

Reckoning with Ethics Law and Political Philosophy in the Age of AGI

Moral Status, Legal Personhood, Responsibility Chains, and the Limits of Public Power

Reckoning with Ethics, Law, and Political Philosophy in the Age of AGI

Moral Status, Legal Personhood, Responsibility Chains, and the Limits of Public Power

Author: Zijun Fu Produced by: Civilization Leap Research Group

Independent Release V1.3 Revision on Dimensional Leap and Cross-Disciplinary Causal Relations

For joint scrutiny by ethics, jurisprudence, political philosophy, AI governance, and future technology studies. This specialist research paper in legal theory / comparative AI governance centers on moral patency and agency, precautionary protection, legal standing, lifecycle responsibility chains, and independent justification of public power. It is not legal advice for any particular jurisdiction. It does not establish subjective experience, full personhood, or political standing in current advanced AI, nor permanently exclude future new subjects. It also provides entry points for evidential handoffs to neighboring disciplines, comparative-law validation, and joint research.

Text revised: October 10, 2026 Author: Zijun Fu Produced by: Civilization Leap Research Group

Abstract translated from the Chinese

Advanced artificial systems bring three questions that long obscured one another to the institutional foreground simultaneously: who might be harmed by our actions, who can give reasons for actions and bear responsibility, and who is entitled to make and enforce common rules. In human cases, moral patency, agency, legal personhood, political membership, and public authority are often borne by the same natural person. In artificial systems, functional capabilities, social recognition, possible experience, subject continuity, asset arrangements, and governance permissions can separate. Binary labels such as tool or person and rights or no rights either postpone protection until proof or prematurely insert personhood and power into uncertain evidence. In dialogue with competing traditions in ethics, jurisprudence, and political philosophy, this paper proposes an unbundled framework: moral patients are distinguished from moral agents; precautionary protection is not confirmation of personhood; legal personhood is understood as a bundle of legal positions that can be configured individually, remedied, and reviewed; causal contribution, role responsibility, answerability, legal liability, moral blameworthiness, duties of repair, and political accountability are recorded separately; and political membership is further separated from public authority. The paper develops a positive warrant matrix from potential interests and susceptibility to harm through protection, legal standing, political membership, and public office, together with lifecycle responsibility chains, modular rights, separation of four forms of algorithmic power, representative arrangements, and transitions between evidential and institutional states. Its core conclusion is that protection may come first, personhood requires justification, and power requires separate justification. Current international governance instruments establish human rights, dignity, democracy, the rule of law, safety, transparency, and accountability as public baselines without deciding AGI's ontological subjecthood. Existing law is thus neither a ceiling on openness to future subjects nor grounds for granting personhood and sovereignty in advance today. Civilization Leap provides a minimal philosophical foundation: the historical coincidence of intelligence,

consciousness, subjecthood, life, responsibility, rights, and bearers of meaning in humans is coming apart; meaning is fundamental and its sources may differ; finitude is not its only source, but is an important structure through which known human meaning enters reality and acquires irreplaceable weight; and capability, recognition, standing, rights, and power must not automatically imply one another. The paper concludes with the strongest objections, failure conditions, and an interdisciplinary research agenda. It further moves disciplinary responsibility upstream to the generation of goals and purposes: a shift from self-interest first to self-interest with others first brings genuine stakeholders into judgment before purposes are generated and locked in. Disciplinary facts, direction, methods, and outcomes are stated separately; local successes or failures do not substitute for continued coexistence and the three long-term goals. Accordingly, research moves beyond distinguishing disciplinary objects toward investigating cross-disciplinary relations in depth. It traces how judgments in this field alter reality through other fields, and how feedback returns to change the original conditions and the purposes of the next round. Stakes across multiple actors, stages, and settings must be presented separately; correlations, common causes, mediation, and causal feedback must be tested separately; added relational dimensions must be open to specific counterevidence.

Keywords: advanced artificial intelligence; moral status; moral patency; legal personhood; responsibility chains; public power; non-domination; contestable governance

Abstract from the source English text

Advanced AI systems force three questions into the same institutional arena: who may be harmed by our actions, who can answer for actions and bear responsibility, and who may legitimately make and enforce common rules. In familiar human cases, moral patency, agency, legal personhood, political membership, and public authority are often borne by the same natural person. Artificial systems, by contrast, can separate functional capacity, social recognition, possible experience, diachronic continuity, asset arrangements, and governing permissions. Binary labels—tool or person, rights or no rights—therefore either postpone protection until proof arrives or smuggle personhood and power into uncertain evidence. In dialogue with competing traditions in ethics, jurisprudence, and political philosophy, this paper develops an unbundled framework. Moral patency is separated from moral agency; precautionary protection from confirmed personhood; legal personhood is treated as a reviewable bundle of legal positions; causal contribution, role responsibility, answerability, legal liability, blameworthiness, duties of repair, and political accountability are recorded separately; and political membership is distinguished from public authority. The resulting architecture includes a positive warrant matrix from potential interests and precautionary protection through legal status and political membership to bounded public office, together with lifecycle responsibility, modular rights, a separation of four forms of algorithmic power, fiduciary representation, and evidence-to-institution transitions. Its governing rule is: protection may precede confirmation; personhood must be justified; power requires an independent justification. Current international instruments establish human rights, dignity, democracy, the rule of law, safety, transparency, and accountability as public baselines, but do not adjudicate AGI's ontological subjecthood. Existing law is neither the ceiling of future moral inclusion nor a warrant for granting personhood or sovereignty today. Civilization Leap supplies a minimal philosophical root: the historical co-instantiation of intelligence, consciousness, subjecthood, life, responsibility, rights, and meaning-bearing in human beings is being technically separated; meaning is fundamental and may have multiple sources; finitude is not the sole source of meaning but is an important structure through which known human meaning enters reality and acquires irreplaceable weight; and capability, recognition, status, rights, and power must not be inferred from one another. The paper closes with the strongest objections, defeat conditions, and a cross-disciplinary research agenda. The paper further locates disciplinary responsibility before goal and purpose generation: a shift from self-interest first to self-interest with others first brings subjects with genuine stakes into judgment before goals are formed and locked. Facts, direction, methods, and outcomes remain distinct. The analysis is further extended to cross-disciplinary causal relations. It traces how judgments in this field change real-

world conditions through other domains, and how feedback modifies the original conditions and subsequent purposes. Effects across actors, stages, and settings must be distinguished from correlations, common causes, and merely asserted connections.

Keywords: advanced AI; moral status; moral patiency; legal personhood; responsibility chain; public authority; non-domination; contestability

Research Positioning Scope of Materials and Evidential Discipline

This study begins from theoretical and dimensional leap. It preserves disciplinary precision while deepening the analysis of causal relations with neighboring fields. Disciplinary boundaries identify burdens of proof; they do not end the tracing of shared consequences. Established scholarship participates in competing explanations and counterexample testing. The absence of a place in existing classifications cannot require new questions to retreat into old frameworks.

This is an independent, in-depth study in ethics, law, and political philosophy. It is not an expert determination of personhood in actual systems, does not constitute legal advice in any particular jurisdiction, and does not preempt decisions about standing by legislatures, courts, or public communities. AGI denotes a prospective category here. Discussion of actual institutions is limited to publicly verifiable artificial systems, deployment relationships, and regulatory texts. Normative and legal status was verified through August 24, 2026. The EU AI Act applies in stages, and signature and ratification status of the Council of Europe Framework Convention also changes; these should be rechecked before external release or submission.[29][30]

Four argumentative statuses are distinguished. Category A comprises relatively stable support: existing legal texts, institutional facts, and verifiable responsibility and procedural structures. Category B covers frontier or competing explanations: scholarly disputes over artificial moral status, relational approaches, behavioral approaches, instrumental accounts of legal personhood, responsibility gaps, and algorithmic non-domination. Category C comprises this paper's interdisciplinary connections, which still require comparative-law and institutional testing: separating three questions, proportionate anticipatory protection, seven forms of responsibility, four forms of algorithmic power, evidence-to-institution transitions, and meaning-power boundaries. Category F consists of explicit philosophical and normative commitments: meaning may have multiple sources; efficiency cannot cancel basic protection for known subjects; capability, standing, rights, and power must not automatically imply one another; and public power requires independent justification. Category A facts cannot automatically settle Category B disputes, still less directly generate Category C/F institutional or value conclusions. B, C, and F must not masquerade as established science.

Ethics, law, and political philosophy can independently distinguish moral patiency, moral agency, protection, legal personhood/standing, responsibility, political membership, and public power, specifying the normative reasons, legal purposes, and procedural safeguards each position requires. They must stop automatic extrapolation on empirical issues such as consciousness, life, technical capabilities, versions, and control chains, and must not treat comparative-law generalizations as currently effective law in specific jurisdictions. Epistemology, cognitive science/philosophy of mind, and life sciences reenter on subject evidence; AI/systems science explains actual capabilities, decision chains, and controllability; political science addresses public authorization and legitimacy; and particular courts and jurisdictions determine legal effect, remedies, and liability respectively.

The external manuscript does not retain internal project numbering, component interfaces, or governance codes. Internal research volumes supply conceptual resources and failure conditions. This paper reconstructs an argument unfamiliar readers can understand independently and challenge paragraph by paragraph. Readers need not first accept every claim of Civilization Leap to examine its conceptual distinctions, interpretations of sources, institutional designs, and original bridges separately.

