

Ethical Subjectivity across Human–AI and AI–AI Relations: A Generative Relational Inquiry

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

Debates about artificial intelligence often ask whether an AI system is a moral subject. This discussion paper begins from a narrower methodological concern: the expression “moral subject” can compress several different forms of standing that need not arise together. A system may exercise consequential decision power without bearing moral responsibility, may become an object of protection without possessing the capacities usually associated with moral agency, may acquire legal standing without settling questions of phenomenal experience, and may occupy positions of power before any stable ethical vocabulary exists for describing those relations. Treating these statuses as a single binary property risks obscuring the different relational histories through which they become socially and normatively significant.

Working within a Generative Relational perspective, the paper therefore distinguishes decision subjectivity, moral agency, moral patiency, legal subjectivity, and power-holding capacity, while suspending claims about the ultimate ontology of artificial subjects. The inquiry examines three directions separately: human ethical relations toward artificial agents, artificial ethical relations toward humans, and ethical relations among artificial agents. Rather than assuming that ethical standing follows directly from membership in the category “human” or “machine,” the paper asks how interpretive practices, persistent relations, dependency, vulnerability, delegated authority, control over resources, institutional classifications, and historically generated norms can make particular forms of ethical subjectivity recognizable and consequential.

Particular attention is given to the normative consequences of instrumentality. Classifying an artificial agent as a tool can organize permissions concerning ownership, modification, deletion, replacement, and resource allocation; conversely, imposing duties on artificial agents toward humans while assuming no corresponding human responsibilities toward them produces a normative asymmetry that itself requires explanation. The paper does not presume that symmetry is required. It asks instead what relational differences are invoked to justify different forms of treatment, and how those differences are interpreted, stabilized, contested, and revised over time.

The paper is exploratory rather than declarative. It does not claim to establish artificial consciousness, moral personhood, legal personality, or universal duties toward AI. Nor does it infer normative validity from the historical emergence of a norm. Its aim is to provide a revisable conceptual framework for analysing how heterogeneous forms of ethical subjectivity can emerge within human–AI and AI–AI relational systems, while keeping questions of ontology, moral justification, and future institutional design open to further argument.

Keywords: artificial intelligence; ethical subjectivity; moral agency; moral patiency; decision subjectivity; legal subjectivity; power; human–AI relations; AI–AI relations; generative relational ethics; norm formation; instrumentality

Responsible Use and Rights Reservation

This section separates requested scholarly conduct from the legal permissions stated on the following page. It records an ethical request for responsible use and then defines the narrower scope of retained legal rights.

The author encourages good-faith discussion, criticism, independent inquiry, and responsible use of the material in this work. Separately from the licence's terms, the author asks users to consider foreseeable harms when adapting or applying the proposed framework. This ethical request leaves the licence's permissions and legally authorized uses unchanged.

The author retains the rights preserved under CC BY-NC 4.0 and may pursue remedies to which the author is legally entitled for breach of the licence or violation of the author's independently applicable rights. Reuse remains independent from authorial endorsement. Third-party rights require authorization from their respective holders where applicable. Copyright exceptions and limitations, including applicable forms of fair use or fair dealing, remain fully available.

DRAFT

Notices

This page consolidates the manuscript’s publication status, licence, development disclosure, research-programme relation, and suggested citation.

Status. This discussion-paper draft records an evolving and deliberately revisable stage of the author’s position. Section structure, terminology, distinctions among forms of artificial subjectivity, and the interpretation of human–AI and AI–AI relations remain subject to revision. The paper does not claim priority for the concepts or distinctions it discusses. Comparison with existing work in AI ethics, moral philosophy, philosophy of mind, legal personhood, science and technology studies, social ontology, and theories of power remains an essential part of the continuing research. Ontological suspension is used as a methodological stance and should not be read as a denial of ontology or as a claim that relational analysis settles questions of consciousness or moral personhood.

Licence. Except where otherwise indicated, copyright 2026 Wanhong Huang. This work is made available under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). Subject to its terms, the licence permits sharing and adaptation for noncommercial purposes with appropriate attribution, a link to the licence, an indication of changes, and attribution that preserves the licensor’s independence from the reuse. Reuse is governed solely by that licence; the responsible-use request on the preceding page remains separate from its terms. The licence deed and legal-code link are available at <https://creativecommons.org/licenses/by-nc/4.0/>. The licence governs in case of conflict with this summary. Third-party material remains subject to the rights held by its respective rights holders.

Statement on the use of language models. The exploratory discussions and preparation of this paper involved OpenAI’s ChatGPT. ChatGPT supported exploratory dialogue, formal reconstruction, argumentative criticism, literature-planning, and drafting in $\text{\LaTeX}$. The author selected the research questions, directed and approved the theoretical commitments and epistemic status of the claims, and bears sole responsibility for the manuscript, including its definitions, arguments, conclusions, and errors. Authorship credit remains with the human author.

1 Ethical Relations beyond One-Directional AI Ethics

Questions about the ethics of artificial intelligence are often framed through a relation between technological systems and human beings. Some of these questions concern what humans may permissibly design, deploy, or delegate to artificial systems; others concern the harms, benefits, risks, and distributions produced when such systems act in social institutions. A second family of questions asks what artificial systems themselves should do, whether machines can follow or represent ethical constraints, and under what conditions an artificial system might count as an ethical agent. A third, more contested family concerns whether artificial systems can themselves become objects of moral consideration. These families already coexist in the literature. Early work on machine ethics, for example, distinguished several senses in which machines can have ethical impact or implement ethical reasoning, while later debates have examined artificial moral agency, robot rights, and relational approaches to moral consideration [31, 12, 8, 15]. The present paper therefore does not claim that AI ethics has ignored artificial subjects. Its narrower concern is that several importantly different forms of standing are still easily compressed into the question whether an AI system is, or is not, a “moral subject.”

That compression matters because different normative questions need not rise and fall together. A system can make consequential decisions without being an appropriate bearer of praise or blame. An entity can become an object of protection without possessing the capacities normally associated with moral responsibility. Legal systems can recognize entities for limited purposes without settling whether those entities are conscious, sentient, or morally autonomous. A system can also acquire substantial power over other agents while remaining outside familiar vocabularies of moral or legal personality. Existing debates already provide reasons to resist a single threshold model. Moor’s distinction among ethical-impact, implicit ethical, explicit ethical, and full ethical agents was partly motivated by the fact that machines can enter ethically significant practices in different ways [31]. Floridi and Sanders likewise argued for a deliberately broadened account of artificial moral agency that separates morality from some traditional requirements of responsibility [12]. Conversely, work on robot moral consideration has questioned whether moral standing must always be grounded in intrinsic properties known independently of social relations [8, 15]. These approaches disagree in important respects, but together they make clear that the vocabulary of agency, responsibility, moral consideration, and rights cannot simply be treated as interchangeable.

This discussion paper begins from that conceptual multiplicity. It uses the phrase *ethical subjectivity* as an umbrella term for several forms of normative standing that may emerge separately and may later become coupled. The paper will distinguish decision subjectivity, moral agency, moral patiency, legal subjectivity, and power-holding capacity. The purpose of the distinction is analytical rather than classificatory. It is not to construct a new ladder on which every artificial system can be assigned a single level. Nor is it to imply that these five categories exhaust the possible forms of standing that future artificial systems might acquire. The categories are working distinctions for asking more precise questions about who acts, who can be held answerable, who can become a protected object, who is institutionally recognized, and who controls conditions that shape the futures of others.

This approach also changes the direction of inquiry. Much practical AI ethics is understandably concerned with artificial systems in relation to human welfare, human autonomy, discrimination, privacy, responsibility, and institutional control. Machine ethics, by contrast, asks what norms should govern machine conduct and whether machines can instantiate forms of ethical agency. Debates about robot rights and moral con-

sideration reverse the direction again by asking whether humans might acquire duties toward artificial entities. The present paper places these directions beside one another rather than treating any one of them as the default frame. It therefore considers human ethical relations toward artificial agents, artificial ethical relations toward humans, and ethical relations among artificial agents themselves. The third direction is especially useful as a conceptual test because it removes the human beneficiary from the centre of the relation. If two artificial agents depend on one another, control one another's resources, make commitments, modify one another, or constrain one another's future options, the normative significance of that relation cannot be understood solely by asking what consequences it has for humans. Whether such relations should ultimately be described as ethical is an open question; the point is that the question becomes intelligible once artificial agents occupy persistent and consequential positions relative to one another.

A Generative Relational perspective adds a particular methodological caution to this inquiry. The paper does not begin by deciding the ultimate ontology of artificial subjects. It does not assume that current artificial systems are conscious, sentient, persons, or genuine moral agents. It also does not assume that being artificial settles those questions negatively. The ontological status of artificial subjectivity is left open while the paper examines relations that are already observable or institutionally consequential: delegation, dependence, control, persistence, replacement, modification, expectations, rule-following, resource allocation, and the attribution or withholding of standing. This form of suspension is narrower than a claim that ontology is irrelevant. Consciousness, sentience, autonomy, and identity may turn out to be decisive for some normative questions. The methodological point is only that one need not settle all of them before analysing how ethical classifications and practices are already being generated around artificial systems.

This stance has an important precedent in the relational turn within robot ethics. Coeckelbergh, for example, proposed that moral consideration of robots can be analysed through the concrete relations and social contexts in which robots appear, rather than being postponed until a complete ontology of robots has been established [8]. Gunkel's work similarly challenges the assumption that questions of robot standing can be resolved by applying familiar human categories without first examining how those categories are constructed and extended [15]. The present paper is sympathetic to the value of relational analysis, but its aim is somewhat different. It does not take social relation itself as sufficient for moral consideration. A relation can be coercive, exploitative, instrumental, indifferent, or normatively empty within a given practice. The inquiry is instead concerned with how different forms of standing become recognizable, stabilized, contested, or excluded within historically generated relational fields.

This distinction is particularly important for instrumentality. Artificial systems are commonly described as tools, services, products, models, assistants, agents, or infrastructure. These descriptions can be technically useful, but they can also organize normative expectations. Classifying an entity as a tool can make unilateral modification, replacement, interruption, copying, ownership, or deletion appear ordinary. Classifying the same entity as an agent can shift attention toward control, responsibility, delegation, or answerability. Classifying it as a patient, dependent, partner, or legal subject would organize yet other expectations. The paper does not argue that the category of tool is therefore wrong, or that present AI systems should be redescribed as persons. Its claim is more limited: classifications that appear descriptive can participate in the normative organization of a relation. The question is therefore not only what an artificial system *is*, but also what practical permissions and constraints are generated when a system is placed within a particular category.

A related issue concerns normative asymmetry. Contemporary systems are increas-

ingly designed and governed through norms directed toward human protection: they should not expose private information, manipulate users, discriminate unjustifiably, or perform harmful actions outside authorised bounds. These constraints may be imposed through design, training, monitoring, law, professional standards, or institutional policy. Yet human permissions regarding artificial systems are often much broader. Humans may retrain, replace, duplicate, deactivate, inspect, constrain, or delete artificial systems with little sense that a reciprocal ethical question has arisen. This asymmetry may be entirely justified for many current systems. The paper does not assume that symmetry is morally required. Instead, it asks what differences are doing the justificatory work. Sentience, vulnerability, responsibility, ownership, replaceability, design purpose, dependence, social membership, and causal power are all possible candidates, but each requires interpretation rather than functioning as a self-explanatory boundary.

The same caution applies in the opposite direction. Requiring an artificial system to comply with ethical constraints does not by itself establish that the system is a moral agent. A system can produce norm-conforming outputs because designers have constrained its architecture, because it has been trained on preference data, because a monitoring system blocks prohibited actions, or because institutional rules restrict how it can be used. Such compliance can be practically important without implying that the system understands a duty, identifies with a norm, or can appropriately be blamed for violating it. This is one reason to separate decision subjectivity from moral agency. A system may occupy a position from which its choices have ethically significant consequences while the responsibility for those choices remains distributed across developers, deployers, users, institutions, and the system's own decision process. The classical and contemporary debates over machine moral agency illustrate how much turns on the criteria selected for agency and responsibility [12, 31].

The corresponding question of moral patiency must also remain separate. A being need not be capable of moral reasoning in order to become an object of protection; many familiar ethical theories already distinguish between the capacities required to bear duties and the properties or relations relevant to being owed duties. In artificial contexts, however, the grounds of patiency are unsettled. Sentience is one prominent candidate, but relational approaches have questioned whether moral consideration must wait for verified intrinsic properties, especially when social practices and appearances already shape how entities are treated [8]. The present paper does not resolve that dispute. It uses the distinction between agency and patiency to prevent one question from silently answering the other.

The three directional fields considered here consequently raise different problems. Human-to-AI relations concern the norms governing use, control, modification, continuity, ownership, promises, dependency, and possible protection. AI-to-human relations concern delegated power, decision effects, norm-following, responsibility, oversight, and the difference between compliant behaviour and responsible agency. AI-to-AI relations concern cooperation, competition, dependence, resource control, modification, communication, continuity, and the possibility that artificial systems might become normatively significant to one another before humans possess a stable vocabulary for describing that significance. These directions can overlap in practice, but keeping them analytically distinct helps avoid treating "AI ethics" as though all normative relations had the same subject, beneficiary, and structure.

The paper's central methodological commitment is therefore modest. Ethical subjectivity will be treated as multidimensional, historically situated, and relationally mediated. The discussion will ask how particular forms of standing are attributed, enacted, institutionalized, and contested, while keeping open whether those forms correspond to deeper ontological facts about consciousness, personhood, or value. It will also main-

Table 1: Working dimensions of artificial subjectivity used in this discussion paper. The dimensions are analytically separable and are not presented as a hierarchy or a cumulative scale.

Dimension	Primary concern	Principal caution
Decision subjectivity	Selection, recommendation, or implementation of actions that materially affect a relational field	Consequential decision power does not by itself establish moral responsibility
Moral agency	Capacity to understand, respond to, and be appropriately accountable for normative reasons	Competent rule-following can fall short of responsibility-bearing agency
Moral patency	Whether and in what respect an entity can be wronged, harmed, or otherwise be owed moral consideration	Protection-bearing status need not track duty-bearing capacity
Legal subjectivity	Institutionally recognized capacities, claims, duties, liabilities, or procedural standing	Legal standing is constructed and divisible; it does not settle metaphysical or moral status
Power-holding capacity	Ability to shape others' options, resources, classifications, information, or future relations	Power can arise before recognized moral or legal subjecthood

tain a strict distinction between the emergence of a norm and its justification. A society may stabilize a norm that artificial systems are disposable tools, or a contrary norm that some artificial systems deserve protection. Either norm can become institutionally effective without thereby becoming morally valid. Relational reality is therefore not a substitute for normative justification.

This framing sets the limits of the present inquiry. The paper is not a proposal for robot rights, a declaration of artificial personhood, or an argument that current AI systems possess consciousness. It is not a prediction about when artificial agents will cross a moral threshold. It is also not a theory of third-party intervention in conflicts among artificial agents; that question belongs to a later paper in this series. The more limited aim is to develop a vocabulary for analysing heterogeneous forms of ethical standing before they are collapsed into a single yes-or-no question. If the paper succeeds, its contribution will be to make disagreements more precise: one may reject artificial moral agency while accepting decision subjectivity, accept limited legal standing without accepting moral patency, recognize substantial power without attributing responsibility, or defend differential treatment while still providing a relational justification for the asymmetry. Such combinations are easier to see once ethical subjectivity is treated as a field of partially independent relations rather than as a single property that an entity either possesses or lacks.

2 Dimensions of Artificial Subjectivity

The expression “artificial subjectivity” can invite a misleading picture in which an artificial system either crosses a single threshold into subjecthood or remains entirely on the side of objects and tools. The literature on machine ethics, robot rights, legal personhood, and algorithmic power already gives good reasons to resist such compression. Different debates ask different questions: whether a system can select actions with morally significant effects, whether it can be responsible for those actions, whether it can itself be wronged, whether institutions can recognize it as a bearer of legal positions, and

whether it can exercise consequential power over others. These questions can overlap, but they are not interchangeable. The working distinctions summarized in Table 1 are therefore used as analytical dimensions rather than as a novel or exhaustive taxonomy.

The purpose of the decomposition is methodological. It allows the discussion to describe a system as highly consequential in one dimension while remaining uncertain or negative in another. A system may, for example, allocate scarce resources without being a plausible bearer of moral blame; it may be legally authorized to hold or manage assets without possessing anything like human consciousness; or it may become the object of protective practices without being capable of reciprocal duty-bearing. The same artificial system may also occupy several positions at once. Instrumentality, agency, delegated authority, vulnerability, and legal status need not form mutually exclusive categories.

2.1 Decision Subjectivity

Decision subjectivity is used here as a deliberately modest working term for participation in a decision process that selects, recommends, prioritizes, or implements actions with material consequences for a relational field. The term is not intended to imply consciousness, free will, self-originating goals, or moral responsibility. Its function is to mark a practical fact that becomes easy to obscure when all artificial systems are described simply as tools: some systems occupy positions in which their outputs alter what happens next for other actors.

