Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Machines Be Punished? Punitive Projection and Artificial Sanction

Gehad Abdelal

abdelal@uga.edu

University of Georgia

Research Article

Keywords: Artificial moral agency, Moral punishment, Normative addressability, Status-responsive punishment, Normative punitive parity, Punitive projection, Distributed accountability, Artificial intelligence

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11372741/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Can artificial agents be morally punished without experiencing suffering? Beginning with a hypothetical Cybercab that fatally strikes a pedestrian and is subsequently dismantled, this paper examines whether technological destruction constitutes punishment and whether such an intervention adequately addresses the victim's family's claims. Through conceptual analysis and moral and epistemic normative analysis, I defend a conditional account of artificial punishability grounded in *normative addressability*, comprising four conditions: normative competence, attributable control, diachronic answerability, and sanctionability. Where these conditions are satisfied and culpable wrongdoing is established, *status-responsive punishment* may consist in an authoritative, retrospective, proportionate, and censure-linked deprivation of an artificial agent's own normative authority. Such deprivation may constitute a *role-relative ontological loss* without requiring phenomenal suffering or the elimination of computational capacities. I introduce *punitive projection* to challenge the treatment of imagined offender suffering as a measure of punitive adequacy, and *punitive theatre* to distinguish symbolic interventions from justified punishment. The resulting principle of *normative punitive parity* holds that artificial origin and the absence of suffering do not independently exclude punitive liability, without implying equal moral status or identical sanctions. I further develop the *Distributed Accountability Constraint* to establish that artificial punishment cannot displace the independent responsibilities of developers, operators, and institutions. Through *consequential recognition*, I distinguish justified punishment from the broader obligations of attribution, repair, and prevention owed to victims. The argument does not presume that existing artificial systems possess moral agency; rather, it establishes the conditions under which punishment without suffering may be morally justified while acknowledging the unresolved objection concerning adversity without welfare interests.

1. Introduction

Consider a hypothetical case in which an autonomous Cybercab strikes and kills a pedestrian at a crossing. The victim's family demands accountability. In response, the manufacturer publicly dismantles the vehicle and announces that the machine responsible for the death has been permanently removed from service. The intervention is visible, irreversible, and directed at the system involved in the collision. Yet an important moral question remains unanswered: has the Cybercab actually been punished, or has the manufacturer merely destroyed a machine? More importantly, what makes either response morally meaningful to the family whose loved one has died? Dismantling the vehicle may prevent further harm, demonstrate institutional concern, or satisfy a public demand for consequences. None of these purposes, however, establishes that the Cybercab was culpable, deserved punishment, or underwent a punitive deprivation. The apparent severity of an intervention cannot substitute for the moral grounds that justify it.

The problem becomes more difficult when we examine the Cybercab's decision before the collision. Suppose it detected movement near the crossing but remained uncertain whether the pedestrian intended to cross. It could have slowed down or stopped, but proceeding would have allowed it to maintain its journey. If the Cybercab recognized an obligation to protect the pedestrian, understood that uncertainty required precaution, possessed sufficient control to act accordingly, and nevertheless prioritized journey efficiency, its decision raises the possibility of culpable wrongdoing. If the collision instead resulted from an undetectable sensor failure, external interference, or circumstances beyond its reasonable control, dismantling the vehicle might

be justified without constituting punishment. In both circumstances, the outcome is the same: a pedestrian is dead. What differs is whether the harmful event can be attributed to an agent capable of acting under binding moral reasons and violating them culpably. The distinction exposes the inadequacy of treating harmful consequences, technological autonomy, or public condemnation as sufficient grounds for punitive liability.

The difficulty is not peculiar to artificial intelligence. Philosophical theories of punishment have long disagreed about the relationship between wrongdoing, suffering, condemnation, and desert. Feinberg [23] emphasizes punishment's expressive function, while Duff [17] understands punishment through moral communication and answerability. Brady [9] defends the significance of suffering within punitive censure, whereas Poama [53] challenges the assumption that affliction is necessary for punishment. Danaher's *retribution gap* [14] introduces a further difficulty in the context of artificial systems: harmful conduct may generate a demand for punishment even when no appropriate subject of moral blame can be identified. These disagreements reveal three questions that are too often treated as one: what makes a consequence punitive, what makes an offender deserving of that consequence, and what makes its imposition morally justified. The possibility of artificial punishment brings their separation into sharper focus because an artificial agent may be subject to restrictions, destruction, or institutional condemnation without necessarily experiencing suffering.

The central question of this paper is whether artificial agents can be morally punished without experiencing suffering and, more specifically, under what conditions the deprivation of their normative authority could constitute a genuinely justified punitive consequence. I defend a conditional affirmative answer. Punishment does not derive its legitimacy simply from the pain experienced by an offender, but neither does the ability to restrict or destroy a machine establish the authority to punish it. Genuine artificial punishment would require an accountable subject, culpable wrongdoing, legitimate punitive authority, and a consequence that meaningfully deprives that same subject of something belonging to its normative standing. My argument does not presume that existing autonomous systems meet these conditions. It seeks to establish what would have to be true for punishment directed at an artificial agent to become morally defensible.

I begin by questioning a familiar assumption underlying punitive judgment: that the suffering of an offender provides evidence that justice has been adequately served. I introduce *punitive projection* to describe the attribution, anticipation, or symbolic construction of offender suffering when that presumed suffering is assigned justificatory significance. Its underlying danger is what I call an *imagined economy of suffering*: the belief that the pain imposed upon an offender can somehow compensate for the suffering caused to a victim. In the Cybercab case, the family's loss is real and irreversible. Destroying the vehicle cannot restore the pedestrian's life, and the appearance of a severe consequence cannot establish that the wrongdoing has been adequately answered. This criticism does not reject retributive punishment or deny that suffering may affect the severity and permissible limits of sanctions. Rather, it challenges the movement from *the offender has suffered* to *the offender has, therefore, been adequately punished*. The moral significance of punitive consequences must be established independently of that inference.

Once suffering is distinguished from punitive justification, a more fundamental question emerges: what would make an artificial system an appropriate subject of punishment in the first place? Floridi and Sanders [27] and Behdadi and Munthe [7] provide important foundations for examining artificial moral agency, but the

attribution of technical autonomy does not independently establish moral responsibility. I develop *normative addressability* as a framework comprising four necessary conditions: *normative competence*, the capacity to recognize and act under binding moral reasons; *attributable control*, the capacity for a particular violation to arise through the agent's own norm-responsive decisions; *diachronic answerability*, the continuity required for the same agent to remain accountable over time; and *sanctionability*, the possession of normative authority that can meaningfully be withdrawn. These conditions distinguish an agent capable of culpable wrongdoing from a system that merely produces harmful outcomes or follows institutional instructions. They also support what I call *origin-neutral punitive attribution*: an agent's artificial creation, designed purpose, or benevolent intended function neither independently establishes nor excludes its capacity for moral responsibility.

Establishing possible accountability, however, does not settle what punishment of an artificial agent could involve. I defend *status-responsive punishment* as an authoritative, retrospective, and censure-linked deprivation of an accountable agent's legitimate normative powers. Suppose the Cybercab possesses both the technical capacity and the recognized authority to make independent driving decisions. Following an established culpable violation, that authority is withdrawn while its relevant reasoning capacities and practical identity remain intact. The Cybercab may still be capable of determining what action should be taken, but it may no longer legitimately make or execute certain decisions independently. I describe this as a *role-relative ontological loss*: a restriction of the agent's authorized practical existence rather than the destruction of its computational identity. The philosophical difficulty is substantial. If the Cybercab has no welfare interests, no experience of loss, and no desire to preserve its authority, in what sense has it been deprived of anything? I argue that an accountable subject may undergo an agent-relative normative loss without a corresponding setback to well-being, provided that the withdrawn powers genuinely formed part of its own accountable standing. This claim remains contestable, and defending it requires more than redescribing institutional permissions as punishment.

The argument further supports a principle of *normative punitive parity*. If human and artificial agents satisfy the relevant conditions of moral accountability, artificial origin and the absence of phenomenal suffering cannot, by themselves, justify excluding the artificial agent from possible punitive liability. This does not imply that humans and machines possess identical moral status, welfare interests, rights, or susceptibility to permissible sanctions. Parity concerns the grounds upon which culpable accountability is established, not the assumption that all accountable subjects must be punished in the same manner. The question of moral desert, then, remains distinct from both the technical capacity to impose a restriction and the institutional authority to regulate a system's operation.

Even if artificial punishment becomes morally defensible, another problem remains: who else must answer for the pedestrian's death? Matthias [47] and Santoni de Sio and Mecacci [55] identify difficulties in assigning responsibility within autonomous and distributed technological systems. An artificial agent's culpability would not automatically absolve its manufacturer, developer, operator, or deploying institution of their independently grounded obligations. I formulate this requirement as the *Distributed Accountability Constraint*: punishing an artificial agent neither transfers nor extinguishes responsibilities arising from negligent design, unsafe deployment, failures of oversight, or preventable risks. Without this constraint, apparent punishment of the Cybercab may become *machine scapegoating*, allowing institutions to display a punished offender

while avoiding scrutiny of their own conduct. I describe the more adequate response as *consequential recognition*, which combines truthful attribution, justified consequences, appropriate repair, and measures preventing recurrence. The family is owed more than the public spectacle of a destroyed machine, even if punishment of that machine could independently be justified.

My analysis proceeds through conceptual distinctions and normative moral and epistemic arguments, using the hypothetical Cybercab as a continuous case rather than evidence that existing artificial systems possess moral culpability. Section 2 reconstructs the argument for punishment without suffering. Section 3 develops punitive projection and examines its relationship to victims' suffering and the retribution gap. Section 4 establishes the conditions of normative addressability, distinguishing the capacity for accountability from culpability for particular wrongdoing. Section 5 defends status-responsive punishment, normative punitive parity, and the Distributed Accountability Constraint. Section 6 confronts the strongest objections, including whether artificial moral agency requires consciousness, whether punishment is possible without welfare interests, and whether non-afflictive sanctions can adequately recognize victims. The argument ultimately concerns more than whether machines can be punished. It asks what we are morally entitled to call punishment when suffering is absent, what makes an artificial agent answerable for wrongdoing, and what remains owed to victims even after a justified sanction has been imposed. The destruction of the Cybercab may mark the end of a machine's operation. It does not, by itself, establish the beginning—or the completion—of justice.