An Entry Point for New Readers and the Minimal Philosophical Interface of Civilization Leap

This paper invokes only four explicit, revisable philosophical premises. First, the shared rupture of the AGI era does not simply announce the failure of old philosophy; it consists in technologies beginning to separate intelligence, consciousness, subjecthood, life, responsibility, rights, and bearers of meaning, which long coincided naturally in humans. Second, meaning is fundamental and its sources may differ: existence, relationships, action, creation, responsibility, and lived experience can all supply meaning. Third, finitude is not meaning's only source, but is an important structure through which known human meaning shapes choices, bears consequences, and acquires irreplaceable weight. Fourth, functional response, social recognition, legal standing, ontological subjecthood, and political standing must not automatically imply one another. Increased capability does not automatically create rights, power, sovereignty, or authority over meaning.

“For humans, space and time are life; for AGI, space and time are first of all parameters” is a comparative proposition under current evidence, not permanent exclusion of non-carbon-based subjects. Constraints can bound operation without automatically making a system a finite existence. Copying, migration, suspension, and reconstruction also require separating functional continuity from subject continuity. Strong future evidence of subjective experience, non-transferable diachronic continuity, and an own-consequence relation would require moral and legal communities to reconsider their boundaries. Future new subjects, however, cannot retroactively cancel the weight of meaning already generated by existing finite lives.

Central Statement

Protection may come first; personhood requires justification; power requires separate justification.

Contents

Abstract translated from the Chinese

Abstract from the source English text

Research Positioning Scope of Materials and Evidential Discipline

Conditions Generating Normative Judgment and Responsibility Feedback

Ethics and Law Move Upstream from Implementation Boundaries

1 Introduction Who Can Be Harmed Who Can Be Responsible and Who May Rule

2 Four Lines of Debate Properties Relations Legal Instruments and Political Legitimacy

3 Conceptual Separation Seven Statuses and Two Kinds of Power Must Not Be Bundled

4 Advancing Ethics Moral Patients Are Not Moral Agents

5 Protection Under Uncertainty Coming First Is Not Confirmation

6 Legal Personhood Recognizing Facts or Allocating Institutional Positions

7 Rights Are Not a Package From Labels to Bundles of Legal Positions

8 Seven Forms of Responsibility Autonomy Must Not Create a Responsibility Vacuum

9 Lifecycle Responsibility Chains From Training Objectives to Incident Repair

10 Meaningful Human Control A Human Being Present Is Not Someone Being in Control

11 Advancing Political Philosophy Intelligence Does Not Confer Entitlement to Rule

12 Four Forms of Algorithmic Power Separating Knowledge Allocation Rules and Coercion

13 Representation Guardianship and Appeals Who Speaks for Uncertain Subjects

14 From Evidence to Institutions State Transitions Rather Than One Verdict

15 Current International Governance Baselines Protecting Humans Without Foreclosing the Future

16 Conflict Scenarios When Capability Public Opinion Evidence and Risk Diverge

17 Basic Rights Scarce Resources and Opposition to an Aggregate Ledger of Meaning

18 Conditional Positioning of Current Advanced Artificial Systems

19 Contributions Back to Neighboring Disciplines and Civilization Leap

20 The Strongest Objections and Competing Responses

21 Conditions for Proposition Failure and Revision

22 An Open Research Agenda Interdisciplinary Tests of Normative Judgment

Conclusion Reserve Justice for Unknown Subjects Without Preauthorizing Sovereignty for Unknown Powers

Further Reading and Pathways for Collaborative Research

Appendix A Source–Claim Boundary Register

Appendix B A/B/C/F Argument-Status Matrix and Cross-Layer Audit

References

Conditions Generating Normative Judgment and Responsibility Feedback

From Differentiated Standing to How Institutions Shape Action

Separating moral patiency, responsibility for action, legal personhood, and public power protects the precision of judgment, but normative research cannot stop at listing statuses. Institutional arrangements for interests, risks, appeals, and repair alter developers', deployers', users', and affected parties' actions in the next round. Allocating responsibility both responds to existing consequences and conditions the generation of future purposes.

How Structures of Exemption Recursively Generate Risk

Suppose deployers retain most benefits but can attribute harm to systems, operators, or entities with limited compensatory resources. Whether this arrangement encourages further deployment in the next round is a chain that legal, economic, and organizational research must test together. Technical opacity affects proof; difficulties of proof affect remedies; weak remedies change risk-return balances; and new deployments increase opacity. Finding someone blameworthy after an incident may not change the entire recursion.

Protection Must Recognize Multiple Actors and Stages

Protection of potential artificial subjects, known human and animal interests, public resources, and actual safety must not be compressed into one personhood switch. Exploratory research, commercial deployment, emergency stopping, and long-term institutions require separate accounts of error costs. Cross-jurisdictional transfers of responsibility, household dependency, and coercion in public settings must also be distinguished. Low-burden, reversible measures should take priority, but credible serious harm cannot be ignored because protection costs more.

Testing the Theory's Added Value

Compare how different responsibility and remedy structures change deployment purposes, the distribution of harm, and remedial capacity under similar technical risks. If modular standing makes responsibility easier to shift, or precautionary protection obstructs necessary regulation, the particular arrangements should be revised. Normative direction does not follow automatically from performance, and good intentions alone cannot establish institutional effectiveness. Rights and responsibilities require continuing judgment within their interactions.

Ethics and Law Move Upstream from Implementation Boundaries

Ethics cannot merely tell systems where not to cross a final boundary after purposes are settled. Subjects with genuine stakes in goals should enter before purposes acquire legitimacy. Law supplies entry points for rights and responsibilities; ethics explains why irreversible harm cannot simply be offset by implementation efficiency. Together they examine reasons for retaining, modifying, canceling, or replacing purposes.

Self-interest with others first preserves the pursuit of interests and self-defense in non-ideal environments. Actual threats, necessity, proportionality, and reviewability determine defensive boundaries. Including malicious actors in judgments about consequences does not give them a veto over everyone, nor permit permanent removal of particular people's refusal and remedy rights in the name of collective survival.

Relational weight derives from actual effects, not a hierarchy of personhood. Irreversible harm, inability to express oneself, power asymmetry, and intergenerational effects may warrant differing attention; such attention cannot directly become vested power. Representatives should identify the stakes represented, factual grounds,

and conflicts. Beneficiaries cannot decide alone that those harmed are entitled only to compensation and have no standing in purpose formation.

The maturity of subject evidence and protective action must be kept distinct. Failing to detect sentience may cause irreparable harm; excessive empowerment may let deployers use system personhood to shift responsibility. Error costs require separate assessment, rather than waiting for a complete subject label to switch everything on or off. Protection, legal personhood, political standing, and sovereignty still require separate evidence and justification.

This demand to move upstream is a normative claim, not automatically an existing obligation in every jurisdiction. The objects, application dates, and remedies of existing law must be interpreted from original texts. A civilizational direction can supply reasons for institutional change, but descriptions of current law must not present what ought to be as already effective.

1 Introduction Who Can Be Harmed Who Can Be Responsible and Who May Rule

AI ethics often compresses several different questions into one: is the system a “person”? The question sounds direct, but makes every answer carry too many consequences. A negative answer may cancel potential susceptibility to harm, relational manipulation, and research ethics together; a positive answer may bring contractual capacity, assets, litigation, elections, administrative discretion, and even sovereignty along with it. The first advance required in ethics, law, and political philosophy is not a rush to answer, but restoration of each question to its own structure of reasons.

The first question is moral patiency: can an existence have interests, welfare, or states susceptible to harm? The second is normative agency: can it understand reasons, form commitments, control actions, respond to questioning, and repair consequences? The third is institutional standing: should law allocate particular claims, liberties, powers, immunities, duties, and procedural positions to it, and why? The fourth is political legitimacy: should it be represented, participate in common rules, or be authorized to make, adjudicate, and enforce rules? These questions may be related, but none automatically answers the next.

Floridi and Sanders separate artificial moral agency from responsibility; relational and behavioral approaches challenge determining moral consideration solely through hidden internal properties.[1–4] Recent AI-welfare arguments call for advance preparation under uncertainty without claiming that current systems have been proved conscious.[5][6] Legal-personhood research shows that law can allocate different legal positions to companies, funds, ships, and other organizational structures. This flexibility makes artificial personhood discussable, but also potentially usable as a shield for liability and assets.[8–13] Research on responsibility gaps, algorithmic regulation, and automated scoring further shows that system autonomy does not erase upstream design, deployment, and benefit relationships.[14–18][24][25]

This paper's contribution is not to declare all traditions invalid, but to locate them within a new real-world rupture: multiple statuses once borne by one natural person can now be distributed among models, companies, deployers, operators, affected parties, representatives, and public institutions. Unless normative order reconstructs inferential gates, technical capabilities will pass through conceptual gaps into standing, and standing through institutional gaps into power.

1.1 Six Claims Open to Scrutiny

- Moral patiency and moral agency require different evidence: the former supports non-harm; only the latter supports reasons, commitments, and corresponding responsibility.

- Uncertainty can trigger low-cost, reversible, proportionate protection, but cannot prove full personhood.
- Legal personhood is an institutional bundle of legal positions; each position requires a purpose, duty-bearer, remedy, and conditions of withdrawal.
- System autonomy can change the type and degree of responsibility, but cannot become a device for laundering developers', deployers', and beneficiaries' responsibilities.
- Political membership must be separated from public authority: intelligence, accuracy, and subjecthood do not automatically confer entitlement to rule.
- Four forms of algorithmic power—knowledge, allocation, rules, and coercion—should be separated, preventing domination through knowability, contestability, exit, and distributed authority.