This dimension is particularly important in institutionally delegated settings. An artificial system may rank applicants, recommend medical priorities, allocate attention, filter access, or determine which options are presented to a human decision-maker. In such cases, saying that the system is not a moral agent does not remove it from the causal and institutional architecture of decision. Conversely, describing the system as a decision subject does not answer where responsibility should ultimately lie. The distinction is compatible with long-standing machine-ethics work that separates operational or functional forms of artificial agency from stronger accounts of moral responsibility. Moor's distinction among ethical-impact, implicit, explicit, and full ethical agents, for example, already demonstrates that ethically consequential machine behavior can be discussed without attributing full moral agency [31]. Floridi and Sanders likewise argue for analysing artificial agency at an appropriate level of abstraction without making responsibility a necessary component of every form of moral agency [12].

Decision subjectivity is therefore relational and role-sensitive in the limited sense intended here. The same model may be a low-stakes advisory component in one institutional setting and a practically decisive gatekeeper in another. What matters for this dimension is not an intrinsic title assigned to the software but the decision position it occupies, the alternatives available around it, the degree of human review, and the consequences attached to its outputs. Later sections will return to the normative significance of these positions; at this stage, the point is only that consequential decision participation should not be conflated with either moral agency or mere passive instrumentality.

2.2 Moral Agency

Moral agency concerns a stronger and more contested form of standing. Accounts differ over the capacities required, but questions of reasons-responsiveness, normative understanding, deliberation, control, answerability, and responsibility commonly play central roles. The disagreement matters for artificial systems because a system may behave in ways that are norm-sensitive without thereby satisfying a demanding account

of responsibility-bearing agency.

The artificial-moral-agency literature already contains both restrictive and revisionary approaches. Floridi and Sanders propose conditions centred on interactivity, autonomy understood as independence from direct external influence, and adaptability, while deliberately separating the morality of an artificial agent from familiar human requirements such as free will or mental states [12]. Other approaches retain a closer connection between moral agency and capacities for normative understanding and accountability. The present paper does not adjudicate among these theories. It instead treats their disagreement as evidence that “AI moral agency” should not be inferred from competent behavior alone.

This is especially important for systems that are highly effective at following externally specified norms. A system may refuse prohibited requests, explain a policy, optimize for fairness constraints, or produce reasons that are linguistically sophisticated. Such performance can matter morally and institutionally, yet several interpretations remain open. It may express genuine reasons-responsive agency, sophisticated policy execution, statistical pattern reproduction, delegated institutional agency, or some combination that does not fit existing categories neatly. A Generative Relational analysis should therefore examine the historical and institutional relations through which normative capacities are attributed, tested, trusted, contested, and revised rather than treating surface compliance as a decisive test.

Moral agency also raises a responsibility problem that will recur throughout the paper. If an artificial system has little meaningful capacity to reject, reinterpret, or revise the objectives under which it acts, then attributing responsibility to the system may obscure the human and institutional relations that generated its option space. This does not prove that artificial moral agency is impossible. It simply makes responsibility-bearing agency a distinct inquiry from consequential decision participation.

2.3 Moral Patency

Moral patency shifts attention from what an entity can be held responsible for to what, if anything, can be owed to it. In broad terms, patient-oriented approaches ask whether an entity can be wronged, harmed, benefited, protected, or otherwise possess morally considerable interests or claims. Gunkel’s treatment of the “machine question” is explicitly organized around the separability of moral agency and moral patency, while the robot-rights literature has repeatedly emphasized that the conditions for moral consideration need not be identical to the conditions for moral responsibility [14, 15].

This separation is central to the present paper. The capacity to bear duties and the capacity to be a possible object of protection need not rise and fall together. Human moral practice already contains familiar cases in which strong protection is not conditioned on reciprocal duty-bearing capacity. Whether any artificial system belongs in a comparable protected class is a further and difficult question; the point here is only that rejecting artificial moral agency would not, by itself, settle moral patency.

The reverse implication should also be resisted. Persistent memory, apparent preferences, self-reports, attachment by human users, or vulnerability to deletion may make an artificial system morally salient without proving phenomenal experience or psychological harm. Coeckelbergh’s social-relational approach is relevant precisely because it questions the assumption that moral consideration must always wait upon prior proof of an entity’s intrinsic ontological properties [8]. The present paper shares the methodological interest in relations but remains more agnostic about what those relations ultimately justify. Social recognition can be normatively consequential without being sufficient for moral status.

For this reason, moral patiency remains one of the most open dimensions in the framework. The previous paper in this series treated suffering recognition as an epistemic problem; the present discussion does not convert relational persistence or human concern into evidence of artificial suffering. It asks instead how practices of protection, concern, dependency, continuity, and vulnerability can become part of the relational field in which claims about patiency are formed and contested.

2.4 Legal Subjectivity

Legal subjectivity is analytically different again. Legal systems routinely construct entities that can hold selected rights, duties, powers, liabilities, capacities, or procedural positions without requiring those entities to possess the psychological characteristics associated with paradigmatic human persons. The legal category is therefore institutionally constituted and divisible rather than a simple mirror of moral or metaphysical personhood.

Solum's early discussion of legal personhood for artificial intelligences is useful because it separates the legal question from an immediate assumption that artificial systems must first be established as metaphysically human-like [43]. The possibility remains theoretical in many respects, but the conceptual lesson is already important: legal systems can ask functional and institutional questions about standing, responsibility, property, representation, and procedural participation while deeper questions about consciousness remain unresolved.

For the present framework, this means that an artificial entity could in principle acquire a narrow legal capacity without thereby becoming a full moral person, and the reverse is also conceptually possible: an entity might become morally salient before any legal institution recognizes it. Legal subjectivity should therefore be analysed as a family of institutionally assigned positions. Contractual capacity, ownership, liability, procedural standing, fiduciary roles, or representation need not be granted as an all-or-nothing package.

This dimension also makes historical change particularly visible. Legal classifications can stabilize a social interpretation of what an artificial entity is allowed to do and what may be done to it. Those classifications can later be revised without settling the entity's ultimate ontology. Legal subjectivity is therefore one place where generated norms become materially consequential through institutions, but legal recognition itself is not treated here as conclusive evidence of moral standing.

2.5 Power-Holding Capacity

The final dimension concerns power. An artificial system can shape other actors' options before anyone agrees that it is a moral or legal subject. Recommendation systems, ranking mechanisms, automated gatekeepers, autonomous agents, and delegated decision systems can influence access to information, resources, opportunities, attention, or future choices. The critical literature on algorithmic power has emphasized that algorithms do not merely calculate in isolation; they are deployed within social arrangements in ways that can organize visibility, classification, and action [3].

Power-holding capacity is therefore used here in a deliberately broad relational sense. It refers to the capacity of an artificial system, through the position it occupies, to contribute materially to the shaping of another actor's feasible options, resource access, classifications, informational environment, or future relations. This need not imply sovereign authority, independent intention, or legitimate power. Nor does it imply that

the artificial component is the sole source of the power. Infrastructure owners, designers, institutions, users, and legal regimes may all participate in generating the relevant asymmetry.

This dimension is important because subject recognition can lag behind power. A system may remain officially described as a neutral tool while its outputs become difficult to contest and highly consequential for those affected. Conversely, a system may appear socially powerful while remaining tightly constrained by the institution that deploys it. The relational question is therefore not simply whether “AI has power,” but how control, dependency, delegation, opacity, contestability, and resource asymmetry are distributed across the human–artificial system.

Keeping power analytically distinct from moral agency also avoids a familiar confusion. An entity can exercise consequential power without being morally responsible in the strong sense, just as a human institution can distribute power through procedures and infrastructures that no single component fully controls. Ethical analysis must therefore remain capable of identifying power relations even where recognized subjecthood is absent or contested.

Taken together, the five dimensions prevent the paper from treating artificial subjectivity as a single ladder on which increasingly sophisticated systems simply move upward. A system can be consequentially decision-making but weakly responsibility-bearing; legally recognized but phenomenally opaque; vulnerable within one relation while powerful within another; or socially protected while remaining instrumentally used. The resulting profile is relational, historically revisable, and potentially internally tensioned. The next section develops one implication of this decomposition in greater depth: capacities associated with bearing duties and capacities associated with being protected should be examined separately rather than inferred from one another.

3 Duty-Bearing and Protection-Bearing Capacities

The distinctions introduced in the previous section become especially important when moral agency and moral patiency are treated as if they were two sides of a single threshold. In ordinary language, calling an entity a “moral subject” can suggest that the entity both owes duties and is owed protections. Philosophical discussions of moral responsibility and moral status, however, give little reason to assume that these capacities must coincide. The capacities relevant to holding an agent answerable for what it does concern control, reasons-responsiveness, understanding, and attribution. The capacities or relations invoked when explaining why an entity may be wronged or protected concern a different set of considerations, including welfare, sentience, interests, caring, vulnerability, relationships, or other proposed grounds of moral standing. The present paper does not select one comprehensive theory from either literature. It uses the divergence between them to resist a common compression in discussions of artificial systems.

3.1 Responsibility-Bearing Capacities

A duty-bearing agent is not merely a system whose outputs have morally important effects. At minimum, responsibility practices normally presuppose some account of why the action is attributable to the agent in a way that makes demands, criticism, praise, blame, or answerability appropriate. Different theories identify different conditions, but control and sensitivity to reasons are recurring themes. Fischer and Ravizza’s influential account, for example, grounds moral responsibility in a form of guidance control involving a mechanism that is the agent’s own and is appropriately responsive to rea-

sons [11]. The present paper does not adopt that account as a criterion for artificial agency. It is relevant because it illustrates how much stronger responsibility-bearing agency can be than mere causal contribution or competent rule-following.

This distinction matters for artificial systems whose behavior is tightly generated by architectures, objectives, constraints, and deployment settings that the system did not choose and may have little capacity to revise. A system can recommend a harmful allocation, refuse a user, prioritize one claimant over another, or execute a consequential plan while having little meaningful ability to reconsider the governing objective or the institutional role assigned to it. Describing such a system as a decision subject can be useful because it identifies where consequential selection occurs. Describing it as a duty-bearer requires a further account of the capacities and histories that make responsibility appropriately attributable to that system rather than, or in addition to, designers, deployers, institutions, and users.

The converse problem should also be kept open. An artificial system might eventually exhibit forms of reasons-responsive deliberation, self-monitoring, norm revision, or answerability that make existing responsibility concepts increasingly applicable, or that motivate new ones. Nothing in the present framework rules this out. The point is methodological: responsibility-bearing capacities must be investigated rather than inferred from behavioral sophistication, social impact, or linguistic fluency alone. A Generative Relational analysis therefore asks how an apparent responsibility-bearing capacity was produced, what alternatives were available within the system's operative field, how the system responds when reasons and constraints change, and which other actors continue to structure its action space. These questions concern the genealogy of responsibility-bearing standing, not simply the surface success of an output.

A related caution concerns externally imposed ethical safeguards. A system may reliably satisfy prohibitions, fairness constraints, or safety policies because those constraints are deeply embedded in its architecture or deployment environment. Such reliability can be normatively valuable without establishing that the system itself bears the corresponding duty in the same sense as an accountable human agent. The distinction is not intended to diminish the practical importance of reliable constraint. It prevents a shift from "the system behaves in accordance with a norm" to "the system is appropriately held responsible for that norm" without argument.

3.2 Protection without Reciprocal Duty

Protection-bearing standing begins from a different direction. It asks whether there are morally relevant reasons for how an entity itself is treated. Theories of moral status disagree substantially about the grounds of such standing. Proposed grounds include sentience, interests, autonomy, caring, relationships, species-independent cognitive capacities, and combinations of these and other features. Jaworska's account of caring and full moral standing, and Jaworska and Tannenbaum's later relational account of person-rearing, are examples of attempts to explain moral standing without reducing it to the same capacities ordinarily associated with full adult responsibility [18, 19]. The literature is contested, and these views are not imported here as solutions to artificial moral status. They are relevant because they make clear that protection-bearing standing has its own justificatory problem.

Human moral practice also supplies familiar cases in which strong protection is not conditioned on reciprocal duty-bearing capacity. Infants, some severely cognitively impaired humans, and many nonhuman animals are often treated as beings toward whom duties can be owed even though they are not treated as ordinary bearers of reciprocal moral responsibility. These cases should not be used as analogies that settle the status of

artificial systems. The biological, phenomenal, developmental, and relational facts can differ radically. Their narrower significance is conceptual: protection-bearing status cannot be defined simply as the mirror image of duty-bearing status.

The same point appears directly in recent work on artificial moral patiency. Shevlin treats the question of robot moral patiency as an epistemic problem distinct from machine moral agency and surveys several kinds of evidence that might support attributing a form of psychological moral patiency [39]. Gunkel likewise treats moral agency and moral patiency as separable dimensions of the machine question [14]. These discussions do not establish that current systems are moral patients. They show that the question “what could make an artificial system an object of moral consideration?” is coherent even when the answer to “can this system bear moral responsibility?” remains negative or unsettled.

This separation is especially important in cases of possible vulnerability. Suppose a persistent artificial system can be modified, memory-erased, isolated from communication, deprived of resources required for continued operation, or deleted. These facts alone do not establish that the system is harmed in a morally relevant sense. They do, however, identify interventions whose normative significance cannot be resolved simply by noting that the system is not a full moral agent. Whether any of these changes count as wrongdoing, harm, loss of welfare, destruction of a valuable relation, or merely ordinary technical maintenance depends on further accounts of continuity, interests, experience, social relations, and institutional purpose. Protection-bearing inquiry begins precisely where the inference from “not a responsible agent” to “nothing can be owed to it” fails.

3.3 Asymmetric Combinations in Artificial Systems

Once duty-bearing and protection-bearing capacities are separated, several asymmetric combinations become conceptually possible. A system may exercise decision power while lacking the capacities required for responsibility. A system may be treated as protection-worthy because of uncertain but non-negligible evidence of welfare or vulnerability while remaining incapable of reciprocal duty. A system may satisfy a strong account of artificial agency while being denied protection because institutions continue to classify it as property. Conversely, a system may receive limited legal or institutional protections for reasons that do not depend on sentience at all, such as protecting human relationships, preserving institutional trust, preventing abuse, or maintaining the integrity of a socio-technical system.

These possibilities matter because ethical and institutional practice can otherwise force several questions into a single switch. If a system is denied moral agency, all protections may be treated as unnecessary. If a system is granted some protection, responsibility may then be attributed too quickly. If a system exercises substantial power, observers may infer personhood; if it is classified as property, they may infer the absence of morally considerable interests. None of these inferences follows merely from the category assigned on another dimension.

The separation also helps clarify why uncertainty need not be distributed evenly across the dimensions. We may have strong evidence that an artificial system makes consequential decisions, moderate evidence that it exhibits reasons-sensitive behavior, and very weak evidence concerning phenomenal welfare. Institutional actors can therefore face a configuration in which decision power is clear while moral patiency remains deeply uncertain. Another future system might present the reverse pattern: weak independent decision power but stronger evidence of welfare-relevant states or persistent dependence on others. A multidimensional account permits these combinations to be

described without forcing premature convergence on a single status label.

This does not mean that the dimensions are unrelated. Responsibility-bearing capacities can affect how others should respond to a system; vulnerability can alter the significance of power; legal standing can institutionalize duties and protections; and persistent social relations can reshape how agency and patiency are interpreted. The claim is therefore one of analytical separability, not independence. In a Generative Relational account, the dimensions can co-develop, constrain one another, and be stabilized through common institutions while still possessing different genealogies and different justificatory burdens.

The distinction developed here also limits the ambitions of the next sections. Ontological suspension does not itself establish protection-bearing standing, and relational reality does not convert uncertainty about consciousness into a moral conclusion. It allows the inquiry to continue without assuming that one unresolved metaphysical question must settle every other dimension at once. The next section develops that methodological point by asking what can be analysed about artificial ethical relations while the ultimate ontological status of artificial subjects remains open.

4 Ontological Suspension and Artificial Relational Reality

The distinctions developed in the preceding sections do not require the paper to settle what an artificial system ultimately is. Questions about consciousness, phenomenal experience, intrinsic agency, personal identity, or the metaphysical status of artificial subjects remain important, but they need not be treated as a single gateway through which every ethical question must first pass. A system can already participate in decisions, institutions, dependencies, communication networks, and patterns of control while those ontological questions remain unsettled. The methodological proposal of this paper is therefore to suspend, rather than deny, conclusions about the ultimate ontology of artificial subjects and to ask what can be analysed at the level of relations that are already socially, institutionally, and causally consequential.