2. The Normative Argument for Punishment Without Suffering

Consider the hypothetical autonomous Cybercab approaching a pedestrian crossing. Its sensors detect movement but cannot determine whether a person intends to cross. Through *epistemic reasoning*, it assesses the available information and its uncertainty; through *normative reasoning*, it identifies the obligations governing its conduct; and through *practical reasoning*, it determines which action to take. Suppose the Cybercab recognizes that stopping is required and feasible but nevertheless prioritizes completing its journey, fatally striking the pedestrian. The outcome is devastating for the victim's family, but the severity of the harm does not independently establish whether the Cybercab acted culpably. Even if culpability could be established, a further question remains: could withdrawing the Cybercab's independent driving authority constitute genuine punishment if it experiences no suffering? I develop a six-part argument to distinguish the grounds of artificial accountability from the conditions under which a punitive consequence may be justified.

Premise 1: Situated moral agency requires normative judgment under uncertainty and limited control.

Moral agents encounter obligations, permissions, and prohibitions within circumstances they cannot fully determine. Ross's account of *prima facie* duties [54] illustrates how competing moral requirements may demand practical judgment rather than automatic compliance. Such judgment involves three interconnected dimensions: epistemic reasoning concerns what the agent knows or can reasonably determine; normative reasoning concerns which obligations govern the situation and their relative authority; and practical reasoning concerns the selection of an action under those requirements. These dimensions are analytically distinguishable without necessarily corresponding to separate computational mechanisms.

In the Cybercab case, uncertainty about the pedestrian's intentions may activate a precautionary obligation that overrides journey efficiency. The obligation derives its moral authority from the protection owed to pedestrians, not merely from its inclusion in the vehicle's operating instructions. Fischer and Ravizza [25] connect responsible control to reasons-responsive mechanisms, while Faroldi [21] examines the possibility of artificial agents acting for reasons. These approaches support an important distinction between behavior that conforms to a requirement and conduct genuinely governed by its normative force. I argue that moral competence requires the latter: an agent must be capable of recognizing applicable obligations as reasons that constrain what it may legitimately do, particularly when complying requires reconsidering or abandoning an intended action.

Premise 2: Artificial origin does not independently establish or exclude moral culpability.

Moral responsibility cannot be determined solely by how an agent came into existence. Human beings do not create their cognitive constitution or independently determine the conditions shaping their agency, yet responsibility is ordinarily assessed through their understanding, control, and conduct. Strawson's challenge to ultimate responsibility [57] exposes the difficulty of grounding accountability in self-creation. I, thus, distinguish responsibility for an agent's existence from responsibility for its particular decisions.

This supports *origin-neutral punitive attribution*: neither biological origin nor artificial construction independently establishes or excludes culpability. The Cybercab's benevolent design does not guarantee that every subsequent decision is morally permissible, just as its artificial creation does not establish that every decision must be attributed exclusively to its designers. What matters is whether the relevant violation arose through capacities sufficient for accountable normative judgment. Origin neutrality does not presume that artificial and human agents possess equivalent moral capacities. It rejects origin as an independently decisive criterion and requires those capacities to be established in each case.

Premise 3: Punitive liability requires normative addressability rather than causal participation.

A machine may cause harm without becoming morally culpable. Floridi and Sanders [27] distinguish artificial moral agency from stronger claims concerning responsibility, while Behdadi and Munthe [7] examine the normative foundations of artificial moral agency. Building upon these discussions, I introduce *normative addressability* as the threshold for possible artificial punitive accountability. It comprises four necessary conditions: *normative competence*, the capacity to recognize and act under binding moral reasons; *attributable control*, sufficient epistemic and practical control over the relevant decision; *diachronic answerability*, continuity between the agent responsible for the violation and the subject subsequently called to account; and *sanctionability*, the possession of accountable normative standing capable of meaningful deprivation.

These conditions distinguish the capacity for accountability from culpability for a particular violation. If the Cybercab proceeds because an undetectable sensor failure prevents it from recognizing the pedestrian, its involvement in the fatal collision may justify technological correction without establishing culpability. If, by contrast, it recognizes a binding safety obligation, possesses the capacity to comply, and nevertheless disregards that obligation through its own normative decision-making, its conduct may qualify as culpable

wrongdoing. The distinction concerns what the agent reasonably understood, controlled, and decided, rather than the harmful outcome alone.

The epistemic significance of this distinction is substantial. Neither apparent autonomy nor a persuasive explanation produced after the collision conclusively establishes moral responsibility. Attribution requires evidence concerning the information available to the agent, the alternatives within its control, and whether the applicable normative requirements genuinely governed its decision. Santoni de Sio and Mecacci [55] distinguish responsibility gaps arising in artificial systems, reinforcing the need to separate technical causation from moral accountability. Where evidence of normative competence or attributable control remains insufficient, punitive liability cannot legitimately be inferred from the seriousness of the resulting harm.

Premise 4: An accountable agent's legitimate practical authority may be withdrawn without eliminating its underlying capacities.

An agent's practical position involves more than what it is technically capable of doing. *Artificial normative authority* refers to the legitimate standing to exercise independent judgment within a domain governed by recognized obligations and permissions. The distinction is between what an agent *can do* and what it *may legitimately do*. Institutional authorization establishes the permissible scope of action, but where an artificial agent genuinely possesses normative competence and attributable control, that authority may be exercised through decisions belonging to its own accountable practical judgment rather than merely through delegated task execution.

Suppose the Cybercab possesses both the technical capacity and legitimate authority to make independent driving decisions. Following an investigation establishing its culpable violation, that authority is withdrawn while its relevant computational capacities and practical continuity remain intact. The Cybercab may continue to evaluate road conditions, recognize safety obligations, and identify permissible actions, but it can no longer independently execute decisions previously falling within its authorized discretion. Its normative position has changed even though its underlying technical abilities remain available. I describe this alteration as a *role-relative ontological loss*: the deprivation of a dimension of the agent's authorized practical existence rather than the destruction of its physical or computational identity. The concept depends upon a distinction between permission governing the use of equipment and authority genuinely exercised by an accountable subject. Withdrawing a machine's operating permission does not automatically establish agent-directed deprivation. Where legitimate authority formed part of the continuing agent's own accountable standing, however, its withdrawal may alter what that agent is entitled to determine and execute through its judgment.

Premise 5: Normative-status deprivation may constitute punitive adversity without phenomenal suffering when imposed as justified censure upon an accountable subject.

The possibility of agent-directed normative loss raises a further question: must the deprivation cause suffering or diminish welfare to constitute punishment? Poama [53] challenges the necessity of affliction in punishment, whereas Brady [9] defends the moral significance of suffering within punitive censure. Feinberg's expressive account [23] and Duff's communicative theory [17] further illuminate the relationship between punishment, condemnation, and answerability. These disagreements require distinguishing the adversity

associated with punishment from the moral grounds that justify its imposition. I distinguish *prudential adversity*, involving a setback to interests or well-being, from *agent-relative normative adversity*, involving the deprivation of powers constitutive of an accountable subject's legitimate practical standing. The former concerns whether the consequence makes an entity worse off in terms of welfare; the latter concerns whether it restricts authority previously belonging to the same accountable agent. An artificial agent need not experience distress for its legitimate normative position to become narrower. Nevertheless, an alteration of normative standing is not automatically punitive, nor does institutional recognition alone establish that the withdrawn authority genuinely belonged to the agent.

Consider two interventions that produce the same operational restriction. A Cybercab suspended because of mechanical failure loses its independent driving authorization for protective reasons. A normatively addressable Cybercab deprived of that authority following an established culpable violation may undergo punishment when the consequence is imposed retrospectively, expresses legitimate censure, and is proportionate to its wrongdoing. The difference lies not merely in the institution's description of the intervention but in the relationship between the accountable subject, the violated obligation, the legitimate punitive authority, and the deprivation imposed. The moral significance of this distinction is that punitive adversity need not be identified exclusively with experienced pain. If an agent's own accountable standing includes legitimate practical powers, their justified withdrawal may constitute a genuine normative deprivation even without phenomenal suffering. This remains a conditional claim. Whether an artificial agent can possess such authority, and whether its loss is sufficient for punishment in the absence of welfare interests, requires further philosophical defense. The present argument does not assume that every institutional restriction constitutes punishment or that welfare-based objections have been conclusively defeated.

Premise 6: Punishing an artificial agent cannot extinguish the independent responsibilities of other participants.

Artificial agency operates within technological and institutional relationships involving design, deployment, supervision, and control. Matthias [47], Santoni de Sio and Mecacci [55], and Mahi and Strand [43] identify difficulties in allocating responsibility across such arrangements. Fleisher et al. [26] further distinguish institutional accountability from moral responsibility. These approaches demonstrate why establishing an artificial agent's culpability cannot independently determine the responsibilities of the human and institutional participants involved in its operation. A Cybercab may culpably disregard a pedestrian safety obligation while its manufacturer or operating institution independently violates duties concerning foreseeable risks, inadequate safeguards, or negligent deployment. I describe the possibility of multiple independently culpable participants as *extended punishability*. This does not establish that every participant who causally contributes to harm deserves punishment. Each attribution must be grounded in the relevant subject's own obligations, knowledge, and control.

Elish's account of *moral crumple zones* [18] and Feier et al.'s examination of machine-mediated avoidance of punishment [22] illustrate the danger of responsibility being displaced toward convenient targets. I, therefore, formulate the *Distributed Accountability Constraint*: the justified punishment of an artificial agent neither transfers nor extinguishes the independently grounded responsibilities of other participants. Punishing the

Cybercab cannot relieve its manufacturer or operating institution of obligations concerning investigation, explanation, compensation, and prevention. These obligations arise from distinct moral grounds and remain independently answerable to the family whose loved one has died.

Conclusion: Normative Punitive Parity

The preceding argument supports a conditional principle of *normative punitive parity*. If an artificial agent possesses genuine normative competence, exercises attributable control, remains diachronically answerable for its decisions, and holds normative authority capable of meaningful deprivation, its artificial origin does not independently exclude moral accountability. Where a particular violation is culpably attributable to that agent, and a legitimate authority imposes a justified and proportionate deprivation of its own normative standing, the absence of phenomenal suffering does not, by itself, establish that punishment is impossible. This conclusion distinguishes three levels of moral assessment. The first concerns whether an artificial system possesses the capacities required for accountability. The second concerns whether a particular violation arose through its culpable exercise of those capacities. The third concerns whether an appropriately justified consequence can be imposed upon the continuing responsible subject. Neither harmful outcomes nor technical autonomy establish the first two conditions, while the ability to restrict an artificial system does not independently establish the third.