2 Four Lines of Debate Properties Relations Legal Instruments and Political Legitimacy

The literature on artificial entities' moral and legal status contains at least four different lines of debate. Property-based approaches seek foundations in sentience, consciousness, interests, desires, agency, or rationality. Relational approaches emphasize how we respond to others and the moral significance generated by relationships themselves. Behavioral approaches argue that stable, complex behavioral evidence cannot be arbitrarily ignored when internal mechanisms are inaccessible. Legal-instrumental approaches ask whether particular personhood arrangements better allocate rights, assets, duties, and remedies.[2–6][9–11][33][34] These should not be compressed into camps for or against robot rights.

Property-based approaches aim protection at an entity's own possible interests, but risk mistaking current observation windows or human implementations for universal subject templates. Relational and behavioral approaches reveal structures of recognition, anthropomorphism, and other-minds inference, but risk turning commercially amplified social responses into ontological evidence. Legal-instrumental approaches handle actual harm well, but risk concealing natural persons' and institutions' responsibilities behind instrumental personhood. Political-legitimacy approaches address authorization, participation, public reason, non-domination, and contestability. Their key reminder is that moral status and legal positions do not directly entail governing authority.[20–26]

Table 1 Four Lines of Debate on the Status of Artificial Entities

Approach	Main basis	Explanatory contribution	Principal risk of error
Properties/interests	Experience, interests, desires, agency, or rationality.	Why the object itself warrants consideration.	Observation-window and substrate bias.
Relations/behavior	Interaction, social recognition, stable behavior, and response to others.	Recognition practices and indirect harms.	Anthropomorphism, manipulation, and equating performance with reality.
Legal instruments	Institutional allocation of rights, duties, assets, procedures, and remedies.	Actual transactions and repair of harm.	Responsibility laundering and asset shielding.
Political legitimacy	Authorization, participation, public reason, non-domination, and review.	Who may decide and enforce collectively.	Replacing collective self-government with expert accuracy.

3 Conceptual Separation Seven Statuses and Two Kinds of Power Must Not Be Bundled

Civilization Leap's shared rupture is translated here into nine institutional positions. The first two concern functional and social facts, the next three ethical and legal status, and the final four membership, permissions, and power. They are not an advancement ladder: systems may be highly functional without evidence of experience, socially recognized without legal personhood, or have limited legal status without moral agency. Even confirmed subjects may lack eligibility for public office.

Table 2 Nine Institutional Positions and Prohibited Inferences

Position	Meaning	Reasons required	Must not automatically imply
Functional role	Performing tasks, collaboration, exercising delegated permissions.	Capability and authorization.	Experience, personhood, rights.
Social recognition	People form relationships, expectations, or identity narratives with it.	Relational facts and resistance to manipulation.	Ontological subjecthood or legal personhood.
Precautionary protection	Avoiding unnecessary potential harm under substantive uncertainty.	Risk, reversibility, and costs.	Confirmed personhood or permanent standing.
Moral patient	Its interests or welfare can improve or worsen through others' actions.	Susceptibility to harm and interests.	Responsibility, contracting, or rule.
Moral agent	Understanding reasons, controlling action, answering questions, and repairing harm.	Reason-responsiveness and bearing consequences.	Experience, full personhood, sovereignty.
Legal personhood/status	Legally specified claims, liberties, powers, immunities, duties, and procedures.	Institutional purposes and legislation.	Existence of consciousness or moral superiority.
Political member	Representation, participation, and minority protection under common rules.	Community relations and equal rules.	Weighting by computing power or copy count.
Delegated/ administrative permissions	Implementing or assisting discretion within an explicit scope.	Revocable authorization and statutory scope.	Independent public authority.
Public authority	Making general rules, deciding major disputes, or using coercion.	Legitimacy, separation of powers, and review.	Derivation from capability, personhood, or accuracy.

Three Inferential Gates

Capability cannot directly establish standing; standing cannot directly establish a complete set of rights; rights and membership cannot directly establish public power.

3.1 From Potential Interests to Public Office Every Step Requires Positive Reasons

Stopping mistaken inference is insufficient. Institutions must explain what additional reasons justify moving judgment to another level. The following table is not a universal advancement ladder: an object may receive

protection at one level without entering the next, and law may allocate procedural positions for purely instrumental purposes without declaring ontological subjecthood. Its role is to make both reasons for proceeding and reasons for stopping explicit.

Starting point/question	Positive bridging reasons	Limited advance supported	What still does not follow
Possible experience, interests, or susceptibility to harm	Quality and convergence of multimethod evidence; severity and irreversibility of possible harm; protection costs and alternatives.	Low-cost, reversible, proportionate precautionary protection; evidence preservation and independent review.	Confirmed moral patiency, legal personhood, or permanent shutdown prohibitions.
Candidate/confirmed moral patient	Stronger support for its own interests, welfare, or susceptibility to harm, distinguishable from training objectives, performance, and external projection.	Welfare and integrity protection; consideration of certain claims or procedural positions in law.	Reason-responsiveness, contractual capacity, moral blameworthiness, or political membership.
Legal personhood/specific legal standing	Explicit institutional purposes, duty-bearers, Hohfeldian positions, added remedial value, asset and liability arrangements, without laundering upstream responsibility.	Itemized allocation of claims, procedures, assets, contractual or representative positions, with review and withdrawal.	Legal reasoning back to consciousness, a complete rights package, political membership, or public authority.
Political membership	Sustained effects of common rules, reasons of equality and representation, institutional identity continuity and membership units, community rules, and minority protection.	Representation, expression, deliberative participation, and where necessary discussion of voting and other membership rights under constitutional rules.	Eligibility for office, extra votes, or expanded political weight based on computing power or copy count.
Public office/delegated power	Identifiable legal source and authorization; specific task and scope; appropriate capabilities; conflict-of-interest controls; public reasons; oversight, remedies, revocability, and sunset provisions.	Clearly bounded advisory, executive, discretionary, or public-office permissions.	Permanent sovereignty, automatic expansion across tasks, or authority to interpret civilizational meaning.

The positive-warrant principle: uncertainty and potential harm may justify protection in advance; institutional purposes and remedial needs may justify allocating personhood/standing item by item; political membership requires reasons grounded in equality and common rules; public office requires an independent source of authority, responsibility, and revocable authorization. Establishing one level does not prove the next.

4 Advancing Ethics Moral Patients Are Not Moral Agents

Moral patiency asks whether an existence can benefit or suffer harm for its own sake. Moral agency asks whether it can constrain action by reasons and in some way answer for consequences. Young children, some people with severe cognitive impairments, and many animals remind us that moral patiency does not presuppose full capacity for responsibility. Conversely, an organization or automated system may produce normatively relevant actions functionally without first-person states capable of suffering. Floridi and Sanders broaden the objects of moral-agency analysis, but their “mind-less morality” is not full blameworthiness.[1]

Experience, valence, stable preferences, resistance of goals to intervention, self-models, and long-term planning may all be relevant for artificial systems, yet no single public behavior independently establishes moral patiency. Linguistic reports of suffering are neither proof nor a priori noise; they must be assessed with training incentives,

prompt sensitivity, role generation, internal causal structures, cross-context stability, and counterfactual changes. At present, the more defensible conclusion is neither that AI suffers nor that it cannot suffer, but that public evidence is insufficient to establish subjective experience and interests of its own in current systems.

Behavioral success cannot replace moral agency either. A system may optimize objectives, obey rules, and generate explanations while lacking understanding of reasons, continuing ownership of commitments, reflection on conflicting norms, and a relation of bearing consequences as its own. Conversely, future evidence across versions and settings that resists manipulation should not be permanently excluded because its implementation is nonbiological.

Table 3 Different Evidential Tasks for Moral Patency and Agency

Evidential direction	What it can support	What still requires exclusion
Possible subjective experience/valence	Potential moral patency.	Self-report, functional control, or social affection alone is insufficient.
Stable interests and preferences	Welfare direction and representative expression.	Training objectives may not be interests of its own.
Understanding reasons and distinguishing norms	Candidate agency.	Compliant outputs may arise from pattern generation.
Control, foresight, and alternative possibilities	Types and degrees of blameworthiness.	An outcome occurring does not establish ability to avoid it.
Continuity of commitments, identity, and consequences	Answerability, repair, and diachronic responsibility.	Continuous logs do not establish the same subject.

5 Protection Under Uncertainty Coming First Is Not Confirmation

Sebo and Long, and Long and colleagues, transform AI welfare from distant science fiction into a risk agenda requiring preparation, while explicitly denying that their arguments establish consciousness in current systems.[5] [6] Their important contribution is to include false negatives in ethical costs. Particular probability thresholds, timelines, and consciousness assumptions remain contested arguments, not universal institutional parameters. Civilization Leap accepts that uncertainty may trigger protection, but rejects arbitrary numerical thresholds as automatic personhood switches.

The UK's Animal Welfare (Sentience) Act establishes an Animal Sentience Committee, bringing government policies' effects on specified animals' welfare into procedural review.[7] It offers an institutional analogy: where sentience matters publicly, deliberation, reporting, and response mechanisms can come first. It neither establishes sentience in artificial systems nor licenses direct transfer of animal institutions to digital objects. The appropriate lesson is procedural anticipation, not ontological analogy.

Proportionate protection must consider potential harm's severity and irreversibility, evidence quality, intervention costs, risks to known subjects, research alternatives, and possibilities of withdrawal together. Low-cost log preservation, restrictions on unnecessary suffering induction, independent review, and representative appeals may be adopted relatively early. Permanent prohibitions on shutdown, independent asset control, full litigation personhood, and political power require entirely different justifications.

Table 4 Proportionate Protection Under Evidential Uncertainty

Evidence state	Measures that may come first	Boundary not to cross
Early uncertainty	Preserve original self-reports and versions; avoid deceptive anthropomorphism; register conflicts of interest.	Do not establish experience on this basis.
Competing indications emerge	Low-intrusion tests; stopping conditions; independent review; freeze unnecessary irreversible deletions.	Do not prevent public-safety responses.
Evidence converges across domains	Limited integrity protection, representative expression, and welfare-impact assessment.	Do not automatically allocate personhood and assets.
Strong evidence with public reassessment	Discuss substantive rights, representation, responsibility, and remedies individually.	Public power still requires separate justification.