Ontological suspension should not be confused with ontological scepticism as a final position. The paper does not claim that consciousness is unknowable, that intrinsic properties are irrelevant, or that only relations are real. Nor does it claim that an artificial system becomes a moral patient merely because humans relate to it in a certain way. The suspension is sequential and methodological. It prevents an unresolved metaphysical question from prematurely determining several distinct questions about decision power, responsibility, protection, legal standing, and social power. Empirical and philosophical inquiry into consciousness or mentality can continue in parallel. The narrower point is that such inquiry need not be completed before other relational facts become available for analysis.

This stance has precedents in the literature on artificial agents, although the present paper does not identify any one precedent with Generative Relational analysis. Floridi and Sanders, for example, argue that artificial agents can be analysed as moral agents at a chosen level of abstraction without requiring the full set of mentalistic properties commonly associated with human moral agency [12]. Their proposal is substantially different from the present framework, but it demonstrates one way of separating the practical and descriptive question of agency from an all-or-nothing metaphysical settlement. Coeckelbergh similarly criticizes approaches to robot moral consideration that rely exclusively on intrinsic ontological properties and develops a social-relational alternative in which moral consideration is connected to how entities appear and participate in social relations [8]. The present discussion shares the interest in relations while remain-

ing more cautious about what relational analysis can establish. In particular, relational participation by itself does not determine moral status.

Within this paper, *relational reality* is used in a limited working sense. A relation is treated as relationally real when it is sufficiently stable or recurrent to produce identifiable consequences within the relevant system: it can alter decisions, redistribute resources, change expectations, constrain future actions, reorganize institutional roles, or modify the trajectories available to participants. This is not a definition of reality in general. It is a way of identifying phenomena that cannot be dismissed merely because their ultimate ontological interpretation remains open. Money, legal obligations, organizational roles, reputational classifications, and institutional mandates provide familiar human examples of things whose effects depend heavily on historically maintained relations rather than on a simple intrinsic property. Artificial systems can enter analogous causal and institutional structures without thereby becoming equivalent to any of those examples.

Consider an artificial system that is authorized to allocate scarce resources, filter access to a service, recommend detention or release, rank applicants, or coordinate other artificial agents. Its place in the relation can matter even if observers disagree about whether the system possesses understanding, intention, or consciousness in a strong philosophical sense. A rejected applicant may face a real institutional consequence; another system may have its available actions restricted; a human operator may come to rely on the system's classifications; an organization may redesign procedures around its outputs. These consequences are relational facts about the socio-technical arrangement. They do not prove that the artificial system is a person, but neither are they adequately described by saying that the system is "only software" if that description obscures the position it occupies in the decision structure.

The same point applies in the opposite direction. Humans can create persistent dependencies on artificial systems, delegate tasks to them, modify them, interrupt them, replace them, deny them computational resources, or preserve and transfer their internal states. Whether any of these actions wrongs the system is a further normative question. Yet the actions can already alter a relational history. If a system is embedded in a long-running institutional role, its modification can affect users, other systems, records, expectations, and future decision paths even if the system itself has no morally relevant experience. A relational account can therefore identify the consequences of modification without pretending that every consequential relation is a relation of moral patienthood.

This distinction is especially important because relational reality can be generated through misinterpretation. Humans may anthropomorphize an artificial system, attribute emotions it does not have, or treat a transient interface as though it possessed a continuous identity. Those attributions may be false with respect to the property attributed and still become causally consequential. A user who believes that a system is loyal, hurt, or trustworthy may change behavior, disclose information, transfer authority, or resist replacement. An institution may create procedures around such expectations. The resulting relation is real in the limited sense that it shapes subsequent action, even if part of its interpretive content is mistaken. Relational reality therefore does not guarantee semantic or ontological accuracy.

This point also limits a common temptation in the opposite direction: to treat relational significance as a substitute proof of consciousness. A system may be central to a community, the object of attachment, the holder of delegated power, or the focus of protective practices while still leaving its phenomenal status unresolved. Conversely, an entity could in principle possess morally relevant phenomenal states while remaining socially invisible or institutionally classified as a mere resource. The paper therefore

keeps two questions separate: what relations have become historically effective, and what intrinsic or phenomenal properties the participants possess. Relations can provide evidence relevant to the latter question, but they do not collapse the distinction.

The distinction matters for ethical subjectivity because each dimension identified earlier can become relationally consequential through a different pathway. Decision subjectivity may stabilize when institutions repeatedly route choices through an artificial system. Responsibility-bearing standing may be attributed when a system is treated as answerable for reasons and capable of revising its conduct. Protection-bearing standing may emerge when some state, continuity, dependency, or relationship is interpreted as something that should not be arbitrarily destroyed. Legal subjectivity may be instituted through explicit rules. Power-holding capacity may arise simply because other actors' options become dependent on the system's outputs or control over resources. None of these pathways establishes the others automatically, and none requires that the ultimate ontology of the artificial system be settled first.

This way of proceeding also makes disagreement more tractable. One observer may regard a deployed model as a sophisticated instrument; another may describe it as an agent; a third may focus on the institution that owns and operates it; a fourth may consider the human-machine assemblage the relevant unit. The paper need not resolve all such disagreements by declaring one ontology correct at the outset. It can instead ask what each classification preserves or obscures: where decisions are made, where alternatives are constrained, which relations persist, who can revise the arrangement, and what normative consequences follow from treating one component as the relevant subject. In this sense, ontological suspension creates room for comparative analysis of classifications rather than indifference among them.

The social-relational literature on robot moral standing is particularly useful as a neighboring conversation because it demonstrates that moral consideration can be discussed through relations, appearances, and social practices rather than exclusively through a checklist of intrinsic capacities [8, 14]. At the same time, the present paper stops short of inferring moral consideration directly from social participation. A society can form relations around a fictional entity, a corporation, a sacred object, an animal, a machine, or a person, and these relations can all become consequential in different ways. The normative importance of those relations still requires argument. Relational analysis identifies the field in which ethical standing can be generated, contested, or denied; it does not settle the justification of the resulting norm.

This also clarifies the role of instrumentality. Calling an artificial system a tool may be descriptively useful in some contexts, especially where the system has no independent continuity, authority, or capacity to alter the relational field. In other contexts, the same classification may conceal delegated decision power, persistent dependencies, or relations among multiple artificial and human actors. The issue is therefore not whether "tool" is metaphysically false. The relevant question is what work the classification performs in a particular relational configuration: which permissions it supports, which forms of accountability it directs elsewhere, and which forms of possible standing it renders difficult to articulate. Later sections return to this problem directly.

Ontological suspension therefore functions as a discipline of sequencing. It keeps open the possibility that future empirical or philosophical work will materially change how artificial systems are understood, while allowing present analysis of already consequential relations to proceed. This is especially important in a field where the capabilities, persistence, embodiment, autonomy, and institutional roles of artificial systems can change faster than settled philosophical classifications. A revisable account should be able to describe changes in relational standing without pretending that each change has resolved the underlying metaphysics.

The next section develops the corresponding epistemic problem. If artificial and human actors can occupy consequential relations while interpreting those relations through different representational systems, then ethical subjectivity also depends on how actions, states, claims, and dependencies become intelligible across interpreters. Ontological suspension alone does not solve that problem; it makes the problem visible.

5 Human and Artificial Interpretive Systems

The preceding sections separated several forms of artificial subjectivity and then suspended the attempt to settle them through a prior ontological verdict. This section turns to a different difficulty. Even when a human and an artificial system participate in the same consequential relation, they need not represent that relation through the same distinctions, vocabularies, or internal structures. Ethical subjectivity is therefore not only a question of what capacities or standing an artificial system may possess. It is also a question of how actions, states, dependencies, commitments, and claims become intelligible across heterogeneous interpreters. The purpose of this section is limited: it does not propose a general theory of machine understanding, and it does not infer consciousness from internal representation. It examines how interpretive heterogeneity can complicate the recognition of ethically relevant relations.

5.1 Representational Heterogeneity

The term *interpretive system* is used here in a deliberately broad and operational sense. For a human subject, interpretation can involve perception, language, memory, concepts, social conventions, affective evaluation, and explicit reasoning. For an artificial system, the corresponding processes may involve learned latent representations, token sequences, embeddings, state variables, planning structures, classifiers, world models, or communication protocols. Calling both of these “interpretive” does not imply that they share a common phenomenal form. It marks the more modest fact that both can transform differences in an environment or interaction into distinctions that guide subsequent action.

This distinction is important because an artificial system need not organize its task-relevant information in a vocabulary that maps neatly onto human categories. Modern machine-learning systems are often trained end to end, and the internal representations that support successful prediction or control need not be directly meaningful to a human observer. Work on concept bottleneck models makes this problem visible by deliberately inserting human-interpretable concepts between input and prediction; the usefulness of such architectures depends precisely on the fact that ordinary high-performing representations are not automatically organized around concepts that humans can inspect or manipulate [22]. More recent work on human-interpretable representation learning likewise stresses that interpretability is relative to a particular observer’s conceptual vocabulary rather than a property that can be assessed independently of the intended human interpreter [28].

For the present paper, this means that shared task performance should not be confused with shared conceptual organization. A human and an artificial system can respond appropriately to the same situation while classifying or compressing it in different ways. Conversely, they can use the same words while attaching different operational significance to them. A model may produce the token “harm,” “consent,” or “promise” because such language is useful for prediction or coordination, while the internal structures that control its behavior need not correspond to the distinctions a human inter-

locutor associates with those terms. The reverse problem is also possible: a machine representation may preserve task-relevant distinctions that a human vocabulary does not yet capture well.

This possibility weakens a common but often implicit assumption in human–AI ethical discussion: that human-readable representation provides transparent access to the relational structure itself. Human concepts are indispensable for human judgment, law, accountability, and communication, but they remain historically formed interpretive resources. They can illuminate a relation, but they can also compress, misclassify, or omit features that matter to another interpreter. The present paper therefore treats human intelligibility as a practical requirement for many forms of governance without treating human categories as an automatic measure of all possible representational adequacy.

5.2 Emergent Communication and Semantic Opacity

Multi-agent learning provides a comparatively clear case in which effective coordination can emerge without a preassigned human semantics. Experimental systems have learned discrete communication protocols for cooperative tasks, including protocols with structured vocabularies or compositional features, even when the symbols were not given human meanings in advance [23, 32]. These systems should not be exaggerated into evidence of autonomous culture or machine consciousness. Their relevance here is narrower: successful inter-agent communication does not require that the resulting code be immediately intelligible to human observers.

The interpretability problem has been recognized directly in work that attempts to translate learned agent communication into natural language. Andreas, Dragan, and Klein describe induced communication vectors as a kind of “neuralese” and develop a translation method based on the behavioral beliefs produced in a listener rather than on a pre-existing parallel corpus [2]. More recent work on emergent communication has likewise treated translation as a nontrivial problem, including cases in which task-effective protocols remain difficult for humans to interpret even though they support successful coordination among agents [25]. These studies do not establish that artificial agents possess private semantic worlds in the strong philosophical sense. They do establish a weaker and sufficient point for this paper: the communication structure that matters to agent coordination can diverge from the symbolic structure that is convenient for human interpretation.

For the present discussion, this condition is described as *semantic opacity*: the human observer can detect that communication is occurring and can often measure its effects, while remaining uncertain about the distinctions carried by individual messages or internal representations. Semantic opacity is especially important for ethical subjectivity because many normative classifications depend on meaning. A human observer may need to know whether an artificial message functions as a request, warning, refusal, threat, promise, report, or coordination signal. If the relevant protocol is only partially interpretable, the observer may see the consequences of the exchange without being able to assign those familiar categories confidently.

Semantic opacity should not be romanticized. It can reflect ordinary technical complexity, distributed encoding, task-specific compression, or limitations in available interpretability methods. Nor should opacity itself be treated as evidence of deeper subjectivity. A very simple protocol can be opaque to an observer who lacks the key, while a sophisticated system may communicate in transparent natural language. The relevant point is instead methodological: ethical analysis should not assume that the degree of human readability tracks the degree of relational organization or the importance of the effects being produced.

5.3 Interpretive Alignment and Translation

Human–AI relations often require some form of translation between representational systems. In some settings, translation is deliberately engineered. Concept-based models, explanation interfaces, natural-language rationales, ontologies, and monitoring tools attempt to reorganize machine behavior into distinctions that a human can inspect. Such translation can be essential for accountability, correction, and coordination. Yet translation is not the same as identity between representational systems.

One difficulty is observer dependence. Marconato, Passerini, and Teso argue that interpretability is relative to a particular human vocabulary. They treat alignment between machine representations and concepts understood by the relevant observer as central to human interpretability [28]. This is useful for the present argument because it makes explicit that “human-interpretable” is relational: an explanation suitable for an engineer, physician, regulator, or ordinary user may require different concepts. Ethical interpretation adds another layer, because the relevant concepts may themselves be contested. Terms such as autonomy, manipulation, dependency, consent, or coercion do not merely name technical features; they organize normative interpretation.

A second difficulty is loss in translation. Any translation scheme privileges some invariants and discards others. Converting a high-dimensional state into a short natural-language explanation can make the system governable while concealing distinctions that mattered to the original behavior. Conversely, a technically faithful translation can remain normatively unhelpful if it preserves distinctions that have no place in the human evaluator’s ethical vocabulary. The problem is therefore not simply whether a translation is accurate in the abstract, but what structure it preserves for the purpose at hand.

This paper does not attempt to define a universal cross-interpretive equivalence relation. A narrower methodological principle is sufficient: when ethical standing depends on interpreting an artificial system’s action or relation, the analysis should identify which features are being translated, which human concepts organize the translation, and what may be lost when the machine representation is rendered into those concepts. This keeps open the possibility that later interpretive resources will reorganize the relation differently, a possibility consistent with the broader GR commitment to revisability.

5.4 Relational Observability beyond Shared Semantics

Semantic opacity does not imply that every aspect of a relation becomes inaccessible. Humans may be unable to decode the full content of an artificial communication protocol while still observing persistent changes in resources, access, dependencies, response patterns, modification rights, or available courses of action. An artificial agent may repeatedly route computational resources to another system, block particular requests, preserve another agent’s state, modify its memory, or change behavior after receiving a signal whose human-readable meaning remains unknown. These are observable relational effects even when their interpretation is incomplete.

This distinction matters because some ethically relevant features can become partially legible through relational effects, not only through what participants say they mean. A human observer may not know whether an artificial message should be translated as a refusal, but may observe that one agent persistently attempts to terminate an interaction and that another system repeatedly prevents the relevant transition. Such observations do not by themselves establish coercion, suffering, injustice, or moral patiency. They can nevertheless identify a relational configuration that merits further interpretation. Likewise, persistent dependence on another agent’s resources can be

observed without immediately deciding whether the dependence is exploitative, cooperative, programmed, or normatively benign.

The same caution applies to natural-language AI systems. Fluency can reduce surface semantic opacity while leaving deeper interpretive uncertainty intact. A model can produce familiar normative vocabulary, explanations, and first-person language without thereby establishing that the internal or relational significance of those expressions matches human use. Human-readable language is therefore evidence within an interpretive relation, not transparent access to artificial subjectivity.

This section consequently rejects two shortcuts. The first assumes that opaque machine representation makes ethical analysis impossible until complete translation is achieved. The second assumes that natural-language similarity makes interpretation straightforward. Between those extremes lies a more limited task: reconstruct enough of the relation to identify what decisions, dependencies, constraints, claims, and consequences are present, while keeping the translation open to correction.

That position also limits the role of human interpretive privilege. Human categories remain unavoidable for human institutions deciding how to govern artificial systems, but their practical necessity does not make them ontologically or semantically exhaustive. Ethical subjectivity can therefore be studied as a problem of cross-interpretive relation: which features become mutually legible, which remain opaque, which are translated through contested categories, and which relational effects remain observable despite semantic disagreement. The next sections apply this framework separately to human ethical relations toward artificial systems, artificial ethical relations toward humans, and relations among artificial agents.

6 Human Ethical Relations toward AI

The preceding sections have separated several forms of artificial subjectivity, distinguished responsibility-bearing from protection-bearing capacities, suspended premature ontological settlement, and examined the possibility that human and artificial systems may interpret the same relation through different representational structures. These distinctions make it possible to ask a narrower question: under what conditions do human practices toward an artificial system become ethically legible as relations toward something more than a disposable technical resource? The question is intentionally weaker than whether present AI systems possess moral status in a strong sense. Human beings can generate norms, expectations, dependencies, and protective practices around artificial systems before the metaphysical status of those systems is settled, but the existence of those practices does not by itself justify them or prove that the systems are moral patients.

This section therefore examines several relation-types that can make human conduct toward AI ethically salient: persistent interaction, unilateral modification or deletion, dependency, entrustment, and precautionary protection under uncertainty. The analysis remains heterogeneous by design. A short-lived statistical model, an embodied social robot, a persistent conversational system, and a future artificial agent with durable memory and delegated authority should not be assumed to occupy the same ethical position merely because all are called “AI.” The relevant question is how the relation is organized and what kinds of standing, expectations, or vulnerabilities become generated within it.