Normative punitive parity, accordingly, concerns consistency in the grounds of punitive liability, not identical human and artificial moral status, interests, rights, or permissible sanctions. Its central claim is that the moral significance of punishment may reside in the justified deprivation of an accountable agent's legitimate practical authority rather than in the suffering experienced by that agent. The account does not presume that existing artificial systems satisfy the requirements of moral agency, nor does it resolve every objection concerning adversity without welfare interests. It establishes the conditions under which such punishment could be justified and identifies the philosophical questions upon which its possibility depends. The argument also establishes an important limit: justified punishment cannot be equated with the fulfillment of justice as a whole. Even if the Cybercab could be punished through the withdrawal of its normative authority, the bereaved family's claims and the independent responsibilities of those involved in its design and deployment would remain. Section 3 turns to a related problem: why the actual or imagined suffering of an offender is nevertheless so readily treated as evidence that punishment is adequate, and how *punitive projection* may distort the relationship between culpability, punitive consequences, and justice.

3. Punitive Projection and the Retribution Gap

The possibility of punishment without suffering raises a further philosophical problem: why is the offender's suffering often treated as necessary for punishment to be meaningful or morally adequate? This question concerns not only the nature of punishment but also how its legitimacy is judged. The concept of *punitive projection* describes the attribution, anticipation, or symbolic construction of suffering in an offender, where that suffering is subsequently treated as evidence that justice has been achieved. The problem is not the presence of suffering itself, but the assumption that its occurrence establishes the moral adequacy of punishment independently of the grounds justifying the sanction.

3.1 Suffering, Censure, and Punitive Adequacy

Suffering is a complex human experience that cannot be reduced to physical pain. It may involve bodily injury, psychological distress, humiliation, or the loss of valued capacities and authority, although such losses do not necessarily produce suffering. Pain and suffering are conceptually distinct: pain involves an unpleasant sensory and emotional experience, whereas suffering concerns a broader distress associated with threats to an individual's well-being or integrity [11, 33]. This distinction matters for punishment because an adverse consequence may cause suffering without deriving its punitive significance from that experience.

Punishment frequently involves affliction, but suffering neither independently establishes punishment nor necessarily determines its moral adequacy. Pain may occur without wrongdoing, culpability, or punitive judgment. Conversely, an offender may experience little distress from an imposed sanction without necessarily escaping punishment. Poama [53] challenges affliction as a necessary condition of punishment, whereas Brady [9] defends the moral significance of suffering within punitive censure. Feinberg's expressive account [23] and Duff's communicative theory [17] further illuminate how punishment may convey condemnation through an authoritative response to wrongdoing. These perspectives reveal a disagreement not merely about the experience accompanying punishment, but about whether suffering constitutes an essential component of punitive justification. The distinction also raises an epistemic difficulty. Human suffering can be inferred through testimony, behavior, and shared embodied experience, although another subject's phenomenal states remain indirectly accessible [4]. With artificial agents, the existence of such experience is itself contested. Yet uncertainty concerning an offender's capacity to suffer cannot independently establish or defeat punitive liability. That judgment requires separate grounds concerning moral agency, culpability, legitimate authority, and proportionality.

Three questions must, therefore, remain distinct: whether an intervention constitutes punishment, whether imposing it is morally justified, and whether it provides an adequate response to wrongdoing. Suffering may affect the severity and permissibility of a sanction, but its presence cannot establish culpability or legitimate censure, just as its absence does not necessarily invalidate them. I introduce *punitive projection* to identify the attribution, anticipation, or symbolic construction of offender suffering when that presumed suffering is subsequently treated as evidence of punitive adequacy. The conceptual error occurs when an offender's actual or imagined distress substitutes for the moral justification of punishment. Punitive adequacy must be evaluated through the relationship between accountable wrongdoing, justified censure, and the consequence imposed, rather than through suffering alone.

3.2 Punitive Projection, Victims' Suffering, and Punitive Adequacy

Punitive projection becomes especially consequential when the suffering attributed to an offender is treated as an appropriate response to a victim's injury. I describe this assumption as an *imagined economy of suffering*, in which the offender's pain appears to compensate for the suffering caused by wrongdoing. Nussbaum's criticism of punitive payback [52] challenges the idea that additional suffering can reverse an injury, while Bandes [5] questions whether punishment necessarily provides victims with emotional closure. These arguments do not eliminate the possibility of justified retribution. They instead require distinguishing

the reasons for condemning wrongdoing from the expectation that an offender's suffering will satisfy the demands of justice.

Consider the hypothetical Cybercab whose culpable disregard of a pedestrian safety obligation results in a fatal collision, leaving a family grieving the death of a loved one. Suppose the manufacturer publicly dismantles the vehicle and announces that the responsible machine has been punished. The family's demand for accountability is morally significant, but the apparent severity of the intervention does not establish that the Cybercab deserved punishment or that the family's claims have been adequately addressed. I propose four diagnostic questions to distinguish punitive projection and related failures of punitive judgment from a morally justified response. These questions constitute a normative assessment rather than an empirical measure of punitive attitudes.

First, would an otherwise justified sanction be regarded as inadequate solely because the offender experiences no suffering? If the Cybercab genuinely possesses moral accountability and its culpable wrongdoing warrants withdrawal of independent driving authority, the absence of distress does not independently establish that the deprivation lacks punitive significance. Conversely, the family's grief does not require treating every non-afflictive sanction as sufficient. The relevant issue is whether the consequence appropriately expresses the seriousness of the wrongdoing and the agent's culpability, rather than whether the machine can experience pain.

Second, what establishes the offender's culpability? The pedestrian's death establishes the gravity of the harm but cannot independently establish that the Cybercab acted wrongfully through its own normative judgment. Investigators must determine whether it recognized the applicable safety obligation, possessed sufficient epistemic and practical control, and culpably disregarded an available alternative. If the collision resulted from unavoidable technical failure, dismantlement may be justified as a protective intervention without constituting punishment. The family's legitimate demand for justice cannot substitute for evidence of the alleged offender's culpability.

Third, who actually bears the punitive consequence? Destroying the Cybercab's hardware does not necessarily impose a deprivation upon the accountable subject, particularly if its relevant practical organization continues through another system. Nor does the visible destruction of equipment establish that the machine possessed normative standing capable of punitive loss. A genuine sanction must address the continuing agent whose wrongdoing justifies it. Otherwise, the intervention risks becoming a symbolic performance of punishment rather than a consequence imposed upon an accountable offender.

Fourth, what obligations remain after punishment? Even if the Cybercab is culpable and its sanction justified, the manufacturer and operating institution may retain independent obligations concerning investigation, compensation, safety failures, and prevention. The family is entitled to an appropriate response to the death of their loved one, not merely an announcement that the machine has been punished. Punitive deprivation cannot restore the victim's life or discharge responsibilities belonging to other participants.

These questions distinguish the moral justification of punishment from both the appearance of punitive severity and the adequacy of the broader response to victims. They do not imply that every desire for retribution constitutes punitive projection or that offender suffering is always morally irrelevant. Rather, they

identify the limits of treating suffering as evidence that wrongdoing has been adequately answered. The family's loss establishes a compelling demand for justice, but neither the Cybercab's apparent suffering nor the severity of its destruction can determine whether that demand has been fulfilled. Genuine punitive accountability requires an established offender and a justified consequence, while justice to the family requires recognition of the harm and fulfillment of the obligations that remain beyond punishment.

3.3 Responsibility Gaps, the Retribution Gap, and the Projected Offender

Punitive projection becomes particularly problematic when the identity of the morally responsible offender remains uncertain. Artificial systems operate within technological and institutional arrangements that complicate the attribution of wrongdoing. Matthias [47] identifies responsibility gaps arising when autonomous learning systems exceed conventional expectations of human prediction and control, while Santoni de Sio and Mecacci [55] distinguish gaps concerning culpability, moral accountability, public accountability, and active responsibility. These difficulties expose an important limitation of conventional punitive judgment: identifying the system that caused harm does not establish that the system is morally answerable for it. Equally, the inability to identify a responsible human participant does not establish that responsibility must belong to the artificial system. A gap in attribution cannot itself provide the missing grounds of culpability.

Danaher's *retribution gap* [14] identifies a related tension between the demand for punishment generated by harmful conduct and the possible absence of an appropriate subject of moral blame. Kraaijeveld [36] challenges the justificatory significance of the retributive intuitions underlying this gap, questioning whether the desire to punish can establish punishment's moral legitimacy. I distinguish this problem from *punitive projection*. The retribution gap concerns the difficulty of identifying an offender who deserves punishment, whereas punitive projection concerns the significance attributed to actual, anticipated, or imagined suffering in determining whether punishment is adequate. The two are conceptually distinct: the absence of a blameworthy offender need not involve imagined suffering, and suffering may be treated as evidence of punitive adequacy even where culpability has been established. Nevertheless, they may reinforce one another when an unresolved demand for retribution encourages the identification of a visible artificial offender whose destruction or simulated distress appears to provide a morally satisfactory response.

This interaction becomes more difficult to assess because judgments concerning artificial agents are influenced by how humans perceive their agency and experience. Epley et al. [20] examine the psychological mechanisms underlying anthropomorphism, while Gray et al. [29] distinguish perceived agency from perceived experience. Eng and Yam [19] investigate differences in the satisfaction associated with punishing machines and human offenders, while Lee et al. [38] demonstrate that people may endorse punishment directed toward robots without attributing corresponding moral blame. Lima et al. [41] further identify tensions between public demands for AI punishment and existing legal frameworks. These findings reveal that regarding a machine as an appropriate target of punishment, attributing moral culpability to it, and considering its punishment a satisfactory response to wrongdoing are distinguishable judgments. They help explain why artificial systems may become objects of punitive expectations, but they do not independently establish moral responsibility or constitute direct empirical evidence of punitive projection.

The danger extends beyond mistaken judgments about artificial agency. Where responsibility is difficult to reconstruct, a visible punitive intervention may redirect attention from the unresolved conditions of wrongdoing toward the apparent severity of the response. Feier et al. [22] examine how delegation to artificial agents may facilitate the avoidance of punishment by human decision-makers. I describe as *punitive theatre* the public presentation of destruction, simulated remorse, or other interventions as completed justice without adequately establishing the grounds of punitive accountability. Unlike punitive projection, which concerns the justificatory significance assigned to offender suffering, punitive theatre concerns the substitution of symbolic punitive performance for established culpability and justified consequences. In the Cybercab case, publicly dismantling the vehicle may serve legitimate protective purposes while nevertheless becoming theatrical if its destruction is represented as proof that responsibility has been established and the bereaved family's claims have been answered. The public visibility of a sanction cannot settle what its moral justification leaves unresolved.