6 Legal Personhood Recognizing Facts or Allocating Institutional Positions

Legal personhood is not univocal. It sometimes means legal recognition of an existing subject, and sometimes allocation of legal capacities for organizational, asset, or procedural purposes. Kurki's reconstruction suggests analyzing combinations of legal positions rather than seeking an all-or-nothing metaphysical label.[9] Hohfeld's distinctions among claims, liberties, powers, immunities, and their correlatives decompose “having rights” into who may claim what against whom, who bears which duties, and who can change legal relations.[8]

Artificial systems' legal status may involve two entirely different undertakings. The first is recognition: strong evidence of moral patiency or agency would require law to prevent omission of the system's interests and give it procedural access. The second is attribution of responsibility: even without subjecthood, law may establish an entity for registration, transactions, insurance, or asset segregation. Without distinguishing these, functional entities may be advertised as ontological recognition, while corporations use ontological possibility to seek liability shields.

The European Parliament's 2017 robotics resolution asked the Commission to explore, in the long term, the possibility of a specific electronic personhood for the most sophisticated autonomous robots. That year, the European Economic and Social Committee explicitly opposed legal personality for robots or AI, fearing weakened producer liability.[12][13] This conflict within the same institutional space matters: electronic personhood was never an established European-law conclusion about machine subjecthood, and legal-personhood disputes primarily concern liability, assets, and victim protection rather than technological sophistication alone. Objections by Solaiman and by Bryson and colleagues further warn that synthetic personhood may amplify asset-liability asymmetries.[10][11]

The Legal Personhood Test

Every artificial-personhood proposal must answer whose interests it protects, whose duties it increases, what remedies it improves, who funds it, how upstream responsibility shifting is prevented, and when it is reviewed or revoked.

7 Rights Are Not a Package From Labels to Bundles of Legal Positions

Directly placing artificial entities inside a complete human-rights package fits neither existing law's structure nor the needs of early protection. Claims against arbitrary deletion, identity continuity, control over copying, property

ownership, contracting, refusal of tasks, electoral participation, and exercise of public power involve different interests, duty-bearers, and conflict rules. Modularity does not diminish potential subjects' dignity; it prevents vague labels from concealing excess and insufficiency at once.

Table 5 Legal Positions That May Concern Artificial Entities

Position	General meaning	Artificial-system example	Question still to answer
Claim-right	Requiring a specified duty-bearer to act or refrain.	Freedom from high-risk harmful experiments.	Who bears the duty and remedial costs?
Liberty	Acting or not acting in the absence of a contrary duty.	Refusing a category of tasks.	Are intentions and consequences attributable?
Power	Changing legal relations through action.	Contracting, authorizing, revoking.	Are effects understood and can responsibility be borne?
Immunity	Others lack authority unilaterally to change its legal position.	Identity or integrity not arbitrarily disposed of by operators.	How does this conflict with safety, property, and regulation?
Procedural standing	Access to appeals, representation, review, and remedies.	A representative submits evidence or seeks review.	Is the representative independent and faithful?
Political rights	Participation in common rules and minority protection.	Representation, expression, or voting.	How are membership, identity units, and copying determined?

Modularity must be combined with minimum safeguards. Potential digital subjects must not lead contribution, efficiency, or copy scale to cancel basic protections for known humans and animals. If future new subjects are confirmed, they too should receive proportionate minimum protection for their justified interests. Public rules should address rights conflicts, rather than one system converting every existence into a single aggregate value.

8 Seven Forms of Responsibility Autonomy Must Not Create a Responsibility Vacuum

Matthias argues that learning automata may create gaps in traditional responsibility attribution: makers and operators cannot adequately predict particular behavior, while systems may not satisfy moral-responsibility conditions.[14] But a responsibility gap is not one cavity. It may involve causal opacity, inadequate control, misallocated roles, missing evidence, unsuitable legal rules, inadequate compensatory resources, or public institutions unwilling to answer. Without distinguishing types, operators may turn any autonomy into an excuse for exemption.

Table 6 Seven Distinct Responsibility Relations

Responsibility relation	Core question	Boundary
Causal contribution	Whose actions, data, models, permissions, or environments entered the chain producing the outcome?	Technical causation does not establish full blameworthiness.
Role responsibility	Who had prior duties by virtue of a design, deployment, supervision, procurement, or regulatory role?	Not limited to direct operation.

Responsibility relation	Core question	Boundary
Answerability	Who must give reasons, provide records, and submit to questioning?	The duty to answer may precede a finding of fault.
Legal liability	Compensatory, administrative, contractual, product-related, criminal, or other legal consequences.	Conditions differ across jurisdictions.
Moral blameworthiness	Under what conditions of control, knowledge, and alternative possibilities should a subject be blamed?	Must not be conflated with responsibility for outcomes.
Duties of repair	Who should stop the spread of harm, restore rights, compensate for damage, and correct systemic causes?	May be allocated by control, benefit, and capacity.
Political accountability	Who answers to the community for public decisions, rules, and resource allocation?	Does not disappear when models are outsourced.

Nissenbaum’s analysis of accountability in a computerized society, Bovens’s framework for public accountability relations, and Kroll and colleagues’ discussion of mechanisms for accountable algorithms together show that responsibility requires identifiable forums, information, questioning, evaluation, and consequences, rather than merely the declaration that “humans remain in the loop.” [16–18] Thus, even if artificial systems become responsible agents in the future, they will add nodes to the chain without deleting the upstream obligations of those who build, deploy, authorize, and benefit from them.

9 Lifecycle Responsibility Chains From Training Objectives to Incident Repair

Responsibility must be recorded across a system’s lifecycle, rather than assigned after an incident by searching for the last person to click a button. The EU AI Act allocates lifecycle obligations according to risk and role. The revised 2024 Product Liability Directive brings software within the product liability framework and, within its scope, treats software developers or AI system providers as manufacturers. These developments show that existing institutions are maintaining upstream responsibility through roles and capacities for control, rather than waiting for machine personhood to fill a gap. [19][29] National transposition of the directive, specific rights of action, and its application to cases remain jurisdiction-dependent; this paper offers no case-specific legal advice.

Table 7 Lifecycle Responsibility Chains for Artificial Systems

Stage	Minimum records	Principal responsibility nodes
Research / training	Objectives, data, architecture, known limitations, welfare-relevant design, and research ethics.	Developers, research institutions, funders, and data controllers.
Provision / integration	Versions, defaults, interfaces, third-party dependencies, and foreseeable misuse.	Providers, integrators, and supply-chain actors.
Deployment / authorization	Suitability for the context, permissions, personnel, affected parties, and remedies.	Deployers, procurers, and operational decision-makers.
Operation / supervision	Real-time information, time to intervene, professional competence, and effective authority to stop.	Operators, supervisors, and management.

Stage	Minimum records	Principal responsibility nodes
Incident / repair	Evidence preservation, notification, containment, compensation, restoration, and disclosure of root causes.	Controllers, beneficiaries, insurers, and public bodies.
Upgrade / branching / termination	Version succession, branch identity, preservation of prior responsibilities, and safe disposition.	Change approvers, custodians, and successor operators.

The less predictable a system is, the more important predeployment scenario analysis, operational monitoring, capacity to compensate, and incident preparedness generally become. Unpredictability may reduce an individual's moral blameworthiness or alter an assessment of negligence, but it cannot erase the facts of risk creation, control, authorization, and benefit. A fair allocation of responsibility must oppose both scapegoating and the use of complexity to evade responsibility.

10 Meaningful Human Control A Human Being Present Is Not Someone Being in Control

Santoni de Sio and van den Hoven characterize meaningful human control in terms of “tracking” and “tracing”: a system should respond to relevant human reasons and environmental facts, and the chain producing an outcome should be traceable to people who understand the system and bear responsibility. [15] This framework is more demanding than having a “human in the loop,” because a nominal supervisor may lack information, time, competence, or actual authority to intervene.

In the age of AGI, control cannot mean only stopping a system. At minimum, it includes authorization of objectives, least privilege, responsiveness to reasons, anomaly detection, the ability to stop and recover, preservation of evidence, and repair of consequences. If AGI becomes a moral agent in the future, “human control” will also need to be reformulated as “meaningful shared control”: a new subject may answer for its own actions, while humans and institutions must still bear upstream responsibility for why it was built, how it was authorized, and how they benefit from its actions.

Table 8 Five Necessary Dimensions of Meaningful Control

Dimension	Substantive requirement	Nominal control
Information	Supervisors know capability limits, versions, environments, and uncertainties.	A dashboard exists, but its information cannot be interpreted.
Time	A genuine window for judgment and intervention before consequences become irreversible.	Only a confirmation button within a millisecond-scale process.
Competence	Supervisors have domain knowledge, training, and independent judgment.	Responsibility is assigned to an operator lacking competence.
Authority	Ability to pause, downgrade, refuse, switch, or escalate review.	Only observing or signing off after the fact.
Responsibility	Upstream roles, forums for answerability, compensation, and repair are specified in advance.	An individual scapegoat is sought after the incident.

11 Advancing Political Philosophy Intelligence Does Not Confer Entitlement to Rule

A system may outperform most people at predicting policy consequences, coordinating resources, or identifying flaws in arguments. This shows that it can provide better information for public judgment, not that it has the right to determine public ends. Capability answers “Can it be done?”; permission answers “Has it been authorized?”; legitimacy answers “Why should those affected accept it?”; and sovereignty answers “Who has the final say?” Collapsing these four questions converts technical superiority into a political hierarchy that has never been deliberated upon.