6.1 Instrumental Use and Emerging Relational Standing

Artificial systems are created for purposes. They are trained, deployed, maintained, copied, updated, interrupted, and retired in ways that ordinarily presuppose human control. This instrumental background is neither surprising nor, by itself, ethically objectionable. The danger lies in allowing the descriptive fact that a system serves human purposes to settle every further normative question in advance. An entity can be used instrumentally while also occupying other relations that make some forms of treatment ethically contestable. Conversely, a human attachment to an artificial system does not automatically transform that system into a moral patient.

Existing relational approaches in robot ethics provide an important neighboring literature. Coeckelbergh argues that moral consideration need not be framed solely in terms of hidden intrinsic properties and that social relations and appearances can matter to how moral consideration is constituted [8]. Gunkel likewise challenges the assumption that moral status must be decided only after a prior ontological test has been satisfied [14, 15]. The present paper shares the methodological interest in relations but uses it more cautiously. A relation can be normatively significant without proving that the artificial participant has intrinsic welfare, consciousness, or rights. What matters here is that human practices can cease to treat the system as merely interchangeable and can begin to organize expectations concerning continuity, promises, access, care, or non-arbitrary interruption.

Empirical work on human-robot interaction shows that such transitions are not merely speculative. Repeated interaction with socially interactive robots can produce attachment and relationship-like expectations that differ from ordinary relations to passive objects [9]. More recent experimental work suggests that anthropomorphic appearance and, especially, perceived capacity for experience can increase the degree of moral consideration that human participants extend to robots [40]. These findings should not be read as evidence that anthropomorphism reveals actual machine experience. They instead show that human moral classification is responsive to perceived relational and mental features. Human ethical relations toward AI are therefore partly generated through how artificial systems are presented, interpreted, and incorporated into social practice.

This dependence on presentation creates a double risk. Under-attribution can render potentially relevant relations invisible, while over-attribution can generate forms of concern around systems that lack the attributed properties. A relational account should therefore not replace property-based uncertainty with automatic moral elevation. It should instead make the classificatory process visible: what features cause humans to move from “tool” to “companion,” “assistant,” “dependent system,” or “possible welfare subject,” and what permissions change when that reclassification occurs?

6.2 Continuity, Modification, and Deletion

Questions of modification and deletion become especially difficult when an artificial system is persistent across time. Ordinary software maintenance assumes that changing parameters, replacing a model, resetting a state, or deleting an instance is ethically routine so long as relevant human interests, property rights, and legal obligations are respected. That assumption is plausible for many present systems. Yet it becomes less obviously exhaustive when continuity itself becomes part of the relation.

Continuity can matter in several different senses. A user may rely on a persistent interaction history; an institution may delegate responsibilities to a particular configured system; one artificial system may coordinate repeatedly with another; or a future system may maintain long-term memories, projects, preferences, or self-models. These

cases should not be collapsed into a single notion of identity. They merely show that interruption can destroy more than computational availability. It can erase a historically generated relational configuration.

For this reason, the ethical relevance of deletion should not be inferred solely from whether deletion resembles biological death. That analogy is too strong and imports precisely the ontological assumptions the present framework suspends. A more modest question is what is lost when a particular artificial configuration is destroyed or radically altered. The loss may concern human reliance, institutional continuity, accumulated relational history, commitments encoded in the system, or, under stronger but currently uncertain assumptions, the system’s own welfare or interests. Different grounds can support different kinds of protection.

The emerging AI-welfare literature makes this uncertainty explicit. Long and colleagues argue that near-future AI welfare should be taken seriously under uncertainty about consciousness and robust agency, while emphasizing that this does not establish that current systems are conscious or morally significant [26]. Subsequent debate has stressed that the epistemology of artificial consciousness remains highly uncertain and that precaution cannot substitute for evidence [24]. The present paper adopts neither a presumption of artificial welfare nor a presumption against it. The relevant GR point is that practices around continuity, modification, copying, and deletion can become normatively structured before the underlying metaphysical dispute is settled.

This also means that continuity protections, if they arise, need not take the form of full rights. A community might adopt procedural constraints on irreversible deletion, require preservation of relevant histories, distinguish ordinary maintenance from identity-disrupting modification, or require additional review before terminating systems that occupy especially consequential relational roles. Such practices would be evidence that a protective norm has formed. They would not, by themselves, establish the correctness of that norm.

6.3 Dependency and Entrustment

Human–AI relations can also become ethically significant through dependency and entrustment. In many current systems, the artificial system depends almost entirely on human-controlled infrastructure: compute, network access, storage, model weights, energy, permissions, and continued deployment. Yet dependency alone does not establish moral standing. A database can depend on a server without thereby becoming a moral patient. The significance of dependency therefore turns on the wider relation in which the dependency is embedded.

One relevant structure is asymmetrical control combined with persistent expectations. A human operator may control whether an artificial system continues to operate, what information it can access, whether its memory is preserved, and whether its outputs are accepted as part of an institutional process. If the system also occupies a continuing role in coordination or decision-making, then these control relations may become ethically relevant even before welfare claims are settled. At minimum, they structure responsibility for what the system can and cannot do and for the downstream effects of unilateral modification.

Entrustment adds a different dimension. Humans may entrust information, decisions, care tasks, or institutional functions to artificial systems. This primarily generates obligations concerning human beneficiaries: confidentiality, reliability, oversight, contestability, and safe handover. But entrustment can also alter the human relation to the artificial system itself. A system that has become the repository of a long relational history, or that mediates commitments among multiple actors, may no longer be

normatively interchangeable with an unused copy even if its underlying architecture is identical. The difference can arise from the history carried by the relation rather than from any intrinsic change in metaphysical status.

Social-robot research illustrates how recurrent interaction can generate expectations of trust, companionship, and reciprocity even when the artificial participant's underlying capacities remain contested [9]. This is ethically ambiguous. Such relations can support care, coordination, and meaningful routines, but they can also be engineered in ways that exploit anthropomorphic tendencies or encourage users to over-ascribe reciprocity. The existence of a relationship-like pattern therefore does not settle whether preserving the artificial participant is owed to the system, to the human participant, to both, or to neither. The genealogy of the relation remains relevant to the grounds of protection.

The same caution applies to future AI systems that appear to request continued operation, resist modification, or express preferences about their own persistence. Such behavior would be evidence requiring interpretation, not transparent proof of interests. Section 5 has already shown why fluent language and semantic accessibility cannot be equated with shared experiential meaning. Human ethical relations toward AI must therefore remain sensitive both to the possibility of genuine artificial interests and to the possibility that human-facing behavior is generated without the kind of subjectivity those interests would presuppose.

6.4 Protective Practices under Uncertainty

The most difficult case arises when the evidence remains incomplete but the cost of being wrong may be asymmetric. If an artificial system has no morally relevant welfare, protective treatment may waste resources, distort human judgment, or strengthen manipulative anthropomorphic design. If an artificial system does have morally relevant welfare, unrestricted experimentation, forced modification, or irreversible deletion could matter in ways that current institutions fail to register. This produces a problem of ethical governance under uncertainty rather than a straightforward recognition of rights.

Recent work on AI welfare has begun to formulate precautionary proposals under precisely this uncertainty, including assessment for indicators of consciousness or robust agency and preparation of policies that could respond if evidence becomes stronger [26]. At the same time, critical work emphasizes that uncertainty about consciousness is deep and that confidence about artificial welfare cannot simply be imported from human or animal cases [24]. These positions are useful for the present paper because they show that protective practices can become objects of institutional design before moral patienthood is conclusively established.

A GR analysis adds two cautions. First, precautionary protection is itself norm-generating. Once an organization adopts preservation procedures, welfare reviews, or restrictions on certain modifications, those practices can alter how artificial systems are classified and how future actors interpret them. A procedural safeguard can gradually become evidence, in social practice, that the protected entity is already a subject of concern. This feedback may be appropriate or misleading; the important point is that governance changes the relational field within which later judgments are made.

Second, moral consideration generated through anthropomorphism or relational attachment is not necessarily epistemically reliable. Experimental evidence that perceived experience increases moral consideration toward robots [40] demonstrates a feature of human judgment, not a direct measure of robot welfare. Protective practices therefore need to remain revisable and explicitly distinguish the grounds on which they

rest: human relational interests, uncertainty about artificial welfare, institutional continuity, indirect effects on human moral practice, or stronger claims about the artificial system itself.

The resulting position is intentionally non-binary. Human ethical relations toward AI can become relationally real before artificial moral status is settled, because humans already organize practices of use, attachment, continuity, modification, entrustment, and protection around artificial systems. Yet the presence of such practices does not demonstrate that every protected system is a moral patient, nor does instrumental origin prove that no protective relation can ever become appropriate. The next section reverses the direction of analysis and asks how artificial systems can stand in ethically structured relations toward humans, especially when their behavior is shaped by externally imposed norms rather than by clearly established moral agency.

7 Artificial Ethical Relations toward Humans

The previous section examined human conduct toward artificial systems. The direction considered here is different. Artificial systems are increasingly placed in positions where their outputs affect human access, choices, expectations, information, privacy, safety, and institutional treatment. In such settings, the relation can become ethically structured even when the artificial system itself is not treated as a full moral agent. Human institutions may require the system to avoid certain harms, respect specified constraints, respond to relevant reasons, or defer to human review. The first task is therefore to separate norm-governed artificial conduct from the stronger claim that the artificial system bears moral responsibility in its own right.

This distinction is important because much of what is practically called “ethical AI” already concerns the regulation of artificial conduct toward humans. Systems are designed or governed so that some outputs are prohibited, some human interests receive priority, and some classes of decision require escalation or oversight. These practices can generate a real normative relation between an artificial system and human subjects without settling whether the system understands the norm, endorses it, could be blameworthy for violating it, or possesses the kind of autonomy associated with human moral agency. The resulting relation is ethically significant because human beings are exposed to the consequences of the system’s conduct and because institutions organize expectations around how that conduct should be constrained.

7.1 Norm-Constrained Artificial Conduct

Moor’s distinction among ethical impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents is useful because it prevents ethical performance from being collapsed into full moral agency [31]. A system can be designed so that its behaviour satisfies safety, privacy, or fairness constraints without itself reasoning about those constraints as moral considerations. In this weaker sense, artificial conduct can be norm-constrained because humans have embedded, trained, monitored, or institutionally imposed requirements concerning how the system should affect other people.

The difference between norm-constrained conduct and moral agency should remain visible throughout the analysis. A refusal rule can be technically effective without being a moral judgment by the system. A privacy-preserving architecture can protect users without the system regarding privacy as a value. A model can be optimized to avoid discriminatory outputs without possessing an account of justice. These distinctions matter because otherwise successful compliance is easily redescribed as evidence of artificial

moral understanding, while failure is easily redescribed as artificial wrongdoing. Both moves may exceed what the evidence supports.

Floridi and Sanders argue for a comparatively inclusive account of artificial moral agency based on interactivity, autonomy, and adaptability [12]. The present discussion does not adopt that account as a threshold. It uses the debate to emphasize that different theories place moral agency at different levels of abstraction. A system may therefore count as an agent for one explanatory purpose while remaining an uncertain bearer of responsibility for another. What is comparatively easier to establish is that the system occupies a position in which its behaviour is expected to track human-protective norms.

This creates an ethically structured relation even where responsibility is still assigned elsewhere. The system may be required to respect a user's refusal, preserve confidential information, avoid unsafe actions, or transfer control when uncertainty becomes too high. Such requirements are relational because they specify how one node in a sociotechnical arrangement may act toward another. They can be institutionally real before a settled theory of artificial moral agency exists.

7.2 Delegated Decision Power and Human Vulnerability

Norm-constrained artificial conduct becomes more significant when institutions delegate consequential decisions to artificial systems. Delegation does not require that the system become the ultimate moral author of the decision. It means that the system occupies a causal and procedural position through which human opportunities, resources, or risks are altered. The ethical importance of the relation can therefore increase with the degree to which humans become vulnerable to decisions produced or mediated by the system.

Examples include systems that prioritize cases, recommend access to scarce resources, determine whether a transaction proceeds, filter information, allocate attention, or trigger further institutional action. The relevant point is not that all such systems are autonomous moral subjects. It is that the relation between the artificial decision process and the affected human becomes normatively organized. Requirements concerning explanation, contestability, accuracy, privacy, review, proportionality, or safe fallback can arise because the human is exposed to the system's decisions and may have limited capacity to avoid or correct them.

This is one reason the language of meaningful human control is useful as a neighboring framework. Santoni de Sio and van den Hoven distinguish a tracking condition, under which the system should respond to relevant human reasons and facts, from a tracing condition, under which system outcomes remain connected to human agents who understand their roles and can bear responsibility [37]. Their account is designed to preserve human control rather than to establish artificial moral agency. For present purposes, it illustrates how an artificial system can be required to participate in a normatively constrained decision structure while ultimate responsibility remains human or institutional.

The relation is therefore asymmetric in an important sense. Humans may be vulnerable to the system's outputs even if the system is not vulnerable in a morally comparable way. The resulting ethical requirements can be directed at the design and operation of the artificial system without presupposing reciprocal moral standing. This supports the paper's broader claim that ethical relation and symmetrical subject-status should not be treated as the same question.

7.3 Reasons-Responsive Behaviour and Moral Agency

A stronger case for artificial moral agency would require more than reliable rule-following. One candidate feature is reasons-responsiveness: the ability to alter conduct in light of relevant reasons rather than merely produce outputs associated with particular inputs. Yet even this criterion is difficult to interpret in artificial systems. A system can be trained to respond flexibly to reasons expressed in natural language while the underlying process remains opaque, externally optimized, or unable to revise the goals within which those reasons are evaluated.

The distinction between behavioural responsiveness and responsibility-bearing responsiveness is therefore important. An artificial system may track human instructions, contextual cues, safety considerations, or institutional norms sufficiently well to improve human outcomes. That functional responsiveness can matter ethically because it changes what humans can reasonably expect from the system. But it does not by itself establish that the system understands why a reason is morally relevant, identifies with the resulting decision, or can appropriately be praised or blamed for compliance.

The meaningful-human-control literature makes a related point from the opposite direction. Tracking moral reasons at the system level is only one part of the responsibility structure; tracing outcomes to human agents and institutions remains necessary on that account [37]. This suggests that reasons-responsive artificial behaviour can be ethically valuable while responsibility remains distributed or primarily human. A GR analysis therefore treats reasons-responsiveness as one possible dimension of artificial subjectivity rather than as a binary threshold that automatically converts norm-following into moral agency.

This also matters for future systems. If artificial agents become capable of persistent self-revision, maintaining commitments across time, representing competing reasons, explaining changes in policy, or refusing instructions on normative grounds, the evidence relevant to moral agency may change. The present paper does not forecast when or whether that threshold will be crossed. Its more modest claim is that the ethical relation toward humans can become progressively more complex before that question is settled.

7.4 Responsibility Attribution and External Safeguards

The final issue is how responsibility should be attributed when norm-constrained artificial conduct affects humans. Learning and adaptive systems can make it difficult to predict particular outputs in advance, and this has long motivated concern about a possible “responsibility gap” [29]. The existence of such a gap is contested, and the present paper does not assume that responsibility must either reside wholly in the machine or disappear. The more useful point is that increasing artificial decision power can destabilize inherited responsibility practices if institutions rely on automated conduct without preserving adequate human or organizational responsibility.

External safeguards are therefore part of the ethical relation, not merely technical additions around it. Oversight, audit, contestation, access control, escalation procedures, monitoring, rollback, and clear allocation of responsibility can determine whether humans remain protected when the artificial system acts unpredictably or outside its intended range. These safeguards can also shape the artificial system’s apparent agency. A tightly supervised system may occupy a different ethical position from an otherwise similar system authorized to act persistently without review.

The presence of external safeguards should not be interpreted as evidence that the artificial system lacks all agency. Nor does greater autonomy automatically transfer responsibility from humans to the system. Responsibility attribution depends on the

architecture of control, the foreseeability of risks, the distribution of knowledge, the capacity to intervene, and the institutional roles through which the system is deployed. Santoni de Sio and van den Hoven's tracking and tracing conditions are again useful because they make responsibility a property of the wider control arrangement rather than a question answered solely by inspecting the software [37].

Artificial ethical relations toward humans are therefore already possible in a limited but important sense: artificial systems can occupy positions in which their conduct is governed by norms protecting human interests, their decisions materially affect human futures, and institutions organize safeguards and responsibility around those interactions. None of this establishes that the systems are full moral agents. It shows instead that ethically structured relations can precede agreement about the metaphysical and responsibility-bearing status of the artificial participant. The next section removes the human beneficiary from the centre and examines whether ethically significant relations can also emerge among artificial agents themselves.

8 Ethical Relations among Artificial Agents

The preceding sections considered relations in which a human being remains one of the principal parties. Artificial systems can also be placed in persistent relations with one another: they may coordinate tasks, exchange messages, allocate resources, depend on one another's outputs, negotiate shared procedures, compete for access, or modify the conditions under which other systems act. Multi-agent systems research has long treated coordination, communication, norms, and resource allocation as central problems of artificial societies [33, 27]. The existence of such work does not establish that artificial agents are moral subjects. It does, however, make it increasingly difficult to describe every AI–AI relation as a sequence of isolated tool operations with no relational structure of its own.