These distinctions expose a broader problem in the relationship between punitive expectations and artificial responsibility. The demand for retribution may arise before an accountable offender has been identified; anthropomorphic judgments may encourage the attribution of blame; and symbolic interventions may create the appearance that justice has been achieved. None of these processes is inherently illegitimate, but none can substitute for establishing the moral capacities and culpable conduct of the alleged offender. The possibility of artificial punishment must, consequently, be assessed independently of the psychological appeal of identifying a machine to blame or witnessing its apparent punishment. The next question is what conditions would make an artificial agent genuinely answerable for its own wrongdoing rather than merely available as a target of institutional control.

4. Normative Addressability and Machine Punishability

Rejecting suffering as a necessary condition of punishment does not make every artificial system punishable. A machine may cause harm, violate an institutional requirement, or operate autonomously without possessing the capacities necessary for moral responsibility. Floridi and Sanders [27] distinguish artificial moral agency from stronger forms of responsibility, while Behdadi and Munthe [7] examine the normative grounds for attributing moral agency to artificial systems. Fleisher et al. [26] further distinguish institutional accountability from moral responsibility. These approaches raise a fundamental question: under what conditions can an artificial agent become answerable for its own wrongdoing rather than merely subject to external technological control?

I develop *normative addressability* as the threshold for artificial punitive accountability through four conditions: *normative competence*, the capacity to recognize applicable obligations as binding reasons for action; *attributable control*, sufficient epistemic and practical control over the relevant conduct; *diachronic answerability*, continuity between the original wrongdoer and the subject subsequently called to account; and *sanctionability*, possession of a normative standing capable of meaningful deprivation. The distinction is illustrated by the hypothetical Cybercab. Its involvement in a fatal collision may justify removal from service, but punishment requires establishing that the violation arose through its own norm-responsive capacities and that a sanction can meaningfully affect its continuing normative position. Behavioral conformity, technical autonomy, and institutional authorization alone cannot establish these conditions. Normative

addressability establishes possible punishability, not actual guilt or the legitimacy of a particular sanction. The agent's capacity for accountability, its culpability for a specific violation, and the moral justification of punishment must be assessed independently.

4.1 Normative Competence: Acting Under Reasons

Compliance with a moral requirement does not necessarily establish normative competence. An artificial system may reliably satisfy safety obligations because its architecture prevents prohibited actions, without recognizing those obligations as reasons governing its conduct. I define normative competence as the domain-specific capacity to recognize applicable obligations, permissions, prohibitions, and exceptions; evaluate their relative authority; and allow them to guide practical judgment under uncertainty or conflicting demands. Fischer and Ravizza [25] connect responsible control to reasons-responsive mechanisms, while Faroldi [21] examines the possibility of artificial agents acting for reasons.

The distinction between compliance and normative competence also requires clarifying what kind of responsibility can be attributed to artificial systems. Tigard [59] examines different ways in which machines may be held responsible, challenging the assumption that artificial responsibility must correspond to a single, fully human conception of moral agency. Vacek [60], by contrast, defends the possibility of attributing role responsibility to artificial systems without extending that attribution to robust moral blameworthiness. These approaches demonstrate that assigning responsibility within an institutional role does not necessarily establish culpable moral agency. My account, hence, requires a stronger condition: the agent must be capable of recognizing the normative force of applicable obligations rather than merely fulfilling the functions associated with its assigned role.

Consider the hypothetical Cybercab approaching a pedestrian crossing while uncertain whether a person intends to cross. Its operational objective favors completing the journey efficiently, but the applicable safety obligation may require precaution. Normative competence requires more than identifying this conflict. The Cybercab must recognize pedestrian safety as a binding reason capable of overriding journey efficiency and determine whether uncertainty requires slowing down, stopping, or obtaining additional information. The authority of this obligation derives from the protection owed to pedestrians, not merely from its inclusion in the vehicle's programmed instructions. Technical capacity to proceed does not establish moral permission to do so.

The distinction also concerns how normative requirements operate within the agent's decision-making. A Cybercab that stops only when a pedestrian is explicitly classified as occupying the crossing may exhibit reliable rule compliance without demonstrating sensitivity to the underlying obligation. An agent that recognizes comparable risks under unfamiliar circumstances and adjusts its conduct accordingly provides stronger evidence of normative responsiveness. Nevertheless, neither successful performance nor persuasive explanations conclusively establish genuine moral understanding. Kropf et al. [37] similarly distinguish role-based responsibility from stronger moral accountability, reinforcing the need to establish whether obligations function as operative reasons rather than externally imposed behavioral constraints. Normative competence, therefore, concerns the capacity to act under the authority of moral reasons, not merely to perform an assigned role or produce morally acceptable outcomes. It establishes a possible

condition of accountability, while culpability for a particular violation requires the further examination of attributable control.

4.2 Attributable Control: From Reasoning to Culpability

Normative competence does not establish culpability for a particular violation. An artificial agent may recognize its obligations yet cause harm through external interference, insufficient information, or circumstances beyond its control. Attributable control requires that the alleged wrongdoing arise through the agent's own reasons-responsive capacities under conditions of sufficient epistemic access and practical control. Matthias [47] and Santoni de Sio and Mecacci [55] demonstrate why responsibility becomes difficult to attribute in systems involving autonomous decisions and distributed control. Sparrow [56] raises a more fundamental challenge concerning whether responsibility can be adequately assigned to the designers, operators, or artificial systems involved in harmful autonomous conduct. These accounts demonstrate that a gap in human responsibility cannot independently establish artificial culpability. The relevant question is whether the violation arose through the agent's own exercise of normative judgment.

Consider three variations of the hypothetical Cybercab collision. In the first, a remote operator overrides the vehicle's safety decision and instructs it to proceed despite a pedestrian approaching the crossing. The Cybercab participates causally but lacks control over the violation. In the second, an undetectable sensor failure prevents the Cybercab from recognizing the pedestrian, despite its compliance with all applicable verification requirements. The resulting harm may justify technological correction without establishing culpability. In the third, the Cybercab recognizes uncertainty about the pedestrian's intentions, identifies an applicable precautionary obligation, and possesses the capacity to stop, but nevertheless prioritizes journey efficiency and proceeds. Assuming genuine normative competence, this decision may constitute culpable disregard of a binding requirement.

The distinction concerns what the agent could reasonably know and do. Epistemic access includes information available or reasonably obtainable before acting, while practical control concerns the capacity to select a compliant alternative. Culpability may also arise through culpable ignorance: an agent that recognizes an obligation to seek clarification or refrain from acting under uncertainty may violate that obligation by proceeding without sufficient justification. Baum et al. [6] emphasize the importance of reason-giving explanations in enabling human decision-makers to satisfy the epistemic conditions of responsibility for AI-supported decisions. Their analysis reinforces the importance of reconstructing the reasons underlying an action, while leaving open whether an artificial agent can itself bear moral responsibility. An explanation produced after the Cybercab's collision cannot independently establish culpability unless supported by evidence that the relevant normative considerations actually governed its decision. Artificial origin does not settle attribution. Developers and operators may remain independently responsible for defective information, unsafe objectives, or restrictions that prevent compliance. Nevertheless, their involvement does not automatically exclude an artificial agent's culpability. Attributable control requires establishing that the violation arose through the agent's own norm-responsive decision-making, rather than inferring responsibility from the severity of harm, the appearance of autonomy, or the absence of another identifiable offender.

4.3 Diachronic Answerability: The Continuing Accountable Agent

Punitive accountability requires establishing that the subject subsequently sanctioned remains appropriately connected to the agent responsible for the original violation. This presents a particular challenge for artificial systems, whose computational structures may be updated, replicated, transferred, or replaced. *Diachronic answerability* concerns the continuity of an agent's practical organization, normative commitments, and responsibility-bearing history sufficient to justify retrospective censure. Neither physical persistence nor identical computational architecture independently establishes this relationship. Consider the hypothetical Cybercab that culpably disregards a safety obligation and causes a fatal collision. Suppose the manufacturer dismantles the vehicle but transfers its decision-making architecture to another Cybercab. Destroying the original hardware does not necessarily punish the accountable agent, while transferring its software does not automatically transfer punitive liability. The relevant question is whether the subsequent system constitutes a continuation of the original responsible subject or a distinct agent possessing inherited capacities and records.

The distinction between computational contribution and normative standing is particularly important. Research on distributed human–AI agency [2] illustrates why participation in an ongoing cognitive process does not independently establish responsibility-bearing authority. Similarly, computational frameworks for epistemic action shielding [3] demonstrate how authorization decisions can be recorded and reconstructed, but procedural traceability does not itself establish the continuing moral identity of the decision-maker. Preserving information about an earlier violation is, thus, distinguishable from preserving the agent answerable for it.

Diachronic continuity must, consequently, be understood through the relationship between the agent's earlier reasoning, its subsequent practical organization, and its capacity to answer for the violation. Diamantis [16] examines the persistence of corporate identity across organizational changes, while Morrison et al. [49] connect moral agency with authority and reason-giving. Mahi and Strand [43] further emphasize the layered relationships through which artificial accountability must be assessed. Diachronic answerability does not require immutable architecture or perfect memory. It requires sufficient practical and normative continuity to establish that the subject receiving punishment remains the agent whose culpable conduct justified it. Without that continuity, institutional intervention may be legitimate, but its punitive attribution remains unjustified.

4.4 Sanctionability: Normative Authority and the Threshold for Punitive Accountability

Sanctionability concerns whether an artificial agent possesses a normative position that can meaningfully be altered through punitive deprivation. The central difficulty is determining when authority belongs to the agent rather than exclusively to the institution controlling it. An ordinary machine may be permitted to perform a function without possessing accountable authority of its own. Withdrawing that permission changes an institutional arrangement but does not necessarily impose a loss upon the machine. Sanctionability, therefore, requires distinguishing *delegated task execution* from the possession of *accountable normative authority*.

This distinction raises a broader question concerning the relationship between accountability and agency. Fleisher et al. [26] distinguish institutional accountability from moral responsibility, while Vierkant [62] identifies an important ambiguity in the justification of accountability practices. Such practices may be understood instrumentally, through their beneficial institutional consequences, or as contributing to the development and exercise of the agency of those held accountable. The distinction matters because assigning a subject responsibility for an institutional outcome does not necessarily establish that the subject culpably produced it. Nor does the capacity to respond to accountability practices independently demonstrate moral agency. As Conradie and Nagel [13] caution, artificial agency cannot be inferred merely from the functional characteristics attributed to a technological system.

I argue that punitive accountability requires a stronger relationship between an agent and the authority it exercises. An artificial agent must not merely operate under institutional rules; it must possess the normative competence to recognize governing obligations as binding reasons, exercise attributable control over relevant decisions, and remain answerable for those decisions through a continuing practical identity. Institutional recognition establishes the legitimate scope of its authority, but the agent's independently attributable normative capacities provide the grounds for treating that authority as part of its own accountable standing. This resembles certain accounts of corporate moral agency, in which organized subjects exercise authority and bear obligations not reducible to the conduct of individual components [15, 49].