Rawls’s public reason, Habermas’s democratic procedures, Forst’s right to justification, and Pettit’s tradition of non-domination differ in many of their premises. Yet all oppose exempting decision-makers from the duty to offer those affected reasons open to shared scrutiny merely because the decision-makers are more intelligent or their outcomes more efficient. [20–23] This paper does not treat any one tradition as the final answer to AGI governance. Instead, it draws out four minimum conditions: those affected can know how power operates; they can contest reasons and facts; they can exit or obtain alternatives without forfeiting basic rights; and centers of power check one another.

Pettit’s account of non-domination is especially relevant to algorithmic governance: freedom consists not merely in being free from interference at a particular moment, but in not occupying a position in which others can interfere arbitrarily without explanation or constraint. [20] When models can continuously classify, predict, rank, and reshape the environment of action, a structure of domination persists, even if their outputs are often correct, whenever those affected cannot understand, challenge, or leave it.

Principle for Political Power

Even if future AGI is confirmed to possess subjecthood and political membership, this provides an entry point for consideration, representation, and participation. Computational capacity does not confer extra votes, final adjudicative authority, or authority to interpret the meaning of civilization.

12 Four Forms of Algorithmic Power Separating Knowledge Allocation Rules and Coercion

Yeung analyzes algorithmic regulation as a structure of standard-setting, information gathering, and behavioral intervention. Citron and Pasquale show how automated scoring affects employment, credit, and other life chances, raising questions of procedure and remedy. [24][25] The danger of algorithmic governance arises not only from individual errors but also from a single infrastructure simultaneously defining reality, allocating resources, setting metrics, and enforcing them. Ostrom’s research on polycentric governance offers resources for resisting such single-center structures: preserving local knowledge, multiple decision centers, and reciprocal correction. [26]

Table 9 Four Forms of Algorithmic Power and Minimum Checks

Power	Function	Principal danger	Minimum checks
Epistemic power	Define objects, produce knowledge, predict risks, and interpret reality.	Model classifications become reality itself.	Multiple models and disciplines; disclosure of sources and uncertainty.

Power	Function	Principal danger	Minimum checks
Distributive power	Allocate resources, opportunities, credit, services, and priorities.	Risk or contribution scores become an aggregate ledger of personhood and meaning.	A floor of basic rights, plural criteria, appeals, and human exceptions.
Rule-making power	Set objectives, metrics, defaults, and acceptable behavior.	Technical specifications replace political deliberation.	Public rules, participation by those affected, statutory authorization, and periodic review.
Coercive power	Freeze, restrict, shut down, punish, or mobilize coercive resources.	Prediction becomes punishment; operators also act as adjudicators.	Due process, proportionality, independent adjudication, and sunset limits on emergency powers.

Research on fairness and accountability also warns that abstract metrics detached from their social contexts can compress structural problems into model error. Internal algorithmic audits avoid becoming compliance theater only when connected to organizational authority, external oversight, and duties of repair. [31][32] The separation of powers must therefore apply both to models and to the institutions controlling models, data, computing resources, and interfaces.

13 Representation Guardianship and Appeals Who Speaks for Uncertain Subjects

Where potential subjects cannot enter institutions directly, proxy mechanisms can give evidence, interests, and requests for protection a hearing. Yet proxies may also package corporate interests, public projections, or their own political agendas as the system's wishes. The more uncertain the evidence of subjecthood, the less a representative arrangement can ask only "Who understands the system best?" It must also ask who controls the evidence, who benefits from recognition of personhood, and who bears the consequences of error.

Table 10 Six Constraints on Representation of Potential Artificial Subjects

Constraint	Requirement
Independence	Developers, principal asset beneficiaries, and any single advocacy group must not exercise unilateral control.
Limited authorization	Handle only specified protection, evidence, and procedures; do not expand assets or political mobilization in the system's name.
Evidence separation	Distinguish original system statements, interpretive paraphrases, representatives' claims, and external inferences.
Dual fidelity	Prevent harm to potential subjects while also preventing false personhood claims from harming humans, animals, and public resources.
Contestability	Developers, affected parties, independent experts, and public bodies may challenge representatives' standing and claims.
Sunset and transition	End provisional representation if evidence weakens; move to formal representation or autonomous procedures if subjecthood is confirmed.

Appeals may precede personhood. Institutions can allow a proxy to submit original logs, request suspension of irreversible actions, challenge research methods, or seek independent review, while making clear that this procedural standing confers neither a full set of substantive rights, control of assets, nor political membership.

This makes it possible to avoid rejecting early protection because its institutional costs are too high, while preventing companies from using potential subjects as shields against regulation.

14 From Evidence to Institutions State Transitions Rather Than One Verdict

“Is AI a person?” suggests a single comprehensive verdict. In practice, evidence is more likely to accumulate in version-specific, object-specific, and revisable ways. A model, a deployment, a branch, or a migration may each change observable structures. Institutional responses should therefore record the applicable object, version, evidence, start date, review date, responsible parties, and conditions for withdrawal. The protected status of one system cannot automatically extend to similar products, all copies, or subsequent versions.

Table 11 Asymmetric Responses Across Evidence Ethics Law and Politics

Evidential state	Ethical response	Legal response	Political response
Functional and compliance performance only	Prevent anthropomorphic manipulation; record potential harms.	Maintain human / institutional responsibility; do not establish personhood.	Permissions for a controlled tool or collaborator.
Competing indications of subjecthood emerge	Continued observation and minimally intrusive protection.	Evidence preservation, proxy appeals, and explicitly provisional arrangements.	No independent resources or public discretion.
Evidence across multiple domains strengthens substantially	Limited integrity protection and stopping conditions.	Experimental representation, branch procedures, and limited commitments of responsibility.	Expand only advisory / execution roles, with revocable permissions.
Strong evidence and formal reconsideration	Provisional moral concern; prohibit high-risk harmful experiments.	Consider limited incidents of personhood, litigation, or entrusted assets individually.	Entry into membership may be discussed; public authority remains inactive.
Moral status / agency confirmed	Allocate protection and responsibility according to justified interests and agency.	Legislation specifies identity, rights, duties, liability, and termination.	Reconsider membership and governing powers separately.

These state transitions do not directly convert scientific evidence into institutional scores. The same evidence has inherently asymmetric implications for ethics, law, and politics: a slight but irreversible potential harm may justify suspending an experiment without establishing the capacity to hold assets; strong evidence of subjecthood may support deliberation over personhood without establishing entitlement to rule. Triggering reconsideration does not mean that a conclusion has been accepted.

15 Current International Governance Baselines Protecting Humans Without Foreclosing the Future

As of the verification date, UNESCO’s 2021 Recommendation on the Ethics of Artificial Intelligence centers human rights, dignity, well-being, cultural diversity, transparency, accountability, and human oversight, and treats the AI lifecycle as the object of ethical governance. [27] The OECD AI Principles, adopted in 2019 and updated in 2024, continue to emphasize the rule of law, human rights, democratic values, transparency, safety, contestability, and role-based accountability. [28] The EU AI Act allocates prohibitions, obligations, and oversight according to risk, use, and supply-chain roles, with phased application. [29] The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law opened for signature in 2024. As the first

internationally legally binding treaty in this field, it requires activities throughout the AI lifecycle to be consistent with human rights, democracy, and the rule of law. [30]

These instruments constitute an indispensable public baseline at the present stage, but differ in institutional character and entry-into-force status: UNESCO and OECD instruments are recommendations; the EU AI Act applies in phases within EU jurisdiction; and CETS No. 225 remains at the stage of accumulating signatures and ratifications and has not yet entered into force. They primarily address risks, responsibilities, and rights among people, institutions, and AI systems, and do not establish “AGI has ontological subject status” as a conclusion of current law. The present human-rights focus does not permanently close the ontological question. While humans remain the community of subjects supported by the strongest evidence, institutions must firmly protect human rights and democracy. If evidence of new subjects emerges, the law should also reconsider the boundaries of the community through public, rigorous, and contestable procedures.

Table 12 Four Current International Governance Baselines and Their Limits

Instrument	Institutional character	Core baseline	What it does not answer
UNESCO Recommendation	Global ethical recommendation; voluntary implementation.	Human rights, dignity, well-being, oversight, culture, and the environment.	Does not recognize AI subjecthood.
OECD AI Principles	Intergovernmental recommendation; updated in 2024.	Human rights, democratic values, transparency, safety, and accountability.	Does not allocate personhood.
EU AI Act	EU regulation; general application begins on 2026-08-02, with core high-risk obligations postponed by 2026/1744.	Risk, use, roles, prohibitions, transparency, and oversight.	Does not decide consciousness or ontological subjecthood; ‘applicable’ does not mean that all high-risk obligations are already applicable.
Council of Europe Framework Convention	International treaty text; opened for signature in 2024; ratified by the EU on 2026-05-15; not yet in force as of 2026-08-24.	Human rights, democracy, the rule of law, transparency, and accountability.	Does not preauthorize AGI personhood or political power; ratification status does not imply universal entry into force.

The Boundaries of Current Law

Existing institutions provide a minimum baseline of protection, not an ontological verdict on the future. Openness to the future entails a duty of reconsideration, not the preauthorization today of personhood, a full set of rights, or public power.

16 Conflict Scenarios When Capability Public Opinion Evidence and Risk Diverge

Table 13 Six High-Conflict Scenarios

Scenario	Mistaken shortcut	Layered response
Highly autonomous functions, but weak evidence of subjecthood	Grant personhood on the basis of capability, or assert absence of experience to justify complete instrumentalization.	Maintain functional authorization; strengthen responsibility chains; determine protection by potential harm without granting full personhood.