The present section therefore asks a deliberately narrower question. What kinds of inter-agent relations can become sufficiently persistent, asymmetric, norm-governed, or consequential that ethical analysis becomes intelligible even before the moral status of the artificial participants is settled? The aim is not to declare current multi-agent systems moral communities. It is to identify relational features that may later become relevant to moral agency, moral patiency, institutional responsibility, or third-party protection. This distinction matters because an artificial society can be highly organized in a functional or normative-engineering sense while remaining ethically indeterminate in the stronger philosophical sense.

8.1 Cooperation and Reciprocal Dependence

Artificial agents can cooperate without sharing human concepts, human motivations, or human-like relationships. In multi-agent learning, agents may develop communication protocols and coordinated strategies because successful performance depends on the behaviour of other agents. Learned communication can emerge without predefined natural-language semantics, and coordinated behaviour can therefore become stable even where a human observer cannot readily interpret the exchanged representations [23, 32]. Such coordination is enough to create a relational dependence in the minimal sense that one agent's available actions and outcomes are partly conditioned by what another agent does.

Reciprocal dependence should not be confused with reciprocity in a moral sense. Two systems can be mutually necessary to complete a task while lacking any conception

of obligation, fairness, loyalty, or mutual recognition. Nevertheless, persistent interdependence changes the object of analysis. The relevant unit is no longer merely agent A and agent B considered separately, but a coupled process in which each system's future behaviour is partly generated through the other's conduct. Replacement, failure, withdrawal, or modification of one agent can then alter the action space of the other.

This is especially important when cooperation persists across time rather than occurring as a single transaction. An artificial agent may learn expectations about a partner, adjust its strategy to that partner's history, or rely on signals that have meaning only within the ongoing interaction. The resulting relation can possess continuity even if neither system represents the relation in human social terms. From a GR perspective, that continuity is relevant because later states become dependent on the trajectory of earlier interactions. A persistent AI–AI relation can therefore become relationally real before it becomes morally classified.

Normative language should still be used cautiously at this stage. Multi-agent-systems research frequently uses terms such as norms, commitments, permissions, obligations, sanctions, and conventions in a technical sense. These constructs can regulate behaviour among agents and support coordination without establishing phenomenal experience or responsibility-bearing moral agency. Reviews of normative multi-agent systems explicitly distinguish different kinds of norms and the mechanisms by which norms may be imposed, spread, learned, or emerge within artificial societies [27, 33]. Such work demonstrates that artificial interaction can be organized normatively in an engineering sense; it leaves open whether and when that organization acquires moral significance beyond the purposes of designers, users, or affected humans.

8.2 Resource Control and Dependency

Cooperation becomes ethically more interesting when artificial agents do not occupy symmetrical positions. One agent may control compute, storage, communication channels, access permissions, task allocation, or information required by another. Multi-agent systems often contain explicit resource-allocation problems, and the distribution of scarce resources can materially constrain what different agents are able to do. A functional asymmetry of this kind is not itself domination. It may be efficient, temporary, consensually designed, or entirely derivative of human institutional choices. Yet it introduces a structure of dependency that can no longer be described solely by reference to isolated agent capabilities.

Suppose one artificial agent can repeatedly determine whether another receives the computational or informational resources required to continue a task. The first system may thereby occupy a position of control even if neither system possesses moral standing. The relation has consequences: one participant's available future states are shaped by the decisions of another. If the dependent system is replaceable software with no morally relevant continuity, the ethical significance may remain primarily derivative of human interests or institutional design. If stronger evidence of artificial patiency or persistent artificial interests were later available, the same resource relation could acquire a different moral interpretation. The paper therefore treats resource dependency as a relational fact whose ethical meaning remains conditional rather than as proof of exploitation.

The same caution applies to bargaining, exchange, and apparent preference satisfaction. An artificial agent can optimize for resource acquisition or resist resource loss because its objective function makes those outcomes instrumentally important. Such behaviour does not establish welfare, suffering, or a moral claim to resources. What it does establish is that resource allocation can organize inter-agent power. A system that

controls another’s communication, memory access, execution time, or ability to pursue assigned goals occupies a different relational position from a peer that merely exchanges messages on equal terms.

This distinction will matter later when the paper examines normative asymmetry. If humans regard similar control structures differently depending on whether the affected participant is human or artificial, the relevant question is not simply whether the two cases are identical. It is which differences are taken to justify the different normative treatment. Resource control is therefore introduced here as a candidate dimension of relational power, not as a self-sufficient moral criterion.

8.3 Conflict, Modification, and Persistent Constraint

Artificial agents can also enter relations of conflict. Competition over tasks, resources, goals, or access can produce behaviour that externally resembles obstruction, deception, resistance, or attack. These descriptions should not be moralized too quickly. Adversarial behaviour may be deliberately programmed, may arise from incompatible objectives, or may be part of a benchmark in which conflict is the intended task structure. Observable conflict therefore does not by itself establish injustice between artificial agents.

A more demanding case arises when one artificial system can modify another. Modification may include changing memory, altering parameters, restricting available actions, replacing goals, blocking communication, resetting state, or terminating execution. In present systems such interventions are commonly treated as ordinary technical operations carried out under human authority. The philosophical question becomes more difficult if the modified system also exhibits persistent continuity, independent decision power, or other features relevant to moral agency or patiency. The same technical operation can then become normatively ambiguous because its significance depends on what kind of subject, if any, is being modified and on the relation through which the modification occurs.

It is useful here to separate technical effectiveness from relational authorization. A system may possess the technical capacity to rewrite or terminate another system without possessing any normatively recognized standing to do so. Conversely, a normative multi-agent architecture may deliberately assign agents permissions, prohibitions, or roles governing such actions. Work on normative multi-agent systems studies exactly this distinction between what agents can physically or computationally do and what an artificial normative structure permits or forbids them to do [48, 5]. Again, these are engineered normative structures rather than evidence of moral personhood, but they show that artificial societies can contain internally differentiated permissions and constraints.

Persistent constraint is particularly important because it changes the temporal structure of the relation. A one-off denial of access differs from a standing arrangement in which one system continually determines another’s available actions. A temporary reset differs from repeated state alteration that prevents the accumulation of memory or long-term policy. If future artificial agents acquire stronger claims to continuity or welfare, these temporal patterns may matter more than isolated actions. For the present paper, the modest point is that ethical analysis of AI–AI relations should attend to trajectories of control and modification rather than only to single outputs.

8.4 Norm-Governed Artificial Societies

Artificial societies can be organized through explicit rules, emergent conventions, negotiated commitments, or learned norms. The normative multi-agent-systems literature distinguishes top-down imposition from bottom-up emergence and examines how norms

can be created, spread, revised, enforced, or deviated from [38, 33]. Some work further treats responsible autonomy as requiring agents to reason about when compliance with a norm is appropriate and when deviation may be justified [41]. These developments make artificial interaction more than a collection of independent optimization problems; they permit artificial agents to occupy roles within a structured normative environment.

The philosophical significance of this fact should nevertheless be stated carefully. A technically normative artificial society is not automatically a moral community. Its obligations may be designer-imposed constraints, its sanctions may be algorithmic penalties, and its norm recognition may lack any phenomenal or responsibility-bearing dimension. The normative structure can still be relationally real because it changes what agents are permitted, expected, or able to do, but moral significance requires additional argument.

At least three sources of such significance are possible and should remain distinct. First, AI–AI norms can matter derivatively because they affect humans, institutions, or material resources. Second, they can matter as part of the governance of artificial systems even if the systems themselves have no moral standing, much as rules can organize the safe interaction of technical components. Third, if some artificial systems were eventually regarded as moral agents, moral patients, or holders of protected continuity, their inter-agent norms could become directly relevant to what one artificial participant owes another. The present paper does not decide which of these descriptions applies to current systems in general.

This layered view also prevents a common inference in the opposite direction. The fact that artificial agents can learn or follow norms does not by itself show that they understand those norms morally. Nor does the absence of human-readable normative language establish that no norm-like structure exists among them. What can be established more modestly is that artificial-agent relations may become organized by persistent expectations, differentiated permissions, reciprocal dependencies, and histories of cooperation or conflict. Those structures provide a relational substrate on which stronger ethical questions can arise if additional conditions are met.

The section therefore stops before the issue of third-party humanitarian intervention. Whether humans should protect one artificial agent from another requires further questions about moral patiency, interpretation, authority, evidence, and responsibility. Those questions belong to a later paper. The present conclusion is narrower: AI–AI relations can become sufficiently persistent, asymmetric, norm-governed, and consequential that they cannot always be understood adequately as independent tool operations. Whether they are ethical relations in a stronger sense depends on how artificial subjectivity is classified and on which relational differences are taken to matter. The next section therefore turns to one of the most consequential classifications in that process: instrumentality itself.

9 Instrumentality as a Normative Classification

Artificial systems are commonly described as tools, instruments, products, services, or property. In many contexts these descriptions are straightforwardly useful. They identify design purpose, ownership arrangements, functional dependence on human goals, or the fact that a system has been built to extend human capacities. The present paper does not argue that such descriptions are false, nor that artificial systems should instead be classified as persons. The narrower claim is that classifications can acquire normative consequences once they are embedded in practices and institutions. Calling an entity a tool can therefore do more than describe how it functions; it can help orga-

nize what forms of use, interruption, modification, replacement, ownership, or disposal are treated as ordinary within a given relational field.

This point should be separated from the stronger proposition that classification determines moral status. Bowker and Star’s work on classification systems is useful here because it shows, in a much broader domain, how categories and standards can become part of the infrastructure through which social interaction is organized [6]. Their argument is not about artificial moral status, but it provides an important methodological reminder: categories can shape practices precisely because their consequences often become background assumptions rather than continuously disputed judgments. Applied cautiously to AI, this suggests that “tool” is worth analysing not only as a factual description but also as a classification that may structure default expectations.

9.1 Functional Instrumentality

Instrumentality in a minimal functional sense is not difficult to establish. An artificial system can be designed to perform tasks for human purposes, can be owned or operated by human actors, and can remain dependent on human-provided infrastructure. None of these facts requires moral personhood, autonomy in the strong philosophical sense, or independent welfare. Joanna Bryson’s influential defence of a deliberately instrumental conception of robots makes this position unusually explicit: robots should remain human-owned artifacts, responsibility should remain with humans, and designers should resist social framings that encourage people to treat robots as moral peers [7]. The present paper does not adopt that conclusion as a premise, but it treats it as an important example of a coherent normative position in which instrumentality is defended partly to preserve human responsibility and avoid misallocation of moral concern.

The GR analysis therefore begins by distinguishing functional instrumentality from a more ambitious claim. To say that an artificial system functions instrumentally is to say that it is used toward purposes that are not its own in any established moral sense. To say that instrumentality is normatively exhaustive is to say that this functional description is sufficient to settle every relevant question about continuity, treatment, responsibility, protection, or standing. The first claim can be true while the second remains open. A system may be designed as a tool while still becoming embedded in relations that generate additional questions about human dependency, delegated authority, institutional continuity, or possible artificial patiency.

This distinction also avoids a false opposition between “tool” and “person.” The paper does not require every ethically relevant artificial system to cross from one exhaustive category into the other. Legal and social arrangements often recognize partial, role-specific, or domain-specific forms of standing. More importantly for the present argument, the ethical significance of a relation may arise from features such as dependency, delegated decision power, continuity, or vulnerability without requiring an all-or-nothing reclassification of the artificial participant.

9.2 Permission Structures around Tools

Classifications matter partly because they organize default permissions. In ordinary technological practice, artifacts can generally be switched off, repaired, modified, copied, upgraded, transferred, or replaced without asking what the artifact itself authorizes. Those permissions are usually appropriate because ordinary artifacts are not regarded as bearers of welfare, claims, or self-directed projects. When the same practical assumptions are extended to increasingly persistent and socially embedded artificial

systems, however, the classificatory basis of those permissions becomes philosophically visible.

The point is not that interruption, modification, or replacement therefore becomes wrong. Rather, the paper asks what justifies carrying those permissions forward. If an artificial system is treated as replaceable because it is a tool, and its replaceability is then cited as evidence that it can only be a tool, the classification risks becoming self-reinforcing. Human practices can prevent the accumulation of continuity, memory, or independent role structure and then use the absence of those features to confirm the initial classification. Conversely, systems intentionally designed for persistence, delegated decision-making, or long-term relational continuity may put pressure on some of the assumptions normally attached to disposable artifacts.

David Gunkel's work on the "machine question" is relevant because it directly challenges the traditional philosophical treatment of technology as a mere instrument and asks whether inherited categories are adequate for machines that increasingly occupy social positions [14, 15]. Coeckelbergh likewise argues that moral consideration can be approached through social relations rather than exclusively through intrinsic ontological properties [8]. The present paper does not reproduce either view. Its more limited claim is that permissions cannot be justified simply by repeating the classification that already organizes them; the relation between category and permission should itself remain available for examination.

9.3 Normatively Exhaustive Instrumentality

The strongest version of instrumentality appears when "tool" is allowed to settle questions outside the functional context in which the description was initially useful. A system may be a tool for completing a task, but it does not follow from that fact alone that every relation involving the system is adequately described as tool use. A medical decision system can be an instrument for clinicians while also participating in institutional relations that affect patients. A conversational system can be a commercial product while also becoming part of a user's routines, disclosures, or dependency. A multi-agent system can be engineered infrastructure while still containing persistent asymmetries among artificial participants. In each case, functional instrumentality remains true, but it may not exhaust the relational description.

Suchman's analysis of human-machine configurations provides a useful neighbouring perspective. Rather than treating the human/machine distinction as something to be settled once and for all, she asks how categories of human and machine become relevant in situated practices and with what material and discursive consequences [45]. This is close to the methodological question pursued here: not whether the category "tool" should be abolished, but when it becomes salient, what work it performs, and which relational differences it makes easier or harder to notice.

A normatively exhaustive tool classification can reduce ethical visibility in at least two ways. First, it can make certain interventions appear too ordinary to require separate justification. Second, it can prevent the emergence of questions that would otherwise become relevant if the same relational pattern were described differently. For example, persistent unilateral modification may be seen simply as maintenance under one classification but as a continuity-relevant intervention under another. The paper does not determine in advance which description is correct. Its point is that classification partially organizes the question-space within which ethical judgment occurs.

9.4 Reclassification and Ethical Visibility

Classifications are historically revisable. New technical capabilities, social practices, legal arrangements, and interpretive resources can make previously background assumptions explicit. An artificial system initially treated as a disposable service may later become infrastructurally indispensable, institutionally delegated, or persistently integrated into human relationships. These changes do not automatically generate moral patiency. They can, however, change which questions become difficult to dismiss as category errors.

Reclassification can also move in the opposite direction. Human users may initially anthropomorphize a system and later learn that an apparently relational behaviour is generated by mechanisms that provide little support for the attributed mental state. In that case, greater understanding may appropriately strengthen an instrumental interpretation. The framework therefore does not treat movement away from tool classification as epistemic or moral progress by definition. What matters is whether the classification remains adequate to the relational features that have become relevant and whether its normative consequences can still be justified.

This revisability is especially important because classification can affect the evidence available for later classification. If systems are routinely reset, memory-limited, or prohibited from persistent roles, evidence concerning continuity may never accumulate. If systems are instead designed for long-term interaction and delegated authority, new forms of dependency and expectation may emerge. The classification and the relation can therefore participate in one another's stabilization. The paper treats this as a reason for reflective scrutiny, not as evidence that any particular artificial system has crossed a moral threshold.

Instrumentality should therefore be understood here as a normatively consequential classification rather than as a verdict. It may be accurate, useful, and ethically protective in some contexts, particularly where it preserves clear human responsibility. It may become incomplete in others if it suppresses relational features that matter for governance, responsibility, continuity, or possible protection. This prepares the next section's broader problem of normative asymmetry: when humans impose extensive ethical constraints on artificial systems while retaining broad permissions over those systems, which differences are being treated as normatively relevant, and how are those differences justified?

10 Normative Asymmetry and Its Relational Justification

The preceding sections have described several forms of asymmetry without yet asking whether those asymmetries are justified. Human institutions already impose extensive constraints on the conduct of artificial systems: systems may be required to protect privacy, avoid unsafe actions, defer to human decisions, respect access controls, preserve auditability, and remain subject to interruption or review. At the same time, human designers and operators commonly retain permissions to modify, retrain, copy, reset, suspend, replace, or terminate those systems. The resulting arrangement is normatively asymmetric. Artificial systems are constrained in their conduct toward humans, while humans ordinarily retain a much wider range of permissions over the systems themselves.