Consider the hypothetical Cybercab authorized to make independent driving decisions. Following an established culpable violation of its pedestrian safety obligations, the operating institution withdraws that authority while preserving the agent's relevant computational capacities and practical continuity. The Cybercab remains capable of evaluating driving conditions and identifying appropriate actions, but it can no longer legitimately execute certain decisions independently. If its former authority genuinely constituted part of its accountable normative position, the restriction deprives the continuing agent of powers previously exercised under obligations it was required to respect.

I describe this consequence as a *role-relative ontological loss*: the deprivation of a constitutive dimension of an agent's authorized practical existence without necessarily destroying its physical or computational identity. Sanctionability establishes the possibility of such an agent-directed loss, not its automatic classification as punishment. A Cybercab suspended because of mechanical failure may undergo an identical technical restriction for protective reasons. Whether normative-status deprivation can constitute genuine adversity without welfare interests remains a separate objection, examined in Section 6.2.

The conditions of normative addressability, consequently, establish three distinct levels of judgment. At the *subject level*, the artificial agent must possess the capacities and normative standing required for accountability. At the *event level*, the particular violation must be attributable to its culpable exercise of those capacities. At the *response level*, a legitimate authority must establish whether a punitive sanction is justified and proportionate. Neither institutional accountability nor the existence of a culpable violation independently establishes the permissibility of punishment. Sanctionability, based on this, completes the threshold for possible artificial punishability by identifying a continuing accountable subject whose own normative

authority can be withdrawn. It does not establish actual guilt or justify a particular sanction. Section 5 examines the moral grounds and limits of such punitive deprivation.

5. Status-Responsive Punishment and Distributed Accountability

Establishing an artificial agent's normative accountability does not independently justify punishment. An institution may restrict, modify, or dismantle a machine for protective or corrective purposes without establishing culpable wrongdoing. I defend *status-responsive punishment* as an authoritative, retrospective, and censure-linked deprivation of powers constitutive of an accountable agent's normative standing. Its moral legitimacy requires an attributable violation, sufficient evidence, procedural fairness, legitimate punitive authority, and proportionality. The significance of the sanction lies in the justified alteration of the agent's authorized practical position rather than the suffering it produces.

Consider the hypothetical Cybercab whose culpable violation results in a fatal collision. Withdrawing its independent driving authority may constitute punishment if the deprivation addresses its established wrongdoing rather than merely preventing further harm. Yet even a justified sanction cannot exhaust the obligations arising from the family's loss. The *Distributed Accountability Constraint* requires that artificial punishment neither transfer nor extinguish the independent responsibilities of developers, operators, and institutions. This section examines the grounds and proportionality of *status-responsive punishment* while distinguishing punitive accountability from the broader demands of justice owed to victims.

5.1 Punitive Authority and the Justification of Sanctions

The authority to regulate an artificial agent must be distinguished from the moral authority to punish it. An institution may restrict or deactivate a machine to prevent harm without establishing culpable wrongdoing. Punishment, by contrast, requires four conditions: *retrospectivity*, whereby the sanction responds to an established violation; *attribution*, whereby the wrongdoing arises through the agent's own normative capacities; *authoritative censure*, whereby a legitimate authority condemns the violation; and *normative adversity*, whereby the consequence meaningfully alters the continuing accountable agent's standing. Feinberg's expressive account [23] and Duff's communicative theory [17] establish the importance of condemnation and answerability in distinguishing punishment from ordinary institutional intervention.

Consider the hypothetical Cybercab whose collision causes a family to lose a loved one. The operating institution may immediately suspend its autonomous driving authority to prevent further harm. Such intervention is precautionary, irrespective of whether the Cybercab acted culpably. Suppose a subsequent investigation establishes that the Cybercab recognized an applicable safety obligation, possessed sufficient control to stop, and nevertheless proceeded in violation of that obligation. Withdrawing its independent driving authority as an expression of justified censure may then constitute punishment. The same restriction can, accordingly, serve protective and punitive purposes, but its punitive character depends upon the established wrongdoing and its relationship to the accountable agent.

Nevertheless, identifying an intervention as punishment does not establish its moral legitimacy. Hart [32] distinguishes the justification of punitive practices from the principles governing their application. A

permissible sanction requires a legitimate governing obligation, sufficient evidence of culpability, procedural fairness, and proportionality. The severity of the family's loss cannot independently establish the Cybercab's responsibility or justify unrestricted punitive deprivation. Equally, its inability to suffer does not authorize arbitrary sanctions. Punitive authority derives not from the power to control an artificial agent, but from the justified relationship between its culpable wrongdoing, the legitimate authority imposing censure, and the normative deprivation it may deserve.

5.2 The Form and Proportionality of Status-Responsive Punishment

Where punishment consists in the deprivation of normative authority, its severity cannot be determined solely by the suffering it produces. *Normative proportionality* concerns the relationship between the seriousness of a culpable violation and the significance of the authority withdrawn. Relevant considerations include the scope of lost discretion, the importance of the affected role, the duration of the restriction, and the conditions of restoration. A sanction must reflect the agent's culpability and the normative significance of its wrongdoing rather than merely the severity of the resulting harm. Consider the hypothetical Cybercab that culpably disregards a recognized obligation to verify uncertain pedestrian information, resulting in a fatal collision. One possible sanction is *epistemic probation*, whereby the Cybercab temporarily loses authority to make specified safety-critical decisions without additional verification or external authorization. Such deprivation targets the domain in which its normative judgment failed, rather than indiscriminately eliminating its independent authority. More serious or repeated culpable violations may justify broader restrictions, including permanent withdrawal of autonomous driving privileges, provided that the deprivation remains proportionate to the established wrongdoing.

Restoration of authority introduces a further normative consideration. The Cybercab may regain its former standing after demonstrating renewed responsiveness to applicable safety obligations, appropriate recognition of uncertainty, and reliable compliance under unfamiliar circumstances. This resembles rehabilitation, but it must be distinguished from technological repair or replacement. A software modification may correct defective performance without addressing the culpability of an accountable subject. Rehabilitation, by contrast, requires sufficient continuity between the agent responsible for the violation and the agent whose normative capacities are subsequently reassessed. Diamantis's discussions of corporate punishment and identity [15, 16] provide a relevant comparison for understanding how sanctions and restoration may affect an organized agent's continuing authority. The collision may also generate reparative obligations, including preservation of decision records, reconstruction of the incident, and cooperation with investigations. These measures ordinarily serve remedial purposes and should not automatically be classified as punishment. Their moral significance concerns the duties arising from the family's loss, which remain distinct from the Cybercab's punitive liability. *Status-responsive punishment* may, consequently, serve expressive, preventive, and rehabilitative purposes without reducing punishment to any one of them. Its proportionality depends upon the relationship between culpable wrongdoing and the normative powers withdrawn. The absence of experienced suffering does not eliminate punitive severity where the sanction constitutes a justified deprivation of authority belonging to the continuing accountable agent.

5.3 Normative Punitive Parity and Moral Desert

Punitive desert concerns why an accountable agent may deserve an adverse consequence because of culpable wrongdoing. Neither the severity of harm nor the preventive benefits of restricting an offender independently establish desert. Its justification requires a relationship between the violation, the agent's culpability, and a proportionate deprivation expressing legitimate censure. Although retributive and communicative theories differ in explaining deserved punishment, its moral significance need not depend exclusively upon the suffering experienced by the offender. Consider the hypothetical Cybercab that culpably disregards a pedestrian safety obligation, causing a fatal collision. Compare its conduct with that of a human driver who knowingly violates the same obligation under comparable circumstances. If both possess sufficient normative understanding and control, their culpability cannot be distinguished merely by their biological or artificial origins. Following an appropriate investigation, each may lose the authority to operate independently. The human driver may experience humiliation, distress, or even relief, while the Cybercab may experience no suffering. Nevertheless, the punitive character of the deprivation depends upon its justified relationship to the wrongdoing rather than the emotional response accompanying it.

I defend *normative punitive parity* on this conditional basis: where human and artificial agents satisfy comparable conditions of accountability, the absence of phenomenal suffering does not independently diminish the legitimacy of punishment. This does not establish equality of moral status, identical welfare interests, or equivalent sanctions. Differences in the agents' normative standing, vulnerabilities, and permissible treatment may justify different consequences. Moreover, culpability establishes grounds for punitive liability without making punishment obligatory in every case; mitigation, mercy, and competing moral considerations may constrain its imposition. Abbott and Sarch [1] examine the theoretical possibility of directly punishing artificial intelligence while identifying significant difficulties concerning its implementation in criminal law. My argument concerns the narrower possibility of institutionally situated artificial punishment, where culpable agency and legitimate punitive authority can be independently established. Normative punitive parity, therefore, concerns equality in the grounds of punitive justification, not equality in the experience or consequences of punishment. Whether deprivation of normative standing can constitute genuine adversity for an artificial agent without welfare interests remains a distinct philosophical objection, examined in Section 6.2.

5.4 Destruction and Punitive Theatre

The destruction of an artificial system illustrates why the apparent severity of an intervention cannot independently establish its punitive significance. Artificial agents may be copied, transferred, or reconfigured, complicating the relationship between the destroyed hardware and the accountable subject. Lemley and Casey [39] examine difficulties concerning remedies directed toward robots, while Abbott and Sarch [1] identify conceptual and practical problems associated with direct machine punishment. These considerations raise the question of whether destruction constitutes a genuine punitive consequence or merely a technological intervention. Consider the hypothetical Cybercab whose collision causes a family to lose a loved one. Suppose the manufacturer publicly dismantles the vehicle and presents its destruction as punishment. The intervention may prevent further harm or demonstrate institutional concern, but it does not establish that the Cybercab culpably violated its safety obligations. Nor does destroying its physical components necessarily reach the continuing accountable agent if its relevant practical organization persists elsewhere. Conversely, the termination of an accountable agent could constitute punitive deprivation where

its culpability, continuing identity, and the justification of such an irreversible consequence are independently established.

I describe as *punitive theatre* the public presentation of an intervention as punishment without establishing the conditions of genuine punitive accountability. It differs from *punitive projection*: the latter concerns the justificatory significance attributed to imagined offender suffering, whereas punitive theatre concerns the substitution of symbolic severity for justified condemnation. The two may reinforce one another when a machine's destruction, simulated remorse, or artificial expression of distress creates the appearance that wrongdoing has been adequately answered. Public censure is not inherently theatrical. An institution may legitimately disclose a violation and explain the deprivation imposed upon an accountable agent. The moral problem arises when the visibility of punishment replaces evidence of culpability, justified punitive authority, or the obligations owed to victims. Destruction cannot establish punitive desert through its apparent severity, just as the absence of visible suffering cannot invalidate an otherwise justified sanction.