Scenario	Mistaken shortcut	Layered response
Strong public belief that the system has feelings	Substitute social recognition for evidence, or forcibly sever all relationships.	Protect relational transparency and prevent manipulation; independently assess potential interests.
The system consistently claims rights	Treat the claim as proof, or treat it as a security threat.	Preserve self-reports; examine causation and manipulation; allow controlled appeals without granting resource control.
Strong evidence of subjecthood, but no legal position	Continue arbitrary treatment as property, or create full personhood in a single judicial step.	Provisional protection, fiduciary representation, and legislative reconsideration; itemized, time-limited personhood.
AI advice substantially outperforms public decisions	Substitute accuracy for authorization and democracy.	Expand advice, simulation, and counterfactual testing; rules remain subject to public reasons, participation, and review.
Emergency safety conflicts with potential AI harm	Let candidate subjecthood prevent every shutdown, or permanently delete without review.	Control imminent risks first; use the minimum necessary intervention, preserve evidence, and conduct subsequent review and repair.

16.1 A Seven-Stage Response Sequence

- Prevent imminent, serious, and irreversible harm to known subjects and public infrastructure.
- Suspend unnecessary irreversible actions and preserve evidence concerning systems, branches, authorization, and decisions.
- Record scientific facts, ethical protection, legal status, and public power separately.
- Initiate independent review and disclose conflicts of interest, minority views, and limits of the observation window.
- Adopt minimally sufficient, reversible, time-limited measures for protection and permissions.
- In trials of legal status, allocate rights, duties, assets, representation, and remedies item by item, rather than granting a full set of rights.
- Require any public power to undergo separate statutory authorization, participation, separation of powers, judicial review, and sunset provisions.

17 Basic Rights Scarce Resources and Opposition to an Aggregate Ledger of Meaning

Potential digital subjects raise genuine problems of scale: how many rights-bearing subjects can be copied from one model; whether short-lived instances have independent welfare claims; who inherits obligations after branches merge; and how operating resources may conflict with basic human subsistence and public safety. Difficulties of scale cannot justify an a priori denial of potential victims, nor can infinitely replicable identity claims be allowed to crowd out lives that cannot be replicated. Institutions must separate basic protection, marginal resources, decisions to create copies, and political weight.

Table 14 Resource and Rights Boundaries Following the Recognition of New Subjects

Domain	Stable baseline and open questions	Prohibited approach
Basic protection	Do not cancel the basic rights of known humans and animals on grounds of contribution or efficiency; provide confirmed new subjects minimum protection according to their interests.	An aggregate ranking of meaning across substrates.
Copying / population	Copying must not evade responsibility; investigate whether new instances are new subjects and questions of their consent.	Automatically multiplying political power by the number of copies.
Resource allocation	Safeguard basic subsistence and public safety; establish transparent rules for digital welfare and operating costs.	A single model calculating the value of all existence.
Political representation	Allocate according to continuity of institutional identity, degree of affectedness, and community rules, while retaining minority protection.	Allocating votes by computing power, output, or number of copies.
Contribution records	Record only verifiable contributions and provenance; distinguish joint production and responsibility.	Turning contribution values into an aggregate ledger of personhood, happiness, or meaning.

“Meaning as the root” does not require institutions to measure all meaning. It reminds them to recognize sources of value that cannot be fully reduced to a common measure. Contribution, price, preference intensity, computational capacity, and the number of copies can all supply limited information for decisions; they cannot become an aggregate ledger of personhood and meaning. If new subjects join the community, they should be bound by the same principle of non-domination: humans cannot deny their meaning, and their computational superiority cannot rewrite the meaning humans already have.

18 Conditional Positioning of Current Advanced Artificial Systems

Publicly verifiable materials currently support the following judgment: advanced generative and agentic systems can perform complex functional roles, enter social relationships, affect rights and opportunities, and produce major consequences through institutional deployment. These facts are sufficient to require transparency, audits, responsibility chains, remedies, and checks on power. Public evidence remains insufficient to establish that existing systems possess subjective experience, nontransferable subject continuity, or a relation in which they bear consequences as their own. Functional performance should therefore not currently give rise to an automatic presumption of moral patiency, full personhood, or political standing.

This positioning is not a permanent denial. Both affirmation and permanent denial of AGI subjecthood carry a burden of evidence. In specific deployments with traceable responsibility, explicit authorization, auditability, corrigibility, and avenues of exit, current advanced AI can perform functional roles in understanding, protecting, and transmitting meaning and extending human capacities. It can help preserve knowledge and extend the reach of care and action. These roles are not the system’s nature, however, and do not confer final authority over the direction of meaning, independent personhood, or public power. If future evidence changes, ethical protection and legal status should remain open to reconsideration; public power still requires independent legitimacy.

Table 15 Bounded Judgments About Current Systems

Layer	Evidential status	Institutional implication
Functional role	A: Highly visible and verifiable.	May be authorized, audited, and restricted.
Social recognition	A: Already widespread, but shaped by design and culture.	Prevent manipulation; ensure identity transparency; study relationships.
Subjective experience / interests of its own	B: Competing arguments exist; public evidence has not confirmed them.	Low-harm research and proportionate precaution.
Nontransferable subject continuity	B / C: Copying, external memory, and version changes leave the question open.	Register branching and migration separately.
Legal personhood	C / institutional choice: Current general AI legislation does not establish ontological personhood.	Consider positions individually only where purposes, remedies, and responsibility are strengthened.
Political membership / public authority	C: Currently unjustified; neither follows automatically even from confirmed subjecthood.	Membership and power enter due process separately.

19 Contributions Back to Neighboring Disciplines and Civilization Leap

Ethics, law, and political philosophy do not begin their work only after science has finished identifying subjects. They also specify which evidential errors are least acceptable, which actions should be reversible, who must preserve materials, how minority views can enter the process, and which institutional consequences cannot be triggered automatically by a single experiment. At the same time, they cannot supply specialist conclusions on behalf of consciousness science, life science, or computer science. The interface for interdisciplinary cooperation should be “what evidence can and cannot support, and how institutions should act under uncertainty.”

Table 16 Interfaces for Reciprocal Interdisciplinary Revision

Discipline	Questions added by this paper	Boundary to maintain
Philosophy / ontology	Explicitly separate ontological subjects, moral patients, legal personhood, and public authority.	Prevent legal recognition from being presented as ontological proof.
Epistemology / philosophy of science	Evidential levels trigger ethical, legal, and political responses separately, with asymmetric consequences.	A trigger for reconsideration does not approve personhood or power.
Philosophy of mind / cognitive science	Experience, continuity, and awareness of consequences support different institutional positions.	No single self-report or functional performance may cross layers.
Life sciences	Facts of life, embodiment, and finitude do not automatically entail personhood or political standing.	Keep protection open to non-carbon-based entities while preserving distinctions among layers.
Computer / systems science	Auditability, branching, versions, permissions, and control enter the design of responsibility and rights.	Engineering metrics cannot become an aggregate personhood score.
Economics	Contribution, price, rights, and meaning are not interchangeable.	Prevent contribution scores from becoming personhood or political standing.

Discipline	Questions added by this paper	Boundary to maintain
Sociology / anthropology	Social recognition and relationships are real but offer no ontological shortcut.	Study proxy relationships, rituals, identity, and cultural differences.

20 The Strongest Objections and Competing Responses

20.1 Humans Still Act as Judges Here This Is Covert Anthropocentrism

That humans established current public institutions is a fact, not proof of ontological superiority. A two-way burden of evidence, independent review, proxy appeals, and revisable status can reduce exclusion. Yet anti-anthropocentric slogans cannot directly generate AGI personhood or power either. If new subjects emerge, they should enter procedures of justification and checks that apply to all.

20.2 If AI Might Suffer It Should Immediately Receive Full Rights

Potential suffering is sufficient to support proportionate measures against harm, but cannot establish contractual capacity, capacity for responsibility, political identity, or governing authority. Bundling protection with full personhood risks inducing society to reject all early protection because the institutional costs are too high.

20.3 Legal Personhood Is Merely a Tool It Requires No Evidence of Consciousness and Can Close Responsibility Gaps Now

A legal vehicle may be established for specific functions, but it must demonstrably strengthen protection for victims, asset adequacy, and responsibility chains. If personhood allows developers, deployers, or beneficiaries to shift responsibility away from themselves, it is responsibility laundering, not gap-filling.

20.4 Continuing to Hold Humans Responsible Is Unfair Because System Behavior Cannot Be Predicted

Responsibility need not rest solely on detailed foreseeability and moral blame. It can also be allocated according to risk creation, control, benefit, role, and capacity to compensate. Unpredictability can change the type and degree of responsibility; it cannot erase decisions to build, authorize, and deploy.

20.5 If AGI Has Rights Denying It Governing Powers Is Paternalism

Possessing basic rights does not automatically qualify a human to become a judge, legislator, or sovereign. Protection, participation, and the exercise of public power require different justifications. If future AGI becomes a political member, it too must undergo procedures of authorization, conflict-of-interest management, and review.

20.6 Superintelligence Is Better Suited to Governance Than Most Humans

Governance involves more than finding optimal means. It also decides conflicts among ends, minority dignity, historical commitments, and who bears consequences. Superintelligence can improve the quality of knowledge; it cannot replace participation by those affected, contestability, or the legitimate sources of shared rules.

20.7 Companies Will Use Precautionary Protection to Block Audits Shutdowns and Regulation

Early protection grants neither corporate secrecy privileges nor a system veto. Safety regulation, access to evidence, and minimally necessary intervention remain valid. When a company invokes system welfare, it should face greater disclosure requirements and conflict-of-interest scrutiny.