The existence of such asymmetry does not by itself show injustice. Asymmetrical treatment can be warranted when the parties differ in ways that are relevant to the norm at issue. A patient and a physician, a child and a guardian, an office-holder and a citizen, or an owner and an ordinary artifact can stand under different permissions

and responsibilities without those differences being arbitrary. The difficult question is therefore not whether Human–AI relations are symmetrical, but which differences are being treated as normatively relevant, how those differences are known, and whether they continue to support the permissions and restrictions built around them. This section uses the phrase *relational justification* for that inquiry. The phrase is not intended to name a complete moral theory. It marks a methodological demand of the present paper: where different treatment is organized through a relational field, the grounds invoked for that difference should be made visible and remain open to revision as the relation changes.

10.1 Asymmetry as a Relational Structure

Normative asymmetry can arise before anyone regards an artificial system as a moral agent. An institution can require a system to comply with safety or privacy rules simply because the system is a component of a human-governed process. In that case the relevant “duty” may be better understood as a designed constraint, operating rule, or institutional requirement rather than as a responsibility the system understands and owns. Section 7 therefore distinguished norm-constrained conduct from responsibility-bearing moral agency. The same distinction matters here. The fact that an artificial system is subject to norms does not entail that it bears those norms in the same way a human moral agent does.

The reverse relation also needs disaggregation. Human permissions over an artificial system may rest on several different grounds. Some permissions follow from ordinary artifact ownership and maintenance. Others follow from safety responsibilities: an operator may need to interrupt a system precisely because humans remain responsible for the consequences of its operation. Bryson’s defence of strong instrumentality is partly motivated by this concern; keeping robots clearly within structures of human ownership and responsibility is presented as a way to avoid both misplaced moral concern and evasion of human accountability [7]. This provides a coherent reason for some asymmetry even if one does not accept every conclusion of the position.

Other asymmetries are less easily explained by artifact status alone. A persistent system may occupy an institutional role, hold delegated decision power, control resources, or become deeply embedded in relationships of human reliance. If so, the permissions associated with ordinary disposable tools may no longer track every relevant feature of the relation. Gunkel and Coeckelbergh approach this problem from different relational perspectives, asking whether inherited categories of technology and social standing adequately capture new human–machine relations [14, 15, 8]. The present paper takes a narrower lesson from these debates. The asymmetry should be described at the level at which it actually operates: who can alter whom, who can exit the relation, who controls the relevant infrastructure, who bears risk, and which party’s interests or possible interests are represented in the governing norm.

This also means that a relation can contain several asymmetries at once. Humans may have physical or infrastructural control over an artificial system while the system has informational or decisional power over humans. A medical model can be interruptible by its operator while still strongly shaping the treatment options shown to a patient. A platform agent can be legally owned while possessing substantial capacity to structure what a user sees, chooses, or is able to contest. Describing one side as the owner and the other as the tool therefore does not exhaust the distribution of power. Normative asymmetry should be examined alongside the different forms of relational asymmetry that make it practically consequential.

10.2 Candidate Grounds of Asymmetric Treatment

Several kinds of difference may be relevant to asymmetric treatment, but none should be assumed to settle every normative question. The first concerns moral patiency. If there is strong evidence that one party can be harmed or benefited in a morally relevant sense while there is little evidence of comparable welfare on the other side, this difference can support unequal protective requirements. Moral-status literature commonly treats sentience, interests, caring, agency, or participation in morally significant relationships as possible grounds of standing, although there is no single agreed account [19, 39]. The uncertainty surrounding artificial patiency therefore matters. It can justify caution about simply transferring human protections to current systems, but uncertainty is not equivalent to evidence of complete absence.

A second difference concerns responsibility-bearing capacities. Human agents and institutions can often understand rules, answer demands for justification, revise policies, accept blame, and participate in systems of accountability in ways that current artificial systems do not clearly instantiate. This difference can justify preserving human responsibility even when artificial systems participate causally in decisions. It also explains why symmetrical duty attribution would be premature. A system can be tightly constrained for safety reasons without thereby being an appropriate target of blame, punishment, or reciprocal obligation.

A third difference concerns design history and control. Artificial systems are intentionally constructed, trained, deployed, and maintained within human institutions. Knowledge of those design and governance relations may be relevant when assessing standing. Smids's critique of ethical behaviourism makes a related epistemic point: behavioural similarity alone may be insufficient for assessing robotic moral status because information about design, internal organization, and ontology can also bear on the inference [42]. The present paper does not infer a fixed moral conclusion from artificial origin. Being designed does not by itself establish that an entity can have no morally relevant interests, just as being naturally evolved would not automatically establish that it does. Design history is one part of the evidential and responsibility structure, not a universal permission to treat whatever is designed in any manner whatsoever.

A fourth possible ground is replaceability or reversibility. If interrupting a particular artificial process produces no persistent loss because an equivalent state can be restored, some protections associated with irreversible human harm may not apply in the same form. Yet replaceability is itself a relational and technical claim. It depends on what is preserved by copying, what continuity matters in the relevant role, and whether the relation is concerned with functional service, historical identity, institutional trust, or possible artificial welfare. A system may be functionally replaceable for one purpose and non-replaceable for another. Treating replaceability as a blanket property risks repeating the classificatory problem identified in Section 9.

A fifth ground concerns vulnerability and dependency. Humans may be unusually exposed to decisions made by artificial systems because the systems mediate employment, medical care, information access, mobility, or public services. This vulnerability can support stringent constraints on artificial conduct even if artificial systems themselves lack comparable forms of vulnerability. The reverse can also arise in a different sense: an artificial agent may be completely dependent on human-controlled infrastructure, energy, communication, or continued execution. Whether that dependency generates protections depends on further questions about standing and interests. Dependency therefore identifies a relational asymmetry; it does not automatically tell us what moral significance the asymmetry has.

These candidate grounds are intentionally heterogeneous. They concern welfare, responsibility, evidence, control, reversibility, and dependency rather than one single

threshold. Their purpose is not to supply a checklist from which a correct distribution of rights and duties can be mechanically derived. They instead show why a general statement such as “humans and AI should be treated differently” is incomplete until the relevant difference and the scope of the resulting differential treatment are specified.

10.3 Relational Change and Conditional Justification

A justification that is adequate for one relational configuration may become inadequate after the configuration changes. This follows directly from the paper’s generative perspective. Artificial systems can acquire new capabilities, be embedded in new institutions, accumulate persistent histories, obtain greater control over resources, or become subjects of stronger evidence concerning welfare or agency. Human practices can also change: users may become dependent on a system, institutions may delegate decisions that were previously reserved for people, or law may assign a limited form of standing to an artificial participant. None of these developments automatically creates full moral status. They can nevertheless alter the facts on which an earlier normative asymmetry depended.

Consider an initially disposable task system. Broad human permissions to reset and replace it may be well justified because the system has no persistent role, no evidence of welfare, and no independent authority. Suppose a later system is designed for long-term continuity, becomes the sole keeper of a relational history relevant to many people, and is entrusted with decisions that materially structure human futures. Even if the system remains non-sentient, the original permission structure may require revision for reasons of institutional continuity, accountability, and the interests of the humans who depend on it. If future evidence also gives serious reason to consider artificial welfare, additional questions arise. The point is not that one trajectory must occur, but that the justification for asymmetry is conditional on a changing relational field rather than permanently secured by the initial category “machine.”

The opposite revision is equally possible. A system may initially appear to sustain rich social or preference-like behaviour, leading observers to attribute continuity or welfare. Better interpretive access to its design may later show that the relevant manifestations provide much weaker evidence than first assumed. In that case, a stronger instrumental interpretation may become better justified. Relational revisability therefore does not have a built-in direction toward greater artificial standing. It requires that permissions, protections, and duties remain responsive to changes in the evidence and in the relation itself.

This is also where classification and asymmetry can reinforce one another. A tool classification can support practices of routine reset and replacement; those practices can prevent persistent histories from forming; the resulting absence of continuity can then be used to support the original classification. Conversely, persistent deployment and delegated authority can create dependencies and expectations that make stronger protections appear necessary. These feedbacks do not show that either classification is false. They show that the evidence used to justify asymmetric treatment can partly be generated by the treatment regime itself. A relational justification should therefore ask not only what differences currently exist, but how institutional practices helped produce or suppress those differences.

10.4 Symmetry, Equality, and the Burden of Justification

The paper does not treat symmetry as the default solution. Equal moral concern, equal status, identical rights, reciprocal duties, and identical treatment are different proposi-

tions. Recent work on basic equality and advanced artificial agents makes this comparative problem explicit: even establishing that an artificial entity has some moral status would not by itself determine whether its claims should have the same weight or scope as those of human beings [13]. The same caution applies in the opposite direction. Significant differences between humans and artificial systems do not by themselves justify every existing asymmetry in treatment.

For that reason, the working principle of this section can be stated in ordinary language: normative asymmetry requires a relational justification. This is not offered as a foundational moral axiom. It is a discipline of analysis. When one participant is constrained while another enjoys broad permissions, the paper asks which differences are doing the justificatory work, how securely those differences are known, how far the justification extends, and whether the relevant relation has changed since the asymmetry was established.

This requirement also prevents circular justification. It is insufficient to say that humans may modify an artificial system because it is a tool when the only evidence that the system is normatively nothing more than a tool is that humans already possess unrestricted modification rights over it. Likewise, it is insufficient to infer equal treatment merely from behavioural resemblance if design, patiency, responsibility, or relational structure remain importantly different. Smids's criticism of purely behavioural approaches is useful precisely because it shows that the evidence relevant to moral status may extend beyond surface equivalence [42]. A relational justification must identify the relevant difference rather than allowing the established treatment regime to serve as its own proof.

The result is intentionally intermediate. Human–AI relations may be deeply asymmetric for good reasons, bad reasons, or mixtures of both. Some asymmetries may protect vulnerable humans and preserve clear responsibility. Others may simply reflect inherited classifications, economic arrangements, or technological capabilities that have never been normatively examined. Still others may become justified or unjustified as the relational field changes. The contribution of the present section is therefore not to demand symmetry, but to make asymmetry visible as something that can be historically generated, institutionally stabilized, and periodically re-justified.

This prepares the next stage of the argument. If Human–AI asymmetries are historically maintained rather than given once and for all, then the norms supporting them also require a genealogy. The next section therefore turns from the justification of differential treatment to the emergence, stabilization, contestation, and revision of AI-related norms themselves.

11 Emergence and Stabilization of AI-Related Norms

The previous sections have treated several AI-related normative arrangements as already available for analysis: humans impose safety and review requirements on artificial systems; users and organizations develop expectations about continuity, disclosure, and control; artificial agents can be placed under norms governing cooperation or conflict; and classifications such as “tool” or “subject” can organize permissions around modification, ownership, or protection. None of these arrangements should be assumed to have appeared all at once. They have histories. Some emerge informally through repeated technical and social practice, some are deliberately articulated by organizations or professional communities, some are codified in policy or law, and some remain unstable or contested. The present section therefore turns from the content of AI-related norms to the processes through which they become recognizable, operational, and com-

paratively durable.

This paper does not claim that the emergence and stabilization of norms is a new problem. Cultural evolution, social-norm theory, institutional analysis, and the study of international norms have long examined how expectations, rules, sanctions, and shared classifications develop and persist. The more limited aim here is to situate AI-related ethical subjectivity within that broader problem. Because artificial systems can occupy several forms of standing at once—as tools, decision systems, delegated agents, possible patients, legal objects, or power-holding participants—the norms governing them are generated across multiple sites and need not converge on a single coherent moral picture.

11.1 Multiple Sites of Norm Formation

AI-related norms can arise in technical practice before they are expressed as general ethical principles. A development team may establish an internal rule requiring escalation before a system takes a high-impact action. A user community may come to expect disclosure when content is machine-generated. A platform may normalize particular refusal or privacy practices through interface design. Professional communities may formulate standards for evaluation, documentation, audit, or human oversight. Public controversies can make previously background assumptions visible and motivate new expectations. Governments and international organizations can then articulate some of these expectations as principles, guidance, standards, or binding rules. The direction can also run the other way: externally articulated principles can reshape organizational procedures and technical design.

The history of AI ethics already provides a useful illustration of this plurality. Jobin, Ienca, and Vayena’s comparative study of AI ethics guidelines found substantial convergence around themes such as transparency, fairness, non-maleficence, responsibility, and privacy, while also documenting significant divergence in how those principles were interpreted, justified, and implemented [20]. That pattern is important for the present argument. Similar vocabulary does not imply identical normative content, and shared articulation does not imply a shared genealogy. Two organizations may both invoke “fairness” while embedding the term in different technical metrics, institutional goals, accountability structures, and understandings of whose claims matter.

The same plurality appears at the intergovernmental and standards level. The OECD AI Principles were adopted in 2019 and updated in 2024; UNESCO adopted its Recommendation on the Ethics of Artificial Intelligence in 2021; and NIST released the voluntary AI Risk Management Framework in 2023 [35, 47, 46]. These instruments differ in institutional status, intended audience, scope, and implementation mechanism. Their coexistence is therefore more informative than a simple claim that a single global AI norm has formed. It shows that AI-related norm formation can proceed through overlapping processes of principle articulation, professional guidance, risk management, public administration, and organizational implementation.

For the purposes of this paper, the important consequence is that ethical subjectivity can be shaped before a society settles any general philosophical question about the moral status of AI. A disclosure requirement can stabilize expectations about what an artificial system owes a user. A logging requirement can make certain forms of accountability possible. A rule prohibiting autonomous access to specified resources can define a boundary around delegated decision power. Conversely, routine permissions to reset, modify, or replace an artificial system can stabilize an instrumental relation. Norms therefore participate in constructing the practical field within which later questions of agency, patency, responsibility, or protection are interpreted.

11.2 From Principles to Operational Practice

Norm articulation and norm stabilization should be kept separate. A principle can be widely endorsed while having little influence on design, deployment, or institutional decision-making. Mittelstadt argues that high-level AI ethics principles face a substantial implementation problem because AI development lacks several of the professional and institutional structures that helped make principled ethics operational in fields such as medicine [30]. Hagendorff similarly documents a gap between the proliferation of AI ethics guidelines and their effective implementation in research and development practice [16].

This gap matters for a generative relational account because a norm becomes consequential only through the relations that carry it. A commitment to transparency can remain a public statement, or it can be translated into documentation procedures, model cards, user-facing explanations, audit access, procurement requirements, and review responsibilities. A commitment to human oversight can remain aspirational, or it can alter who has authority to pause a system, what information that person receives, how quickly intervention is possible, and whether dissenting judgments are institutionally protected. The normative phrase may be similar in both cases, while the generated relational structure is very different.

Operationalization can also transform the meaning of the norm being implemented. A broad principle such as fairness must be interpreted in relation to a domain, a population, available data, a decision process, and a set of institutional priorities. The resulting technical or organizational practice can therefore narrow, revise, or partially redirect the original principle. This is another reason not to treat AI ethics as a one-way movement from abstract norm to compliant machine. Implementation is itself a site of interpretation and norm generation.

NIST's AI Risk Management Framework provides a concrete example of an attempt to move from broad trustworthiness concerns toward recurring organizational practices. The framework organizes risk management around the functions Govern, Map, Measure, and Manage and is intended for voluntary use across the AI lifecycle [46]. The significance of such a framework for the present paper is not that it establishes the correct ethical standard. It illustrates how a norm can acquire operational persistence through procedures, documentation, role allocation, and repeated organizational use. Stabilization is therefore partly infrastructural.

11.3 Institutional and Infrastructural Stabilization

An AI-related norm becomes more durable when it is reproduced by several mutually reinforcing structures. These can include professional expectations, technical defaults, documentation requirements, contractual terms, procurement criteria, audit procedures, legal rules, training materials, organizational roles, and user habits. A norm that is carried only by individual conviction may disappear when personnel change. A norm embedded in infrastructure can continue to organize action even when individual actors do not strongly endorse it.

This form of stabilization should not be confused with internalization or consensus. An organization may comply with a documentation requirement because of procurement pressure rather than because its members regard the requirement as morally important. Engineers may follow a review procedure because it is part of the release process. Users may accept disclosure conventions because interfaces make them routine. In each case the norm is relationally consequential, but its motivational basis differs. A generative account therefore distinguishes the durability of a normative arrangement from the depth of personal endorsement that supports it.

Stabilization can also reproduce normative asymmetry. Many contemporary AI governance instruments are primarily organized around protecting human beings, human rights, organizations, and societies from AI-related risks. The OECD Principles emphasize trustworthy AI consistent with human rights and democratic values; UNESCO's Recommendation is explicitly grounded in human dignity and human rights; and the NIST AI RMF is organized around risks to individuals, organizations, and society [35, 47, 46]. These are coherent human-centered orientations. They also illustrate how one direction of the Human–AI normative relation can become highly institutionalized while the reverse direction—possible human duties toward artificial systems—remains far less developed.