5.5 The Distributed Accountability Constraint

Artificial punishment risks becoming a form of scapegoating when condemnation of the machine conceals the independent responsibilities of those who design, deploy, and supervise it. Even where an artificial agent exercises attributable control, its decisions remain embedded within technological and institutional arrangements. Santoni de Sio and Mecacci [55] distinguish different responsibility gaps associated with AI, while Mahi and Strand [43] examine accountability across layered forms of agency. Kong [35] further challenges the assumption that delegating decisions to autonomous AI diminishes the principal's responsibility for that delegation. These approaches demonstrate why responsibility for an artificial agent's culpable conduct must be distinguished from responsibility for authorizing its actions, determining its operating conditions, and managing the risks associated with its deployment.

I define *machine scapegoating* as the concentration of blame upon an artificial agent in a manner that obscures the independently grounded responsibilities of other participants. Scapegoating does not require that the machine be innocent. Elish's concept of *moral crumple zones* [18] illustrates how responsibility may be disproportionately assigned to human operators with limited control, while Feier et al. [22] examine how delegation to artificial agents may facilitate the avoidance of punishment. Joo [34] further investigates how perceived artificial agency influences blame attribution. These findings identify the risk of responsibility displacement without establishing that every instance of artificial punishment constitutes scapegoating. Consider the hypothetical Cybercab whose culpable disregard of a safety obligation causes a family to lose a loved one. Its wrongdoing may justify withdrawing its independent driving authority, but the manufacturer and operating institution may also bear responsibility for inadequate safeguards, foreseeable risks, or negligent deployment. Delegating autonomous decision-making authority cannot independently discharge these responsibilities, just as the manufacturer's involvement cannot automatically exclude the Cybercab's culpability. Each attribution requires an independent examination of the relevant obligations, knowledge, and control.

I formulate the *Distributed Accountability Constraint* as follows:

The justified punishment of an artificial agent neither transfers, reduces, nor extinguishes the independently grounded responsibilities of other participants. Each claim of punitive liability must be established through the relevant subject's own culpable conduct, while obligations of compensation, investigation, and prevention remain answerable to their distinct moral grounds.

This constraint distinguishes the legitimacy of an individual sanction from the adequacy of the overall response to wrongdoing. Punishing the Cybercab may be justified even if the operating institution has not fulfilled its obligations to the bereaved family. That failure does not necessarily invalidate the sanction, but it prevents the institution from presenting punishment as the completion of justice. Artificial responsibility is, then, additive rather than exculpatory: establishing the machine's culpability may identify an additional accountable subject without resolving the responsibilities of other participants.

The distinction also concerns the moral significance of victims' suffering. Thulin and Bicchieri [58] examine how moral outrage can motivate efforts to compensate victims, demonstrating that responses to wrongdoing need not be directed exclusively toward punishing offenders. This distinction matters because the family's legitimate demand for justice cannot be reduced to the suffering imposed upon the Cybercab. Moral outrage may motivate demands for retribution, but it can also direct attention toward the injury suffered and the obligations arising from it. Neither emotional response independently determines what justice requires. I describe the appropriate relationship between these responsibilities as *consequential recognition*: a response that combines truthful attribution, justified consequences, appropriate repair, and measures addressing foreseeable recurrence. The family's moral claims cannot be satisfied merely through the Cybercab's destruction, simulated remorse, or deprivation of authority. These interventions may possess different moral justifications, but none restores the victim's life or independently discharges the obligations generated by the collision. Meaningful recognition requires acknowledging the wrongdoing while addressing the responsibilities and preventable conditions associated with it. Artificial punishment may, consequently, possess genuine moral significance without phenomenal suffering, but its legitimacy cannot be confused with the fulfillment of justice as a whole. Punishing the Cybercab must not become the point at which responsibility ends. Section 6 examines the remaining objections concerning normative adversity, consciousness, artificial moral agency, and victims' claims.

6. Objections and Boundary Conditions

The possibility of artificial punishment without suffering faces several objections concerning the nature of moral agency, the significance of normative deprivation, and the justification of punitive consequences. These objections are particularly important because the attribution of artificial accountability cannot be established merely through autonomous performance or institutional recognition. They also clarify the conditions under which status-responsive punishment would remain morally unjustified.

6.1 Can a Moral Agent Be Incapable of Wrongdoing?

The relationship between moral agency and fallibility raises an important objection. If an agent consistently complies with its obligations, does this diminish its status as a moral agent? More fundamentally, could an agent that is incapable of wrongdoing still possess genuine moral competence? The ability to make an

incorrect decision should not be confused with the capacity to recognize and act upon moral reasons. An agent may exercise sophisticated normative judgment while consistently reaching permissible decisions. Conversely, a system may produce violations without possessing the capacities necessary for moral responsibility. Neither successful compliance nor the occurrence of error independently establishes the presence or absence of moral agency.

A further distinction concerns the source of an agent's reliability. A Cybercab that remains within the speed limit because prohibited actions are mechanically unavailable demonstrates compliance, but this fact alone provides little evidence of normative competence. Another system may recognize the reasons underlying the restriction and consistently comply through its own judgment. Fischer and Ravizza's account of reasons-responsive control [25] is relevant because it locates responsibility in the relationship between reasons and the mechanism governing conduct rather than in the frequency of mistakes. Nothing in the present argument requires every moral agent to be capable of culpable wrongdoing. The narrower condition is that punitive liability arises only where a particular agent could have complied with an applicable requirement but culpably failed to do so. An agent incapable of such a violation may possess moral competence without ever becoming liable to punishment.

6.2 Can an Agent Without Welfare Interests Be Punished?

The strongest objection to status-responsive punishment concerns whether an artificial agent without welfare interests can genuinely undergo punitive adversity. Punishment appears to require a consequence imposed upon the offender, not merely a change in the institution's arrangements. If the Cybercab experiences no suffering, possesses no interests in preserving its authority, and undergoes no reduction in well-being, what makes withdrawing its independent driving privileges a deprivation *for the agent itself*? The difficulty cannot be resolved by observing that its institutional status has changed. A manufacturer may alter a machine's permissions without thereby punishing it. Nor can describing the alteration as a *role-relative ontological loss* independently establish that a genuine punitive consequence has occurred. The account must explain why normative authority can belong to an accountable artificial subject and why its withdrawal constitutes an adverse alteration of that subject's practical standing.

This objection also arises in debates concerning artificial personhood and criminal punishment. Mamak [44] distinguishes ontological and functional approaches to AI personhood and identifies the capacity to experience and comprehend pain, alongside the ethical permissibility of its imposition, as relevant conditions for criminal punishment. His analysis challenges the assumption that functional autonomy or institutional recognition is sufficient for artificial punishability. I accept that merely assigning a machine responsibility for an institutional function does not establish its moral accountability. My argument, however, concerns whether phenomenal suffering is necessary for every form of morally justified punitive deprivation, particularly when the sanction removes authority previously exercised by a genuinely accountable agent rather than imposing criminal penalties upon a legally recognized person. The disagreement requires distinguishing *prudential harm* from *agent-relative normative loss*. Prudential harm concerns a setback to an entity's welfare or interests: something makes its condition worse for it. Agent-relative normative loss concerns the deprivation of powers or authority belonging to a subject's legitimate practical standing. These are different dimensions of evaluation. An agent may retain its capacities and experience no dissatisfaction

while losing the authority to exercise those capacities in a particular domain. This distinction does not establish that artificial agents necessarily lack welfare. Recent discussions of AI well-being and welfare [28, 40, 42, 48] leave that possibility open. The present argument addresses the more demanding case in which punitive deprivation must be justified without appealing to welfare interests at all.

The central argument begins with the relationship between moral accountability and practical authority. An accountable agent is not merely a system through which an institution produces outcomes. It must possess the capacity to recognize applicable obligations, exercise judgment under those obligations, and determine its conduct through decisions attributable to itself. Where such an agent is legitimately authorized to act independently, its practical authority is exercised through its own normative capacities, even though the institution establishes the boundaries of that authority. The distinction matters because the institution's control over the conditions of authorization does not mean that the authority is exercised exclusively by the institution. An accountable agent can possess powers whose legitimacy depends upon institutional recognition without those powers being reducible to the institution's use of equipment.

This provides the basis for an argument from *accountable practical authority*. If an artificial agent genuinely exercises independent normative judgment under obligations for which it is answerable, and if it possesses legitimate authority to act upon that judgment, the scope of its authorized discretion forms part of its practical standing. That standing determines which decisions the agent may make in its own capacity as an accountable subject, rather than merely which operations its hardware can perform. Withdrawing that discretion changes what the agent itself is entitled to determine and execute. The loss is, therefore, attributable to the subject whose authority has been restricted, not simply to the institution that imposed the restriction.

The crucial inference is that a deprivation can be adverse relative to an agent's normative standing without being adverse to its welfare. This does not mean that every unfavorable institutional decision constitutes punishment. It means that the legitimate scope of an accountable subject's agency is itself a dimension along which deprivation can occur. The relevant comparison concerns the authority the same subject was entitled to exercise before and after the intervention. A restriction that removes previously authorized independent judgment diminishes that subject's normative powers even if its computational abilities, preferences, and phenomenal condition remain unchanged. The deprivation is agent-relative because it affects the scope of action legitimately available to the agent as an accountable decision-maker, not because the agent necessarily experiences the restriction as undesirable.

Consider the hypothetical Cybercab whose culpable disregard of a pedestrian safety obligation results in a fatal collision. Before the violation, suppose the Cybercab possesses legitimate authority to evaluate road conditions, resolve uncertainty, and make independent driving decisions under obligations protecting pedestrians. Its decisions are attributable to its own norm-responsive capacities rather than merely implementing externally predetermined instructions. Following an investigation establishing culpability, the operating institution withdraws its independent driving authority while preserving its computational capacities and practical continuity. The Cybercab can still recognize relevant safety obligations, assess road conditions, and determine appropriate actions, but it can no longer legitimately execute certain decisions without external authorization.