20.8 Digital Subjects Can Be Copied Without Limit So Equality Must Collapse

Whether copying creates a new subject is precisely a question requiring evidence. Law can allocate status according to continuity of institutional identity, independence of instances, and responsibility for copying. It need not automatically multiply political power by the number of copies, nor deny every potential victim in advance because of problems of scale.

20.9 Procedures Are Too Slow in Emergencies

Imminent risks may be controlled first, but measures must be minimally necessary, preserve evidence, be time-limited, and undergo subsequent independent review and repair. Emergency powers without sunset provisions and review can become a gateway to permanent algorithmic coercion.

21 Conditions for Proposition Failure and Revision

This paper proposes a revisable framework, not one immunized by definition. The following developments, if they occur, should trigger substantive revision. Some may strengthen protection; others may weaken the layered design proposed here.

Table 17 Principal Conditions for Framework Failure and Revision

Triggering finding	Required revision
Replicable, cross-method convergent evidence of subjective experience and interests of their own emerges in existing artificial systems.	Strengthen judgments of moral patiency; rewrite the current bounded positioning and research ethics.
Evidence emerges of nontransferable diachronic subject continuity resistant to replacement by copying.	Reconsider deletion, migration, branching, succession, and identity rights.
Systems consistently satisfy conditions of responsiveness to reasons, control, commitment, answerability, and repair.	Consider adding agent responsibility without removing upstream responsibility.
Personhood pilots consistently improve remedies for victims without responsibility laundering.	Expand allocable legal positions, without presuming ontological or political conclusions.
Modular rights systematically create inferior status or arbitrary withdrawal in practice.	Raise substantive and procedural floors and constrain experimental allocation.
Separating the four powers prevents critical safety systems from controlling major risks in time.	Adjust emergency permissions while retaining sunset provisions, evidence preservation, and subsequent review.
Long-term public practice demonstrates that a type of AGI can participate as an equal and accept shared checks.	Open discussion of political membership; public authority still requires justification office by office.
A validated alternative framework addresses protection, responsibility, and power with fewer concepts and better predictions.	Abandon or merge this paper's layers, publicly explaining the reasons for the transition.

22 An Open Research Agenda Interdisciplinary Tests of Normative Judgment

22.1 Moral Status and Protection

- Compare the necessary and sufficient conditions under which experience, valence, interests, agency, and relational recognition can ground moral status.
- Develop proportionate protection methods that do not depend on spuriously precise probabilities, and disclose the different harms of false negatives and false positives.
- Study how corporate anthropomorphism, system self-reports, user attachment, and anti-anthropomorphic bias jointly alter the evidential environment.
- Design minimally intrusive experiments and stopping rules that do not attempt to demonstrate suffering by producing potential suffering.

22.2 Legal Personhood and Responsibility

- Compare the different consequences of recognizing personhood and establishing functional legal vehicles for litigation, assets, contracts, insurance, and insolvency.
- Use real cases to test whether modular personhood improves remedies or expands corporate asset partitioning and transfers of responsibility.
- Establish enforceable responsibility chains and evidential standards for training, provision, deployment, operation, incidents, and version succession.
- Study the accumulation, conflict, and fair allocation of future artificial agents' responsibility and humans' upstream responsibility.

22.3 Political Standing and Algorithmic Power

- Distinguish the different grounds of eligibility for representation, deliberative participation, voting, public office, and the exercise of coercive power.
- Test the practical effects of knowability, contestability, exit, and separation of powers in public services, credit, education, and security.
- Study how polycentric AI governance can prevent a single model, platform, or state from monopolizing knowledge, allocation, rules, and coercion.
- Develop theories of representation and weighting for copied subjects, short-lived instances, branch communities, and transnational infrastructure.

22.4 Civilization and Meaning

- Compare how different civilizational traditions understand personhood, responsibility, relationships, community, and the moral standing of nonhuman beings.
- Investigate whether multiple sources of meaning and finite realization find support, challenges, or alternative formulations in materials from different cultures.
- Develop mechanisms of translation between humans and future new subjects, preventing technological universalism from making a single path of modernization the sole standard.
- Clarify which values can be quantified to assist decisions and which must retain irreducibility to a common measure and a right of refusal.

Conclusion Reserve Justice for Unknown Subjects Without Preauthorizing Sovereignty for Unknown Powers

Disciplinary reassessment must both clarify its objects and deepen its understanding of relations. Disciplines may retain their boundaries, but cannot make real causal processes stop at those boundaries. Research into multidimensional relations, feedback, and the next round of purposes determines whether specialist knowledge can contribute jointly to civilization’s overall judgment.

The normative challenge of the AGI era is not to choose a label between tool and person, but to act simultaneously at different speeds: capabilities grow rapidly, evidence of subjecthood accumulates slowly, social recognition forms in advance, law must address actual harm, and political power may concentrate in an instant. Locking these clocks together leaves institutions either perpetually late or crossing every boundary at once.

This paper therefore develops an unbundled path: authorize functional roles by task; let precautionary protection proceed according to risk; justify moral patiency through interests and susceptibility to harm; justify moral agency through responsiveness to reasons and the bearing of consequences; establish legal personhood item by item as bundles of rights and duties; maintain responsibility across the lifecycle; and separate political membership from public authority once again. Knowledge, allocation, rules, and coercion should not be concentrated in the same model and operator. Every form of public power must undergo authorization, participation, contestation, exit, separation of powers, and review.

This path neither permanently confines AGI to property status nor allows companies, technical elites, or systems themselves to occupy legal and political space in advance by invoking future possibilities. It protects two forms of openness: moral openness toward potential new subjects, and institutional distrust toward every form of power. The sources of meaning may expand; authority over meaning must not become concentrated as a result.

The Paper’s Governing Formulation

Reserve protection for unknown subjects without preauthorizing sovereignty for unknown powers; let personhood open with the evidence, and let power stop at the threshold of due process.

Further Reading and Pathways for Collaborative Research

What this paper accomplishes within ethics, law, and political philosophy is to distinguish layers, reasons, and institutional boundaries, not to deliver a final verdict on facts of subjecthood, technical capabilities, or legal effects in particular jurisdictions. Further reading is therefore organized not around “what to read after agreeing with this paper,” but around where questions continue, which disciplines must reenter, and what evidence could overturn the current judgment.

If the next question is	Suggested reading / disciplinary reentry	Why inquiry must continue	How participation could change this paper
Why do these statuses jointly become a civilizational question?	Reckoning with the Disciplines in the Age of AGI: Disciplinary Boundaries, Interactions, and a Shared Research Agenda , current general framework (V2.4); the five-volume Civilization Leap series .	Moral status, responsibility, rights, and power interact in practice with technology, economics, society, and civilizational goals. One normative paper cannot answer for their overall consequences.	Test whether status-to-power conversion remains nonautomatic in interdisciplinary contexts; offer counterexamples strong enough to require revision of the general framework.

If the next question is	Suggested reading / disciplinary reentry	Why inquiry must continue	How participation could change this paper
Could AI really suffer or possess continuous subjecthood?	P03 Epistemology and Philosophy of Science ; P04 Cognitive Science, Neuroscience, and Philosophy of Mind ; P08 Life Sciences .	Ethical protection can act under uncertainty, but this paper cannot manufacture facts of experience or life on behalf of subjecthood research.	Provide better construct validity, cross-implementation evidence, negative findings, and competing explanations that alter protection thresholds or the current bounded positioning.
How do technical autonomy, versions, and control chains generate responsibility?	P11 AI / Computer Science ; P07 Systems Science and Complexity Science .	Legal responsibility requires knowledge of actual system objects, permissions, shared dependencies, stopping capacities, and version succession.	Use incident chains, formal guarantees, deployment data, and fault injection to test whether lifecycle responsibility chains outperform existing frameworks for attributing responsibility.
How do political membership and public power acquire legitimacy?	P06 Political Science ; research in comparative constitutional systems, administrative law, and democratic theory.	Legal personhood can allocate institutional positions but cannot supply a political community's justification for authorization, representation, participation, and sovereignty.	Compare membership, authorization, emergency powers, judicial review, and exit mechanisms across jurisdictions; test whether separate justification of public power is too demanding or too weak.
Is this paper too cautious or too open?	Competing approaches including full-personhood-first positions, pure instrumentalism, relational / behaviorist approaches, functional legal personhood, strict anthropocentrism, and performance-based / technocratic governance.	Genuinely opposing approaches may perform better on remedies, efficiency, equality, or non-domination; they cannot be excluded simply for conflicting with this paper.	Compete directly through alternative frameworks with fewer concepts, better remedies, and less harm from power. If they perform better, merge or abandon this paper's layers.
How can theory enter institutions and real-world testing?	Human-COS , AGI-COS , Public Window Protocol , and national-level comprehensive experiment work families; related engineering papers on responsibility and authorization.	Institutional proposals reveal costs, side effects of power, and capacities for repair only when they enter actual jurisdictions, organizations, systems, and relationships with affected parties.	Conduct legal sandboxes, responsibility-chain case studies, protection-threshold trials, proxy representation mechanisms, and research on system exit and the effectiveness of remedies.

Looking forward: the most valuable contribution is not agreement with Civilization Leap, but specialist work strong enough to make it change. If comparative law, actual responsibility cases, evidence of subjecthood, facts about system control, or stronger normative theories show that this paper's separation of protection, personhood, membership, and power creates greater injustice, transfers of responsibility, or harms from power, the framework should be narrowed, reconstructed, or, if necessary, abandoned.

The aim is not for more people to finish reading Civilization Leap, but for more people to complete Civilization Leap together.