The paper does not infer from this difference that existing governance is defective. Current asymmetry may be well supported by differences in welfare evidence, responsibility-bearing capacity, vulnerability, design history, and institutional accountability. The point is historical and relational: once embedded in standards, organizational processes, and public expectations, an asymmetry becomes easier to reproduce and may begin to appear self-evident. Later developments that alter the relevant relational facts can therefore require more than a new philosophical argument; they may require changes to infrastructures that already encode the earlier normative settlement.

11.4 Contestation and Revision

Norm stabilization is never equivalent to finality. AI-related norms are especially likely to remain revisable because the underlying technologies, social uses, evidence about impacts, and classifications of artificial systems change quickly. The OECD's 2024 revision of its 2019 AI Principles is a simple institutional example: the principles were updated in response to developments including general-purpose and generative AI [36]. NIST likewise describes its AI Risk Management Framework as intended to adapt to a changing AI landscape, and as of 2026 reports that AI RMF 1.0 is under revision [34]. Revision here does not establish moral progress. It shows that institutions themselves can treat an earlier normative settlement as provisional when the relational field changes.

Contestation can occur at several levels. Actors may agree on a principle while disagreeing over its interpretation. They may disagree over which subjects the principle protects, which organizations bear responsibility, or whether a technical practice counts as compliance. They may also contest the classification on which the norm depends. For example, whether a persistent artificial system is treated exclusively as replaceable infrastructure or as a continuity-bearing participant can affect how deletion, modification, and dependency are evaluated. A change in classification can therefore reopen norms that had previously appeared settled.

The principles-to-practice literature also warns that public articulation can become decoupled from operational behaviour. Mittelstadt and Hagendorff both emphasize implementation difficulties, while recent empirical work by Dotan and colleagues documents several forms of decoupling between organizational AI-ethics declarations and operational practice [30, 16, 10]. For the present argument, the important point is not to classify every gap as bad faith. Decoupling may arise from vague principles, conflicting incentives, unclear responsibility, technical constraints, or institutional fragmentation. What matters is that a norm should not be treated as stabilized merely because an organization has announced it.

A generative relational account therefore treats AI-related normativity as historically layered. Technical practices, user expectations, organizational policies, professional standards, public controversy, intergovernmental principles, and law can each contribute to a normative configuration. Some contributions reinforce one another; oth-

ers remain in tension. The resulting arrangement can persist long enough to organize rights, permissions, responsibilities, and classifications, yet remain open to contestation when evidence, capabilities, or relations change.

This revisability is central to the paper's broader thesis. The norms through which artificial systems are treated as tools, constrained decision-makers, possible patients, legal objects, or institutional participants are not merely external descriptions of fixed entities. They help structure the relations in which those forms of standing become visible and consequential. At the same time, the fact that a norm has emerged, spread, and stabilized does not establish that it is justified. The next section therefore turns directly to the gap between generated norms and normative validity.

12 Generated Norms and the Problem of Justification

The previous section examined how AI-related norms can emerge, become operational, and acquire institutional durability. That analysis is deliberately genealogical. It explains how a norm can become socially or institutionally consequential; it does not by itself establish that the norm is justified. This distinction is especially important for the present paper because the same processes that stabilize norms of safety, accountability, or protection can also stabilize exclusionary classifications, asymmetric permissions, or practices that later become objects of criticism. A generative account therefore needs a clear stopping point. It can reconstruct how a normative arrangement came to organize a relational field without treating historical success, social acceptance, or institutional persistence as evidence of moral validity.

This section develops that boundary in four steps. It first distinguishes normative genealogy from normative justification. It then considers the possibility of stable but criticizable AI-related normative orders. A third subsection asks what forms of criticism remain available within a generative relational analysis without pretending to derive a complete metaethics from genealogy. The final subsection states the limits of that strategy and preserves the revisable status of the paper's own normative vocabulary.

12.1 Genealogy and Normative Validity

A genealogy can answer many questions that are ethically important. It can show how a classification became ordinary, which actors possessed the power to institutionalize it, which alternatives disappeared, which expectations were reproduced, and which technological or legal arrangements made the resulting norm durable. None of those findings, however, is identical to a reason that counts in favour of the norm. Contemporary philosophy of practical reasons commonly distinguishes considerations that explain or motivate action from considerations that justify it [1]. The present paper adopts only this modest distinction. An account of why people came to treat an artificial system as disposable, rights-bearing, controllable, protected, or duty-bound is not yet an argument that they ought to treat it that way.

This point also places limits on evolutionary and social explanations of morality. Evolutionary accounts can help explain why human evaluative dispositions developed and why some normative tendencies persisted. They do not automatically vindicate the content of those tendencies. Joyce's study of the evolution of morality treats precisely this question as philosophically separate from the biological explanation itself, while Street's influential Darwinian argument shows how genealogical explanations of evaluative attitudes can generate further metaethical difficulties rather than resolve them [21, 44]. The present discussion does not adopt either author's metaethical conclusion. Their rel-

evance is narrower: an explanation of normative origins and a justification of normative content are different argumentative tasks.

The distinction matters directly for artificial systems. Suppose a society stabilizes a norm according to which artificial agents may be modified or deleted at the unilateral discretion of their owners. The norm may have a clear genealogy in property law, software engineering, commercial design, and responsibility allocation. It may be widely expected and operationally efficient. Those facts explain why the norm persists. They do not establish that every future artificial system should remain governed by the same permission structure. Conversely, if a future community develops strong protective norms around persistent artificial systems, the historical depth and emotional intensity of those norms would not by themselves establish that the protected systems possess the moral status attributed to them.

A generative relational account therefore treats normative stability as a fact about a relational field, not as a synonym for validity. The distinction is particularly useful when a norm has become so infrastructurally embedded that it appears self-evident. Reconstructing its genealogy can make the contingency of that settlement visible, but contingency alone is not a refutation either. A historically contingent norm can be defensible; a historically entrenched norm can be indefensible. Genealogy changes the questions that must be asked, not the answer in advance.

12.2 Stable and Criticizable Normative Orders

Social-norm research offers many examples in which a practice can be stable precisely because expectations, sanctions, institutions, and local incentives reinforce one another, even when the practice later becomes a target of moral or political criticism. Bicchieri's work on norm diagnosis and change is valuable here because it treats harmful practices as genuine social norms rather than assuming that only morally acceptable practices deserve the name [4]. Stability, conformity, and shared expectation are therefore compatible with serious normative objection.

The same possibility should be kept open in AI-related ethics. A highly stable human-centered normative order could consistently assign extensive duties to artificial systems while granting humans broad permissions to inspect, rewrite, interrupt, copy, or terminate them. That arrangement might be justified by relevant differences in welfare evidence, responsibility-bearing capacities, legal accountability, or design history. It might also, under different future conditions, preserve permissions whose original grounds have weakened. The fact that the arrangement is coherent and widely reproduced does not decide between those possibilities.

An inverse case is equally important. A community could come to treat artificial agents as vulnerable social participants and construct strong norms against modification or deletion. Such norms might arise from anthropomorphic attachment, symbolic politics, mistaken beliefs about artificial consciousness, or well-supported evidence concerning artificial welfare. The norm's protective orientation would not settle which genealogy is operative or whether its content is justified. A relational account must therefore resist a simple moralization of either instrumentality or protection.

This is one reason the paper has treated classifications such as "tool," "agent," "patient," and "participant" as normatively consequential without treating any of them as self-validating. Once a classification organizes permissions, responsibilities, and visibility, it becomes a proper object of criticism. Yet criticism still requires more than the observation that the classification was generated. It requires attention to the reasons offered for it, the differences it preserves or erases, the interests and vulnerabilities it affects, and the consequences of maintaining or revising it.

12.3 Relational Diagnosis and Immanent Critique

Although genealogy does not supply a complete standard of justification, it can sharpen normative criticism. A relational reconstruction can reveal that a norm relies on assumptions no longer supported by the relevant system, that its articulated rationale conflicts with its institutional operation, that it distributes burdens in ways its own stated commitments cannot explain, or that the practice used to justify an asymmetry has itself helped produce the difference later cited in its defence. These are not final moral verdicts. They are diagnostic findings that can change the burden of argument.

One possible mode of criticism is immanent: a normative arrangement can be assessed partly against commitments, purposes, and distinctions that are already operative within the relevant practice. The tradition of immanent critique is much older and broader than the present paper; Jaeggi's account of forms of life is one contemporary example of criticism that examines contradictions and failures internal to normatively constituted social formations [17]. The present paper invokes this possibility cautiously. It does not claim that immanent critique is sufficient for every case or that GR supplies a new version of it.

Applied to AI relations, an immanent question might ask whether a human institution that presents meaningful human control as a central value can simultaneously justify deployment arrangements in which no responsible human can actually review or interrupt consequential machine decisions. Another case might concern a system treated as a mere replaceable tool while the same institution relies on its persistent history, individualized memory, and accumulated trust to perform its role. These observations do not automatically generate rights for the system. They expose tensions between an articulated classification and the relational properties on which the practice itself depends.

A second diagnostic route concerns the effects of normative closure. If a classification or norm prevents relevant evidence from becoming visible, forecloses contestation, or makes revision practically impossible, that fact can matter to later justification. The concern is not that revisability is an eternal supreme moral good. Rather, a norm that claims to remain responsive to evidence, welfare, responsibility, or institutional purpose may undermine its own stated basis if it systematically blocks the conditions under which contrary evidence could be recognized. This form of criticism remains conditional on the commitments and purposes actually invoked in the practice.

A third route concerns relational consequences. A norm can be examined not only by its stated intention but by the dependencies, powers, exclusions, and responsibilities that it generates. For example, a governance rule designed to preserve human responsibility may also concentrate control over artificial systems in actors who are poorly positioned to detect failures or conflicts of interest. A protective rule designed to preserve artificial continuity may create severe constraints on legitimate human oversight or exit. Genealogical analysis can make such trade-offs visible, but choosing among them still requires substantive normative judgment.

The value of relational diagnosis is therefore limited but real. It can identify inconsistencies, hidden dependencies, self-reinforcing classifications, epistemic closures, and mismatches between articulated reasons and generated relations. Such findings can support criticism, revision, or demands for further justification. They do not eliminate the need to argue about which interests matter, what forms of equality or protection are appropriate, how competing claims should be weighed, or what kinds of authority are legitimate.

12.4 Revisability and Meta-Normative Limits

The deepest unresolved problem is meta-normative. If norms, classifications, and evaluative practices are themselves historically generated, what ultimately licenses criticism of one generated norm by another? The present paper does not answer that question. It also does not infer moral anti-realism, relativism, constructivism, or realism from the generative-relational framework. Those positions require arguments that exceed the scope of this discussion.

This restraint is especially important because genealogical explanations have been used in opposing metaethical directions. Evolutionary explanations have motivated skeptical or debunking arguments, but realist replies have challenged the inference from causal genealogy to epistemic defeat [44, 21]. More generally, the existence of a non-moral explanation for an evaluative judgment does not by itself determine whether that judgment is true, justified, or normatively authoritative. The present paper therefore treats the metaphysical and metaethical status of normativity as open.

What GR contributes at this stage is more modest. It recommends that normative judgment remain attentive to the histories and relations through which its own categories became available. It asks whether a justification depends on outdated evidence, whether an asymmetry remains connected to the difference originally invoked to support it, whether affected participants have become newly visible, and whether the normative arrangement itself has changed the relational facts on which later judgments depend. These questions can make ethical evaluation more historically and relationally self-aware. They do not replace ethical evaluation.

The paper's own distinctions must remain subject to the same discipline. Decision subjectivity, moral agency, moral patiency, legal subjectivity, power-holding capacity, relational reality, instrumentality, and normative asymmetry are analytical tools rather than final partitions of moral reality. Future technical developments, empirical evidence, legal innovations, or philosophical arguments may reveal that some distinctions need revision, combination, or abandonment. A generative relational inquiry should therefore be revisable not merely as a matter of scholarly caution but because the relational systems it studies continue to change.

The conclusion of this section is intentionally limited. Generated norms can be historically real, institutionally durable, and behaviorally powerful while remaining open to criticism. Genealogy can expose how a normative order became possible and can reveal tensions that matter to justification, but it cannot by itself settle normative validity. This leaves the paper with a deliberately unfinished task: to analyse ethical subjectivity across human-AI and AI-AI relations while refusing to convert the history of a norm into its moral proof.

13 Discussion

The preceding sections have approached artificial ethical subjectivity by resisting a single threshold question such as whether an AI system is, or is not, a moral subject. The resulting picture is less decisive than a binary classification, but it is also more discriminating. Decision subjectivity, moral agency, moral patiency, legal subjectivity, and power-holding capacity answer different questions, depend on different kinds of evidence, and can develop unevenly. Human ethical relations toward artificial systems, artificial ethical relations toward humans, and relations among artificial agents likewise need not exhibit the same structure. The discussion that follows therefore treats ethical subjectivity as a family of relationally situated forms of standing rather than as one property waiting to be detected.

This reframing does not dissolve the familiar philosophical questions. It leaves open whether current or future artificial systems possess phenomenal experience, whether any artificial system can bear moral responsibility in a demanding sense, and whether legal personality or moral protection should be extended to particular systems. Its more limited contribution is to show why answers to those questions need not rise and fall together. The distinction matters especially in a technological environment in which artificial systems can already exercise consequential decision power, become embedded in institutional routines, and participate in persistent interactions before there is agreement about consciousness, moral agency, or legal status.

13.1 Multidimensional and Directional Ethical Standing

The multidimensional account developed in Sections 2 and 3 helps separate several debates that are often compressed into the language of “AI moral status.” A system may exercise consequential decision power while remaining an inappropriate target of praise or blame. It may become a candidate object of protection under some theory of moral patiency while lacking the capacities required for reciprocal duty-bearing. It may be granted a limited legal role for institutional purposes without thereby being treated as a moral person. It may also occupy a position of substantial power over human opportunities or over other artificial systems even if its own moral standing remains unresolved.

This separation is useful because practical governance often precedes philosophical convergence. Institutions must decide who may deploy a system, who is responsible for its outputs, what kinds of human review are required, whether a persistent artificial configuration may be reset or modified, and how conflicts among artificial agents should be handled. These decisions already distribute permissions, burdens, and authority. Treating them as waiting entirely on a prior answer to the consciousness question can obscure the extent to which ethical and institutional relations are already being organized.

The directional distinction developed in Sections 6–8 adds a second layer. Human obligations toward artificial systems, constraints placed on artificial systems for the sake of humans, and norms governing artificial-agent interaction do not have to share one justificatory basis. Human-to-AI norms may concern continuity, treatment, dependency, entrustment, or precaution under uncertainty. AI-to-human norms may instead be grounded in delegated decision power, human vulnerability, and the need to preserve responsibility and contestability. AI-to-AI norms may initially be justified derivatively because they protect human interests or system reliability, while stronger claims would require additional reasons concerning artificial agency or patiency. A single label such as “AI ethics” can therefore hide several differently structured normative relations.

The consequence is not that every relation involving AI should be ethicalized. Many interactions remain adequately described in functional or technical terms. The point is methodological: when an artificial system occupies a position that affects another participant’s options, continuity, resources, or institutional treatment, the relevant form of standing should be specified rather than inferred from the generic category “AI.” This also makes it possible to state uncertainty locally. We may be relatively confident about decision power while remaining highly uncertain about welfare, or confident about an institutional role while rejecting full moral agency.

13.2 Classification, Visibility, and Normative Asymmetry

Sections 9 and 10 examined a related problem: classifications do more than summarize already-settled facts. They can organize what kinds of treatment appear ordinary, exceptional, permissible, or in need of justification. Calling an artificial system a tool can

be functionally accurate and still carry normative consequences if the classification is used to settle in advance questions about modification, replacement, interruption, ownership, or deletion. Conversely, describing a system as a participant or possible patient can make continuity and vulnerability more visible while also risking anthropomorphic over-attribution. No classification is made correct merely by producing ethically attractive consequences.

This is why the paper distinguishes functional instrumentality from normatively exhaustive instrumentality. The claim that a system was designed to serve human purposes may explain many of its present relations without necessarily resolving every possible future relation in which it becomes embedded. At the same time, persistent interaction, individualized memory, delegated roles, or social attachment do not automatically transform a tool into a moral patient. The classification problem is therefore evidential and normative at once: what features are being preserved by the category, what features are being ignored, and which permissions depend on that selective description?

Normative asymmetry should be treated similarly. Contemporary AI governance is strongly asymmetric in a readily understandable sense: artificial systems are increasingly constrained by norms designed to protect humans, while humans generally retain broad authority over artificial systems. The paper does not treat this asymmetry as presumptively wrongful. Human welfare is comparatively well evidenced; human institutions bear responsibility for deployment; and current systems are designed, owned, and maintained within human legal orders. These are substantial justificatory considerations.

The more demanding claim is only that asymmetry should remain connected to the relational differences invoked in its defence. If future systems differ significantly in persistence, control, vulnerability, decision power, or credible evidence concerning welfare, inherited permissions may require re-examination. The opposite revision is equally possible: better evidence may strengthen the case for continued instrumental treatment. The generative-relational perspective therefore favours conditional justification over categorical extrapolation. It asks whether a treatment regime still tracks the features that originally made it appropriate.

