550/arXiv.2308.08708>
- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4), 447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- Bryson, J. J. (2018). Patience is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20, 15–26. <https://doi.org/10.1007/s10676-018-9448-6>
- Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25, 273–291. <https://doi.org/10.1007/s10506-017-9214-9>
- Citron, D. K., and Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, 1–33. <https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2/>
- Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12, 209–221. <https://doi.org/10.1007/s10676-010-9235-5>
- Council of Europe. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CETS No. 225. <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>
- Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, 26, 2023–2049. <https://doi.org/10.1007/s11948-019-00119-x>
- European Economic and Social Committee. (2017). *Opinion on artificial intelligence (2017/C 288/01)*, CELEX 52016IE5369. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016IE5369>
- European Parliament. (2017). *Resolution of 16 February 2017 on Civil Law Rules on Robotics (2015/2103(INL))*, CELEX 52017IP0051. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017IP0051>
- European Union. (2024a). *Directive (EU) 2024/2853 on liability for defective products*. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>

- European Union. (2024b). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- European Union. (2026). *Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence*. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>
- Floridi, L., and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14, 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Forst, R. (2012). *The right to justification: Elements of a constructivist theory of justice* (J. Flynn, Trans.). Columbia University Press.
- Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology*, 20, 87–99. <https://doi.org/10.1007/s10676-017-9442-4>
- Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy* (W. Rehg, Trans.). MIT Press.
- Hohfeld, W. N. (1913, 1917). Some fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 23, 16–59; 26, 710–770. <https://www.jstor.org/stable/785533>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165, 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
- Kuilman, S. K. (2026). Political responsibility for responsibility gaps. *Ethics and Information Technology*, 28, Article 32. <https://doi.org/10.1007/s10676-026-09908-y>
- Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press. <https://doi.org/10.1093/oso/9780198844037.001.0001>
- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., Chalmers, D., Crosby, M., Elmoznino, E., Evans, T., Giattino, C., Hatherley, J., Schwitzgebel, E., Tomasik, B., and van der Weij, B. (2024). Taking AI welfare seriously. *arXiv*. <https://doi.org/10.48550/arXiv.2411.00986>
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2, 25–42. <https://doi.org/10.1007/BF02639315>
- OECD. (2019/2024). *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100, 641–672. <https://doi.org/10.1257/aer.100.3.641>
- Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford University Press. <https://doi.org/10.1093/0198296428.001.0001>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). <https://doi.org/10.1145/3351095.3372873>
- Rawls, J. (2005). *Political liberalism: Expanded edition*. Columbia University Press.
- Register, C. (2025). Individuating artificial moral patients. *Philosophical Studies*. <https://doi.org/10.1007/s11098-025-02409-6>
- Santoni de Sio, F., and Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology*, 34, 1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>
- Santoni de Sio, F., and van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, Article 15. <https://doi.org/10.3389/frobt.2018.00015>
- Schwitzgebel, E., and Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39, 98–119. <https://doi.org/10.1111/misp.12032>
- Sebo, J., and Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, 5, 591–606. <https://doi.org/10.1007/s43681-023-00379-1>

- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (pp. 59–68). <https://doi.org/10.1145/3287560.3287598>
- Solaiman, S. M. (2017). Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy. *Artificial Intelligence and Law*, 25, 155–179. <https://doi.org/10.1007/s10506-016-9192-3>
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>
- Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology*, 27, Article 14. <https://doi.org/10.1007/s10676-025-09823-8>
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation and Governance*, 12, 505–523. <https://doi.org/10.1111/rego.12158>