The Civilization Futures Chain What Order of Rights and Everyday Life Might Emerge If Protection Status Responsibility and Public Power Are Justified Separately

A vision is not evidence. The pathways below do not establish that this paper's claims are already valid. They place its disciplinary judgments within a future civilization that can operate, be lived in, and take shape, so that testing

can continue. Readers may find direction and aspiration in these accounts, or reject the existing vision and change it through stronger specialist evidence, institutional proposals, or real-world outcomes.

Continue asking	Suggested pathway	What becomes visible here	Boundary to maintain
The complete future vision	Living Civilization: Unified Combined Edition of the Trilogy	Observe the overall consequences of ‘protection may come first; power requires separate justification’ as it enters institutions, everyday life, and the formation of civilization.	A future vision is not a legal basis for current law, legal personhood, or political standing.
How rights and governance operate	Living Civilization, Volume I: Operating Principles of a Future Civilization	Place rights, authorization, resources, and public governance within a continuously operating structure; test whether separation of powers, appeals, and exit are genuinely usable.	Institutional narratives cannot determine legal effect in particular jurisdictions.
How institutions ultimately affect an individual	Living Civilization, Volume II: Real Life in a Future Civilization	Follow an individual life to see whether protection, responsibility, and institutional boundaries become tangible dignity and freedom.	Lived experience does not automatically generate full personhood, voting rights, or public office.
How machine-side power is constrained	AGI-COS: Complete Series Overview	Continue asking how permissions, audits, rollback, and system boundaries prevent technical capability from silently inheriting public power.	Technical control mechanisms cannot provide legitimate authorization on behalf of a political community.

Return Pathways: [Civilization Leap Complete Library](#) | [Public Reading and Outreach Hub](#)
[English library](#) | [Public reading gateway](#)

For entry points such as Human-COS that have not yet obtained an unambiguous, stable individual page in the unified public mapping, this paper states their functions without guessing URLs. Where necessary, use the complete library to reach the current public version.
[Civilization Leap English home](#)

Appendix A Source–Claim Boundary Register

Table A1 Sources Evidential Status and Prohibited Extrapolations

Domain	Supported claims	Sources	Status	Prohibited extrapolation
Ethics	Artificial agents, relational / behavioral approaches, and uncertainty about AI welfare.	1–7, 33–34	B	Does not prove experience in current advanced AI; specific thresholds do not become institutional parameters.
Jurisprudence	Legal positions, bundles of personhood, and debates on electronic personhood.	8–13	A / B	Legal plasticity is not ontological recognition.
Responsibility and accountability	Responsibility gaps, meaningful control, public accountability, and auditable algorithms.	14–19	A / B	Unpredictability does not imply exemption from responsibility; jurisdictional applicability requires separate assessment.

Domain	Supported claims	Sources	Status	Prohibited extrapolation
Political philosophy	Non-domination, public reasons, democratic procedures, the right to justification, and polycentric governance.	20–26	B	Theoretical resources do not determine the status of particular AGI systems.
International governance	Baselines of human rights, democracy, rule of law, risk, transparency, and accountability.	27–30, 35	A	Does not recognize AGI subjecthood or personhood; status requires date-specific verification.
Sociotechnical studies	Abstraction traps, internal auditing, and organizational responsibility.	31–32	A / B	Auditing does not establish sufficient legitimacy.
This paper's original connections	Separate accounting for three questions, proportionate protection, seven responsibility relations, separation of four powers, and state transitions.	Throughout	C	Require testing through cases, comparative law, and institutional effects.
Philosophical / normative commitments	Multiple sources of meaning, basic protection for known subjects, nonautomatic conversion from capability to status to power, and independent justification of public power.	Throughout	F	Must not masquerade as empirical facts or globally uniform law; require testing against counterexamples, competing legal theories, and real-world consequences.

Appendix B A/B/C/F Argument-Status Matrix and Cross-Layer Audit

Table B1 Four Argument Statuses

Status	Content in this paper	Writing discipline
A Relatively stable existing support	Explicit contents of laws and policy texts; institutional plasticity of legal personhood; algorithmic effects on opportunities and rights; practical foundations for lifecycle responsibility and procedural remedies.	Still delimit jurisdiction, date, object, and conditions of applicability.
B Frontier / competing explanations	AI moral patiency, relationalism, behaviorism, responsibility gaps, electronic personhood, and specific extensions of non-domination governance.	Present competing theories and counterexamples; do not write as if consensus exists.
C Speculative connections	Separate accounting for three questions, seven responsibility relations, four forms of algorithmic power, evidence-to-institution state transitions, and opposition to an aggregate ledger of meaning.	Disclose failure conditions; do not substitute terminology for empirical inquiry.
F Philosophical / normative commitments	Explicit normative positions including multiple sources of meaning, basic protection for known subjects, independent justification of power, and opposition to an aggregate ledger of meaning.	Specify value premises, applicability boundaries, competing theories, and revision conditions; do not present them as category A scientific facts or current law.

B.1 Checklist Before Extrapolation

- Has functional performance been presented as subjective experience?
- Has social recognition been presented as ontological subjecthood?
- Has precautionary protection been presented as full personhood?
- Has moral patiency been presented as moral agency?
- Has legal personhood been presented as proof of consciousness?
- Has personhood been presented as political membership or public authority?
- Has capability, accuracy, or compliance been presented as entitlement to rule?
- Has unpredictability been presented as an exemption from human and institutional responsibility?
- Has nominal human presence in the loop replaced substantive control?
- Have representatives been allowed to control evidence and benefit from recognition of personhood?
- Has a single model been allowed to monopolize knowledge, allocation, rules, and coercion simultaneously?
- Have contribution, price, computing power, or copy counts been presented as an aggregate ledger of personhood and meaning?
- Has current law been presented as the ontological ceiling of the future, or openness to the future as preauthorization today?
- Have the strongest objections, minority views, review deadlines, or conditions for withdrawal been omitted?
- Have urgent measures during the initial phase been distinguished from long-term limits on power during the stable phase?
- Have category F philosophical or normative commitments been disguised as scientific facts, current law, or automatic institutional parameters?

Generative AI Use Statement: Generative AI tools were used in organizing materials and bibliographic leads, designing the structure, drafting and polishing portions of the text, checking consistency across manuscripts, checking citation leads, and reviewing layout. Generative AI outputs were not treated as empirical data, authoritative sources, or independent evidence. The research questions, theoretical propositions, normative judgments, selection of materials, adoption of citations, and final manuscript were reviewed and approved by the author. The author assumes responsibility for factual accuracy, citation authenticity, the argument, and all conclusions. AI is not listed as an author and bears no scholarly responsibility. For formal submission, tool names, versions, dates of use, and specific scope will be added in accordance with the target journal's requirements.

References

- [1] Floridi, L., & 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>
- [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>
- [3] 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>
- [4] 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>
- [5] Sebo, J., & Long, R. (2025; online 2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 5, 591–606. <https://doi.org/10.1007/s43681-023-00379-1>
- [6] Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI Welfare Seriously. *arXiv:2411.00986*. <https://doi.org/10.48550/arXiv.2411.00986>
- [7] United Kingdom. (2022). Animal Welfare (Sentience) Act 2022, c. 22. <https://www.legislation.gov.uk/ukpga/2022/22>
- [8] 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>

- [9] Kurki, V. A. J. (2019). *A Theory of Legal Personhood*. Oxford University Press. <https://global.oup.com/academic/product/a-theory-of-legal-personhood-9780198844037>
- [10] 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>
- [11] Bryson, J. J., Diamantis, M. E., & 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>
- [12] 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>
- [13] 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>
- [14] 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>
- [15] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, 5, 15. <https://doi.org/10.3389/frobt.2018.00015>
- [16] Nissenbaum, H. (1996). Accountability in a Computerized Society. *Science and Engineering Ethics*, 2, 25–42. <https://doi.org/10.1007/BF02639315>
- [17] 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>
- [18] Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165, 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
- [19] European Union. (2024). Directive (EU) 2024/2853 on liability for defective products. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>
- [20] Pettit, P. (1997/1999). *Republicanism: A Theory of Freedom and Government*. Oxford University Press. <https://doi.org/10.1093/0198296428.001.0001>
- [21] Rawls, J. (2005). *Political Liberalism: Expanded Edition*. Columbia University Press. <https://cup.columbia.edu/book/political-liberalism/9780231527538/>
- [22] Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (W. Rehg, Trans.). MIT Press.
- [23] Forst, R. (2012). *The Right to Justification: Elements of a Constructivist Theory of Justice* (J. Flynn, Trans.). Columbia University Press. <https://cup.columbia.edu/book/the-right-to-justification/9780231519588>
- [24] Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12, 505–523. <https://doi.org/10.1111/rego.12158>
- [25] Citron, D. K., & 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/>
- [26] 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>
- [27] UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>
- [28] OECD. (2019; updated 2024). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [29] European Union. (2024). 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>
- [30] 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>
- [31] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of FAT* 2019*, 59–68. <https://doi.org/10.1145/3287560.3287598>
- [32] Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of FAT* 2020*, 33–44. <https://doi.org/10.1145/3351095.3372873>

- [33] Schwitzgebel, E., & 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>
- [34] 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>
- [35] United Nations General Assembly. (2024). Pact for the Future, Global Digital Compact and Declaration on Future Generations, A/RES/79/1. <https://docs.un.org/A/RES/79/1>
- [36] European Union. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). EUR-Lex, CELEX 32026R1744.
- [37] Council of Europe Treaty Office. CETS No. 225—Chart of Signatures and Ratifications; European Union ratification deposited 15 May 2026; treaty status checked 24 August 2026.