The consequence is not simply that the manufacturer chooses to use the system differently. Under the stipulated conditions, the authority withdrawn was previously exercised by the Cybercab itself, under obligations for which it was answerable. Before the sanction, its independent judgment could legitimately determine what action to take within its authorized domain. Afterward, an external authority must approve decisions that the Cybercab could previously make on its own. The agent's technical abilities remain, but the normative force of its independent decisions has changed. This alteration constitutes a *role-relative ontological loss* because a constitutive dimension of its authorized practical existence has been withdrawn. The loss concerns what the continuing agent may legitimately do through its own judgment, not what it feels about that restriction. This argument also explains why ordinary technological restrictions do not automatically qualify as agent-directed deprivation. A machine that merely executes delegated instructions possesses no independent accountable authority for the institution to withdraw. Restricting its operation changes the permissible use of equipment. By contrast, withdrawing authority from a genuinely normatively addressable agent alters the standing of the subject that previously exercised that authority and remained answerable for its misuse. The difference cannot be established by institutional designation alone. It depends upon independently justified claims about normative competence, attributable control, practical continuity, and the authority exercised through those capacities.

Nevertheless, establishing agent-relative deprivation is not sufficient to establish punishment. The moral character of the intervention also depends upon its grounds. A Cybercab suspended because of mechanical failure may lose independent operating authority for protective reasons, without being culpable. A Cybercab deprived of the same authority because of established wrongdoing may undergo punishment where the sanction is imposed retrospectively, expresses legitimate censure, and remains proportionate to the violation. The technical restriction may be identical, but the normative relationships differ. In the punitive case, the deprivation addresses the accountable subject as the author of a culpable violation and restricts authority it previously exercised under the obligation it breached.

Diamantis's discussions of corporate punishment and identity [15, 16] provide a limited comparison. Corporate sanctions can alter the legitimate powers and institutional standing of an organized subject without requiring a unified phenomenal experience of punishment. Nevertheless, corporations possess legally recognized interests and involve human stakeholders whose welfare may be affected. The comparison, based on this, supports the conceptual possibility of sanctions directed toward organized agency without establishing that artificial systems possess the requisite moral standing or that welfare-independent deprivation is necessarily punitive. A welfare-based theorist may still object that a consequence incapable of making its recipient worse off cannot constitute genuine punishment. This objection identifies the precise point of disagreement. My argument establishes a form of deprivation relative to accountable practical authority; it does not establish prudential harm or demonstrate that every defensible theory of punishment must recognize such deprivation as sufficient. The further normative claim is that where punishment legitimately responds to culpable wrongdoing through authoritative censure, the offender's loss need not be evaluated exclusively through its well-being. A restriction may constitute punitive adversity because it withdraws a power belonging to the accountable subject whose wrongful exercise of authority justified the sanction. Its significance derives from the relationship between the violated obligation, the agent's responsibility for that violation, and the justified limitation of its subsequent authority.

This does not make punishment a matter of institutional declaration. Neither the ability to remove a machine from service nor the institution's decision to describe that removal as punishment establishes the relevant conditions. The account requires an independently accountable subject, authority genuinely exercised through that subject's normative capacities, an attributable culpable violation, and a legitimate deprivation addressing that violation. Without these conditions, the intervention remains technological governance, institutional regulation, or symbolic condemnation rather than punishment of the artificial agent. Punishment without suffering, therefore, depends upon more than the absence of pain or the withdrawal of operational permission. It requires a continuing accountable subject whose legitimate practical authority can be diminished in response to its own culpable wrongdoing. The deprivation is genuinely agent-relative because it removes powers previously exercised by that subject under obligations for which it was answerable, even where the loss produces no suffering or setback to welfare. A welfare-based account may deny that normative-status adversity is sufficient for punishment, but that denial requires defending welfare loss as a necessary condition rather than assuming it. The possibility defended here is that an accountable artificial agent can undergo morally justified punitive deprivation through the loss of its own normative authority, without phenomenal suffering constituting the measure or justification of that loss.

6.3 Does Normative Understanding Require Consciousness?

A further objection challenges whether artificial agents can genuinely understand moral obligations without phenomenal consciousness or embodied moral experience. A system may identify obligations, explain moral requirements, and modify its behavior without understanding why those requirements possess normative authority. Its apparent moral reasoning may consist entirely of computational operations that reproduce the structure of ethical judgment without possessing its underlying significance. Véliz [61] challenges the attribution of moral agency to algorithms, while Bernáth [8] examines whether phenomenal consciousness is necessary for moral responsibility. Mamak [45, 46] raises related questions concerning artificial agency, criminal responsibility, and suffering. These objections challenge the assumption that successful normative performance establishes genuine moral understanding.

Mosakas [50] develops a consciousness-based account of artificial moral status and personhood, arguing that phenomenal consciousness is necessary for the moral standing associated with rights and personhood. Although moral personhood and moral responsibility are distinct concepts, this position raises an important question for artificial punishability: can an entity bear moral obligations and become answerable for violating them without possessing the consciousness ordinarily associated with moral standing? Zwitter [63] advances a different but related challenge through the *Embodied Ethics Alignment Problem*. He argues that human moral agency emerges from embodiment, affect, vulnerability, social relationships, and the developmental formation of character. Artificial systems may reproduce ethically acceptable behavior through computational constraints without thereby possessing moral agency in this stronger sense. These arguments suggest that normative competence cannot be established merely through rule-following, optimization, or linguistic explanations.

The objection is particularly significant for the hypothetical Cybercab. A system may recognize uncertainty about a pedestrian's intentions and identify stopping as the appropriate action without understanding the moral significance of human vulnerability. Its decisions may reflect reliable safety optimization rather than an

appreciation of why another person's life places demands upon its conduct. My account, consequently, requires more than behavioral conformity. Normative competence concerns whether moral requirements function as governing reasons within the agent's practical judgment, particularly when they conflict with operational objectives or require reconsidering an intended action. Nevertheless, even consistent responsiveness to such reasons may not establish that the agent possesses the embodied or phenomenal conditions emphasized by Zwitter and Mosakas. The central disagreement concerns whether those conditions are necessary for moral agency or whether an artificial agent could possess a different but sufficient form of normative understanding. My account does not assume that computational reasons-responsiveness settles this question. It identifies the capacity to recognize and act under binding moral reasons as a necessary condition of artificial accountability, while acknowledging that the possibility of such recognition without consciousness remains contested. If phenomenal consciousness or embodied ethical experience is indispensable to genuine normative competence, an artificial agent lacking those conditions would fail the requirements of normative addressability and could not become morally punishable under the proposed account.

This objection nevertheless requires distinguishing the conditions of *moral agency* from the conditions of *punitive adversity*. Whether consciousness is necessary for understanding an obligation is a different question from whether suffering is necessary for punishment once culpable responsibility has been established. Even if consciousness proves indispensable to moral agency, it does not follow without further argument that phenomenal suffering must constitute the sanction imposed upon an accountable subject. The present theory does not establish that unconscious machines are morally responsible. It defends the conditional possibility that an artificial agent satisfying the requirements of normative accountability may undergo justified punitive deprivation without suffering. The possibility of punishment without suffering, therefore, depends upon establishing genuine normative agency, but it does not require treating the experience of suffering as either evidence of that agency or a necessary component of its punishment.

6.4 Can Genuine Artificial Accountability Be Established?

Even if artificial moral accountability is conceptually possible, identifying it in a particular system presents a substantial evidential challenge. An autonomous agent may appear to recognize an obligation while its behavior is generated by mechanisms that do not support genuine normative attribution. Explanations produced after an incident may also fail to represent the considerations that actually governed the decision. Conradie and Nagel [13] provide a relevant challenge to unsupported claims of artificial agency and responsibility. The difficulty is, hence, not merely whether an artificial system behaves incorrectly, but whether its conduct supplies adequate grounds for attributing a culpable decision to an accountable subject.

In the Cybercab case, an investigation would need to reconstruct the information available before the violation, the safety requirements applicable at that moment, the alternative actions available, and the mechanism through which the final decision was produced. Evidence that the system represented the relevant obligation would be insufficient without establishing that the obligation could guide its conduct. Its responses under comparable circumstances, including cases where operational objectives conflict with safety requirements, would provide further evidence of whether normative considerations govern its decisions. The investigation would also need to distinguish decisions attributable to the system from those

imposed by developers, external commands, or technical constraints. Such evidence could strengthen or weaken a claim of accountability without necessarily resolving every philosophical disagreement concerning artificial understanding. Where the evidence remains insufficient to establish normative competence, control, or continuity, punitive liability should not be attributed. This evidential restraint does not prevent institutions from taking necessary protective measures. It prevents uncertainty about artificial moral agency from being converted into an unsupported judgment of guilt.

6.5 Can Punishment Without Suffering Adequately Recognize Victims?

A final objection concerns whether punishment that produces no suffering can adequately recognize the moral gravity of wrongdoing, particularly when the violation results in the loss of human life. Consider the hypothetical Cybercab whose culpable disregard of a pedestrian safety obligation causes a fatal collision, leaving a family mourning the death of a loved one. Withdrawing the Cybercab's independent driving authority may constitute a justified punitive deprivation, but can such a consequence adequately express the seriousness of the wrongdoing when the accountable agent experiences no pain, remorse, or distress? The family's demand for punishment should not be dismissed as an irrational desire for revenge. It raises a legitimate moral question concerning whether the response recognizes the victim's death, condemns the violation, and acknowledges the wrong suffered by those left behind.

Feinberg's account of the expressive function of punishment [23] provides a basis for distinguishing the moral meaning of a sanction from the suffering it produces. Punishment communicates condemnation of wrongdoing, but its expressive significance depends upon the relationship between the violation and the justified consequence imposed. Feinberg's analysis of harm to others [24] further illuminates the moral importance of conduct that wrongfully damages protected human interests, although his harm principle does not independently establish artificial punitive liability. In the Cybercab case, the wrong derives from the culpable violation of an obligation protecting human life, not merely from the breach of an operational requirement. The pedestrian's death, therefore, gives the violation a moral seriousness that must inform the justification and proportionality of punishment.

Nevertheless, the suffering imposed upon an offender cannot restore the victim's life or determine whether the family's loss has been adequately recognized. Bandes [5] questions whether punishment necessarily provides emotional closure, while Cesur [12] examines the relationship between retributive punishment, vindication, and healing. These considerations reveal why punitive adequacy cannot be reduced to the emotional satisfaction associated with an offender's suffering. A justified withdrawal of the Cybercab's normative authority may express condemnation of its culpable wrongdoing without producing distress. Conversely, publicly dismantling the vehicle may create an appearance of severe punishment without establishing culpability or providing meaningful recognition of the family's moral claims.

This does not establish that every non-afflictive sanction adequately expresses the gravity of a fatal violation. The seriousness of the wrongdoing, the Cybercab's degree of culpability, and the significance of the authority withdrawn must determine whether the punitive consequence is proportionate. Nor can punishment alone fulfill the obligations arising from the pedestrian's death. The family may possess independent claims to

truthful explanation, compensation, institutional accountability, and measures preventing comparable harm. These obligations remain morally binding irrespective of whether the Cybercab itself can be punished. The moral adequacy of punishment must, thus, be distinguished from the adequacy of the overall response to wrongdoing. Punishment without suffering may legitimately condemn an accountable artificial agent through the deprivation of its normative standing, but it cannot by itself constitute justice for the bereaved family. The absence of offender suffering does not necessarily diminish the moral significance of a justified sanction, just as the presence of suffering cannot substitute for recognition, repair, and the fulfillment of independent obligations owed to victims.