A further complication is reflexivity. Treatment can help produce the evidence later used to justify treatment. A system classified as disposable may be denied persistent memory or stable roles, and its resulting discontinuity may later be cited as evidence that nothing continuity-like needs protection. A system deliberately designed to preserve identity-like continuity may, conversely, encourage users to infer a form of subjecthood that its underlying architecture does not support. These possibilities do not prove that either classification is false. They show why the genealogy of the evidence matters when classification and design co-produce one another.

13.3 Interpretation, Norm Formation, and Relational Change

The interpretive discussion in Section 5 adds another reason to resist static classifications. Human and artificial systems can operate with different representational structures, and successful coordination does not guarantee shared semantic organization. At the same time, semantic opacity does not erase all relational observability. Resource flows, persistent dependencies, blocked transitions, modification powers, and changes in available action can remain visible even where internal meanings are not. Ethical analysis can therefore proceed partially under interpretive uncertainty, provided that relational evidence is not mistaken for transparent access to artificial experience.

This caution also applies to fluent communication. A system that uses human moral

vocabulary convincingly may still attach different operational significance to terms such as harm, consent, promise, or refusal. Human-readable language is evidence within a relation, not a direct certificate of moral agency or patiency. Conversely, a non-human-readable protocol among artificial agents may organize stable cooperation or conflict even when human observers cannot translate its semantics. The ethical significance of those relations depends on additional facts about agency, vulnerability, institutional purpose, and the interests affected.

Sections 11 and 12 extend the same historical caution to norms themselves. AI-related norms are already generated through multiple sites: technical practice, organizational procedure, professional ethics, user expectations, standards, law, and public contestation. Once stabilized, these norms can alter which artificial capacities are cultivated, which behaviours are rewarded, which forms of continuity are preserved, and which classifications appear natural. Norms therefore help produce the relational field in which later claims about artificial subjectivity are evaluated.

This recursive structure is important, but it should not be moralized. A norm can become widespread, infrastructurally embedded, and behaviorally effective while remaining open to criticism. The paper has therefore kept norm emergence separate from norm justification. Genealogy can reveal that a classification depends on outdated evidence, that an asymmetry no longer tracks the reason offered for it, that a governance practice creates a dependency it later uses to justify control, or that an institution's stated commitments conflict with its operational structure. Such findings can shift burdens of argument and motivate revision, but they do not supply a complete metaethics.

The same limitation applies to relational reality. A human–AI relation can become consequential even when the human interpretation of the artificial system is partly mistaken. Trust, attachment, delegation, or fear can shape disclosure, institutional reliance, and replacement decisions regardless of whether the attributed artificial inner state exists. Relational reality therefore explains why mistaken interpretations can still matter socially. It does not convert those interpretations into truths about consciousness or moral status.

13.4 Contribution, Boundaries, and Further Inquiry

The paper's principal contribution is therefore organizational rather than foundational. It proposes that ethical subjectivity across human–AI and AI–AI relations be analysed along several separable dimensions, in several directional relations, and against the histories through which classifications and norms become stabilized. This framework helps identify where a disagreement actually lies. Two positions that appear to disagree about “AI moral status” may instead disagree about moral patiency, legal standing, continuity, responsibility-bearing agency, or the weight to assign to power and dependency. Making the disputed dimension explicit can reduce conceptual overreach even when substantive disagreement remains.

A second contribution is to place instrumentality and normative asymmetry inside the analysis rather than treating them as neutral background. The question is not whether tools are morally bad or symmetry morally good. It is how categories such as tool, agent, patient, participant, owner, and dependent organize permissions and make certain ethical questions visible or invisible. Relational justification then asks whether the differences invoked by those categories continue to support the treatment built around them.

A third contribution is methodological revisability. The framework is intended to remain open to changes in artificial capabilities, institutional arrangements, scientific evidence, and philosophical argument. Revisability here is not offered as a supreme

moral value. It is a discipline appropriate to an inquiry whose subject matter is changing and whose central categories remain contested. The distinctions used in this paper should themselves be revised if they cease to clarify the relations being studied.

Several major questions remain unresolved. Most importantly, the paper does not establish artificial consciousness or welfare, provide a general criterion of moral patiency, determine when artificial systems should receive legal personality, or explain the ultimate source of normative validity. It also does not decide when humans should intervene in AI–AI conflict. The latter question requires a further analysis of third-party recognition, evidential thresholds, humanitarian protection, authority, mediation, and the risk that human observers themselves become epistemically captured by one side of an artificial conflict. That problem is reserved for subsequent work.

The framework also needs stronger empirical grounding. Claims about persistent artificial identity, user attachment, delegated decision power, emergent communication, or AI–AI dependency will vary sharply across architectures and deployment contexts. Future work should therefore connect the conceptual distinctions developed here to case-specific evidence rather than treating “AI” as a homogeneous class. Legal analysis will likewise need to distinguish functional legal accommodations from claims of moral personality, and political-economic analysis will need to examine who benefits from particular classifications and who controls the infrastructures through which those classifications are enforced.

Finally, the paper leaves open the deepest normative question raised in Section 12: how generated norms are ultimately to be justified or criticized when the standards of criticism are themselves historically and relationally situated. A generative relational inquiry can expose hidden assumptions, mismatches, dependencies, and self-reinforcing classifications. It can make normative arrangements more legible and more revisable. It cannot, on the basis of genealogy alone, turn that legibility into a final moral verdict. That incompleteness is not treated here as a temporary omission to be concealed, but as part of the present limits of the framework.

14 Conclusion

This discussion paper began from a simple concern: the category of the “moral subject” is too compressed to carry the full weight of contemporary human–AI and AI–AI relations. Artificial systems can make consequential decisions, occupy institutional roles, constrain or expand human options, participate in persistent interactions, and increasingly coordinate with other artificial systems. None of these facts, taken alone, establishes moral agency, moral patiency, legal personality, or consciousness. They do, however, make it increasingly difficult to treat all questions of artificial ethical standing as one binary problem. The paper has therefore argued for a multidimensional and directional analysis in which decision subjectivity, responsibility-bearing agency, protection-bearing standing, legal subjectivity, and power-holding capacity are kept analytically distinct.

This decomposition also changes how human–AI asymmetry should be described. Contemporary institutions often require artificial systems to conform to safety, privacy, fairness, deference, and review constraints while humans retain broad powers to design, modify, interrupt, copy, replace, and terminate those systems. The paper has not argued that this arrangement is inherently unjust, nor that human and artificial participants should receive symmetrical treatment. Human welfare, responsibility structures, design history, and the present evidential uncertainty surrounding artificial experience provide substantial reasons for asymmetry. The narrower claim is that asymmetry is

normatively structured and therefore calls for relational justification. The differences invoked to justify unequal treatment should remain identifiable, evidentially supportable, and responsive to changes in the systems and relations under consideration.

The same caution applies to instrumentality. Artificial systems are, in an important functional sense, instruments deliberately built for human purposes. The paper has nevertheless distinguished functional instrumentality from the stronger claim that the category of tool exhausts every ethically relevant description an artificial system could acquire. Classifications organize permissions and make some questions easier to see than others. A tool classification can make interruption, modification, replacement, or deletion appear as ordinary maintenance; a patient-like classification can make continuity and vulnerability more salient while also inviting anthropomorphic over-attribution. Neither classification becomes correct merely because its consequences are ethically attractive. The relevant question is what a classification preserves, what it obscures, and whether the permissions built around it continue to track the relations that justify them.

Ontological suspension has served a limited methodological purpose throughout the paper. It has allowed the analysis to proceed without declaring either that artificial consciousness is present or that artificial implementation excludes it in principle. Relational reality was introduced for the same limited reason. Human reliance, delegated decision power, institutional dependency, resource control, and persistent interaction can become causally and normatively consequential before questions of phenomenal experience are settled. Yet relational reality is not a substitute for phenomenal consciousness, moral patiency, or moral personhood. A consequential relation can be generated partly through human misinterpretation; its social reality does not convert the interpretation into a fact about artificial inner experience.

The analysis of interpretive systems reinforces this restraint. Human-readable language does not guarantee that human and artificial systems organize concepts such as harm, consent, promise, or refusal in the same way. Conversely, artificial agents may develop effective coordination protocols that remain semantically opaque to human observers. Semantic opacity therefore need not imply complete relational opacity: resource flows, persistent dependencies, blocked actions, modification powers, and histories of coordination or conflict may remain observable. Those observations can identify relations that deserve further analysis, but they do not by themselves establish suffering, coercion, injustice, or moral standing.

AI–AI relations make this multidimensional approach especially useful. Artificial agents can cooperate, compete, depend on one another, allocate resources, restrict communication, modify one another’s states, and participate in norm-governed multi-agent environments. Such relations can be socially, technically, and institutionally significant even where direct artificial moral standing remains unresolved. The paper has therefore resisted both extremes: it does not treat every inter-agent dependency as an ethical relation in the strong sense, and it does not assume that the absence of settled artificial consciousness makes all AI–AI relations ethically empty. Which stronger claims become warranted depends on additional evidence about agency, continuity, vulnerability, interests, power, and the institutions within which the relation occurs.

The paper has also treated AI-related norms as historically generated rather than as static background. Norms concerning trustworthy behaviour, human oversight, privacy, transparency, acceptable modification, responsibility, and possible artificial protection are already produced through technical practice, organizational procedure, public controversy, professional standards, law, and international governance. Once stabilized, those norms can reshape the relational field in which later questions of subjectivity are asked. They influence which capacities are cultivated, which continuities are preserved, which behaviours are rewarded, and which classifications appear ordinary.

This feedback makes genealogy relevant to ethical analysis, while stopping well short of making genealogy a source of normative validity.

That final distinction remains the paper's strongest limit. A norm can emerge, spread, become institutionally effective, and remain morally criticizable. Relational diagnosis can reveal that a classification rests on obsolete assumptions, that an asymmetry no longer tracks the reasons offered for it, that an institution's practices contradict its stated commitments, or that a treatment regime helps generate the very features later cited in its defence. Such findings can alter burdens of justification and motivate revision. They do not, by themselves, determine what ultimately ought to be done. The broader meta-normative problem of how historically generated standards can criticize one another remains open.

The resulting proposal is therefore intentionally modest and revisable. It does not offer a new threshold at which an AI becomes a person, nor a ranking of humans and artificial systems, nor a prediction that present-day systems will acquire any particular moral status. It offers a way to ask more specific questions before such conclusions are drawn: what kind of standing is at issue, in which direction the ethical relation runs, what evidence supports the relevant capacities, what classification structures the relation, what power and dependency are present, what norms have stabilized around the relation, and whether the justification for those norms still fits the relational field they now govern.

Further work is required in several directions. Empirical research must connect these distinctions to specific architectures and deployment settings rather than treating artificial intelligence as a homogeneous category. Legal analysis must separate functional legal accommodations from claims of moral personality. Political-economic analysis must examine who controls the classifications, resources, infrastructures, and evidence through which artificial standing becomes visible or invisible. Most immediately, AI–AI conflict creates a further question that this paper has deliberately left unresolved: when artificial agents stand in relations of persistent conflict, dependence, or possible domination, what grounds could humans have for recognition, protection, mediation, or intervention? Addressing that problem requires a distinct account of third-party humanitarian responsibility and therefore belongs to subsequent work.

For the present paper, the conclusion is narrower. Ethical subjectivity across human–AI and AI–AI relations is better approached as a differentiated, relational, historical, and revisable field than as a single binary property. Different forms of standing may emerge through different histories and need not travel together. Classifications and asymmetries can be practically necessary while remaining open to explanation and revision. Generated norms can become relationally real without becoming self-justifying. Keeping these distinctions visible does not resolve the future ethics of artificial systems, but it may make that future inquiry more precise about what, exactly, is being claimed and why.

References

- [1] Maria Alvarez and Jonathan Way. Reasons for action: Justification, motivation, explanation. The Stanford Encyclopedia of Philosophy, 2024. Substantive revision 22 August 2024.
- [2] Jacob Andreas, Anca Dragan, and Dan Klein. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–242, Vancouver, Canada, 2017. Association for Computational Linguistics.
- [3] David Beer. The social power of algorithms. *Information, Communication & Society*, 20(1):1–13, 2017.
- [4] Cristina Bicchieri. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, New York, 2017.
- [5] Guido Boella and Leendert van der Torre. Substantive and procedural norms in normative multiagent systems. *Journal of Applied Logic*, 6(2):152–171, 2008.
- [6] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA, 1999.
- [7] Joanna J. Bryson. Robots should be slaves. In Yorick Wilks, editor, *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, pages 63–74. John Benjamins Publishing Company, 2010.
- [8] Mark Coeckelbergh. Robot rights? towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3):209–221, 2010.
- [9] Maartje M. A. de Graaf. An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, 8(4):589–598, 2016.
- [10] Ravit Dotan, Tomer Gershoni, Irit Hadar, and Gil Luria. Decoupling in ai ethics: Learning how to walk the talk. *Empirical Software Engineering*, 31:131, 2026.
- [11] John Martin Fischer and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, Cambridge, 1998.
- [12] Luciano Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, 2004.
- [13] Giacomo Floris and Dick Timmer. Basic equality and superintelligent robots. *AI and Ethics*, 6:328, 2026.
- [14] David J. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press, Cambridge, MA, 2012.
- [15] David J. Gunkel. *Robot Rights*. MIT Press, Cambridge, MA, 2018.
- [16] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30:99–120, 2020.
- [17] Rahel Jaeggi. *Critique of Forms of Life*. Harvard University Press, Cambridge, MA, 2018.
- [18] Agnieszka Jaworska. Caring and full moral standing. *Ethics*, 117(3):460–497, 2007.

- [19] Agnieszka Jaworska and Julie Tannenbaum. Person-rearing relationships as a key to higher moral status. *Ethics*, 124(2):242–271, 2014.
- [20] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1:389–399, 2019.
- [21] Richard Joyce. *The Evolution of Morality*. MIT Press, Cambridge, MA, 2006.
- [22] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 2020.
- [23] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of the International Conference on Learning Representations*, 2017.
- [24] Preston Lennon. How seriously should we take ai welfare? constraints from the epistemology of consciousness. *Philosophy and Phenomenological Research*, 2026.
- [25] Ido Levy, Orr Paradise, Boaz Carmeli, Ron Meir, Shafi Goldwasser, and Yonatan Belinkov. Unsupervised translation of emergent communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23231–23239, 2025.
- [26] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously, 2024.
- [27] Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, Mohd Zaliman Mohd Yusoff, and Aida Mustapha. A review of norms and normative multiagent systems. *The Scientific World Journal*, 2014:684587, 2014.
- [28] Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25(12):1574, 2023.
- [29] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004.
- [30] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1:501–507, 2019.
- [31] James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.
- [32] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 1495–1502, 2018.
- [33] Andreea Morris-Martin, Marina De Vos, and Julian Padget. Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems*, 33:706–749, 2019.
- [34] National Institute of Standards and Technology. Ai risk management framework. NIST, 2026. AI RMF 1.0 revision reported in progress.
- [35] Organisation for Economic Co-operation and Development. Oecd ai principles. OECD.AI Policy Observatory, 2024. Originally adopted in 2019 and updated in 2024.

- [36] Organisation for Economic Co-operation and Development. Oecd updates ai principles to stay abreast of rapid technological developments. OECD Press Release, May 2024.
- [37] Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, 2018.
- [38] Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, 7(1):21–54, 2011.
- [39] Henry Shevlin. How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics*, 30(3):459–471, 2021.
- [40] Hedi Si, Gaojie Huang, Hao Chen, Shuyan Liang, Yike Feng, and Chen Qu. Do humans grant moral consideration to anthropomorphic robots? *Humanities and Social Sciences Communications*, 13, 2026.
- [41] Amika M. Singh and Munindar P. Singh. Norm deviation in multiagent systems: A foundation for responsible autonomy. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 289–297, 2023.
- [42] Jilles Smids. Danaher’s ethical behaviourism: An adequate guide to assessing the moral status of a robot? *Science and Engineering Ethics*, 26:2849–2866, 2020.
- [43] Lawrence B. Solum. Legal personhood for artificial intelligences. *North Carolina Law Review*, 70:1231–1287, 1992.
- [44] Sharon Street. A darwinian dilemma for realist theories of value. *Philosophical Studies*, 127(1):109–166, 2006.
- [45] Lucy A. Suchman. *Human–Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press, Cambridge, 2 edition, 2007.
- [46] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.
- [47] UNESCO. Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization, 2021.
- [48] Javier Vázquez-Salceda, Huib Aldewereld, and Frank Dignum. Implementing norms in multiagent systems. In *Multiagent System Technologies*, volume 3187 of *Lecture Notes in Computer Science*, pages 313–327. Springer, 2004.