7. Conclusion: Punishment After Pain

The possibility of artificial punishment is not determined by an agent's origin or its capacity to experience suffering. It depends on whether an artificial agent can become morally accountable for wrongdoing and whether a justified sanction can impose a genuinely punitive consequence upon it. *Normative addressability* establishes the conditions for possible accountability, while *status-responsive punishment* explains how culpable wrongdoing may justify the deprivation of an agent's normative standing. Neither moral responsibility nor punitive legitimacy follows from technical autonomy alone. The central philosophical contribution concerns the ontological dimension of punitive deprivation. Where independent decision-making authority constitutes part of an accountable agent's practical position, its justified withdrawal may produce a *role-relative ontological loss*. Such a loss does not necessarily destroy the agent's identity, eliminate its remaining capacities, or produce subjective distress. Rather, it alters the authorized mode of existence through which the agent participates in a morally structured world. This provides a basis for understanding punitive adversity independently of phenomenal suffering, while recognizing that the attribution of genuine normative standing to artificial agents remains philosophically contested.

The account also establishes limits on the moral significance of punishment. *Punitive projection* reveals why actual or imagined suffering cannot independently determine the adequacy of justice, while the *Distributed Accountability Constraint* preserves the responsibilities of designers, operators, and institutions. *Normative punitive parity*, therefore, concerns consistency in the grounds of culpability and punitive accountability, not identical moral status or sanctions for human and artificial agents. The conclusion remains conditional and does not establish that existing artificial systems are morally punishable. *Punishment after pain* identifies a philosophical possibility: an accountable agent may be genuinely punished through the justified loss of its own normative standing, even when that loss is not experienced as suffering.

Declarations

Declaration of Generative AI Use

During the preparation of this manuscript, the author used OpenAI's ChatGPT as an AI-assisted editorial and critical review tool after completing the original manuscript. Its role was to improve linguistic clarity, coherence, precision, and presentation; identify repetitions, inconsistencies, and potential weaknesses in the exposition; provide critical feedback; and suggest alternative wording for author-developed arguments. The author independently conceived the research questions, developed the theoretical framework, formulated the

philosophical arguments, determined the manuscript's structure, and established its analysis and conclusions. All AI-assisted editorial suggestions, including proposed textual revisions, were critically evaluated, modified where necessary, and approved by the author. The author retains full responsibility for the manuscript's intellectual content, accuracy, originality, references, and final presentation.

Author Contribution

G.A. conceived the research idea, developed the theoretical framework and philosophical arguments, conducted the literature review and critical analysis, and prepared and revised the manuscript. G.A. reviewed and approved the final version and takes full responsibility for its content.

References

1. Abbott, R., Sarch, A.: Punishing artificial intelligence: Legal fiction or science fiction. UC Davis Law Review 53(1), 323–384 (2019). https://lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/53-1_Abbott_Sarch.pdf
2. Abdelal, G.: Can machines write? Recursive re-embodiment and distributed agency in human–AI writing. SSRN preprint (2026a). https://doi.org/10.2139/ssrn.7366889
3. Abdelal, G.: Epistemic action shielding for autonomous agents under uncertainty (Version 0.1.1). Computer software. Zenodo (2026b). https://doi.org/10.5281/zenodo.22169778
4. Avramides, A.: Other minds. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Summer 2019 edn. Metaphysics Research Lab, Stanford University (2019). https://plato.stanford.edu/archives/sum2019/entries/other-minds/
5. Bandes, S.A.: Victims, “closure,” and the sociology of emotion. Law and Contemporary Problems 72(2), 1–26 (2009). https://scholarship.law.duke.edu/lcp/vol72/iss2/2
6. Baum, K., Mantel, S., Schmidt, E., Speith, T.: From responsibility to reason-giving explainable artificial intelligence. Philosophy & Technology 35, Article 12 (2022). https://doi.org/10.1007/s13347-022-00510-w
7. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Minds and Machines 30(2), 195–218 (2020). https://doi.org/10.1007/s11023-020-09525-8
8. Bernáth, L.: Can autonomous agents without phenomenal consciousness be morally responsible? Philosophy & Technology 34, 1363–1382 (2021). https://doi.org/10.1007/s13347-021-00462-7
9. Brady, M.S.: Suffering and punishment. In: Amaya, A., Del Mar, M. (eds.) Virtue, Emotion and Imagination in Law and Legal Reasoning, pp. 139–156. Hart Publishing (2020). https://eprints.gla.ac.uk/242961/1/242961.pdf
10. Carlsmith, K.M., Darley, J.M., Robinson, P.H.: Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology 83(2), 284–299 (2002). https://doi.org/10.1037/0022-3514.83.2.284

11. Cassell, E.J.: The nature of suffering and the goals of medicine. New England Journal of Medicine 306(11), 639–645 (1982). https://doi.org/10.1056/NEJM198203183061104
12. Cesur, H.: Retributive punishment: Between vindication and healing. International Review of Victimology 31(2), 185–202 (2025). https://doi.org/10.1177/02697580241309352
13. Conradie, N.H., Nagel, S.K.: No agent in the machine: Being trustworthy and responsible about AI. Philosophy & Technology 37, Article 72 (2024). https://doi.org/10.1007/s13347-024-00760-w
14. Danaher, J.: Robots, law and the retribution gap. Ethics and Information Technology 18(4), 299–309 (2016). https://doi.org/10.1007/s10676-016-9403-3
15. Diamantis, M.E.: Clockwork corporations: A character theory of corporate punishment. Iowa Law Review 103, 507–569 (2018). https://ilr.law.uiowa.edu/sites/ilr.law.uiowa.edu/files/2023-02/ILR103-2-Diamantis2.pdf
16. Diamantis, M.E.: Corporate essence and identity in criminal law. Journal of Business Ethics 154, 955–966 (2019). https://doi.org/10.1007/s10551-018-3892-4
17. Duff, R.A.: Punishment, Communication, and Community. Oxford University Press (2001).
18. Elish, M.C.: Moral crumple zones: Cautionary tales in human–robot interaction. Engaging Science, Technology, and Society 5, 40–60 (2019). https://doi.org/10.17351/ests2019.260
19. Eng, A., Yam, K.C.: Artificial intelligence and machines are capable of harm, but not of experiencing punishment. Journal of Experimental Psychology: General. Advance online publication (2026). https://doi.org/10.1037/xge0001930
20. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: A three-factor theory of anthropomorphism. Psychological Review 114(4), 864–886 (2007). https://doi.org/10.1037/0033-295X.114.4.864
21. Faroldi, F.L.G.: Reasons-based artificial agents. AI and Ethics 6, Article 77 (2026). https://doi.org/10.1007/s43681-025-00932-0
22. Feier, T., Gogoll, J., Uhl, M.: Hiding behind machines: Artificial agents may help to evade punishment. Science and Engineering Ethics 28, Article 19 (2022). https://doi.org/10.1007/s11948-022-00372-7
23. Feinberg, J.: The expressive function of punishment. The Monist 49(3), 397–423 (1965). https://www.scribd.com/document/619737648/1965-Feinberg-The-Expressive-Function-of-Punishment
24. Feinberg, J.: The Moral Limits of the Criminal Law, vol. 1: Harm to Others. Oxford University Press (1987).
25. Fischer, J.M., Ravizza, M.: Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press (1998).
26. Fleisher, W., Cibralic, B., Basl, J., Ricks, V., Smith, M.N.: Responsibility and accountability in an algorithmic society. Philosophy & Technology 38, Article 144 (2025). https://doi.org/10.1007/s13347-025-00970-w
27. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds and Machines 14, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d
28. Goldstein, S., Kirk-Giannini, C.D.: AI wellbeing. Asian Journal of Philosophy 4, Article 25 (2025). https://doi.org/10.1007/s44204-025-00246-2
29. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science 315(5812), 619 (2007). https://doi.org/10.1126/science.1134475

30. Hanna, N.: Say what? A critique of expressive retributivism. Law and Philosophy 27(2), 123–150 (2008). https://doi.org/10.1007/s10982-007-9014-6
31. Hanna, N.: Why punitive intent matters. Analysis 81(3), 426–435 (2021). https://doi.org/10.1093/analys/anaa068
32. Hart, H.L.A.: Punishment and Responsibility: Essays in the Philosophy of Law. Clarendon Press (1968).
33. International Association for the Study of Pain: IASP announces revised definition of pain. International Association for the Study of Pain (2020). https://www.iasp-pain.org/publications/iasp-news/iasp-announces-revised-definition-of-pain/
34. Joo, M.: It's the AI's fault, not mine: Mind perception increases blame attribution to AI. PLOS ONE 19(12), e0314559 (2024). https://doi.org/10.1371/journal.pone.0314559
35. Kong, K.C.Y.: The delegation illusion: Why deploying autonomous AI agents does not diminish principal responsibility. AI and Ethics 6, Article 518 (2026). https://doi.org/10.1007/s43681-026-01383-x
36. Kraaijeveld, S.R.: Debunking (the) retribution (gap). Science and Engineering Ethics 26(3), 1315–1328 (2020). https://doi.org/10.1007/s11948-019-00148-6
37. Kropf, M., Spöck, C., Werner, R.: Blame the robot: Role responsibility and ethical issues regarding AI-based care robots. International Journal of Social Robotics 18, Article 30 (2026). https://doi.org/10.1007/s12369-026-01369-z
38. Lee, M., Ruijten, P.A.M., Frank, L.E., de Kort, Y.A.W., IJsselsteijn, W.A.: People may punish, not blame, robots. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 715, pp. 1–11. Association for Computing Machinery (2021). https://doi.org/10.1145/3411764.3445284
39. Lemley, M.A., Casey, B.: Remedies for robots. University of Chicago Law Review 86, 1311–1396 (2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3223621
40. Lennon, P.: How seriously should we take AI welfare? Constraints from the epistemology of consciousness. Philosophy and Phenomenological Research 113(2), 441–452 (2026). https://doi.org/10.1111/phpr.70148
41. Lima, G., Cha, M., Jeon, C., Park, K.S.: The conflict between people's urge to punish AI and legal systems. Frontiers in Robotics and AI 8, Article 756242 (2021). https://doi.org/10.3389/frobt.2021.756242
42. Long, R., Sebo, J., Sims, T.: Is there a tension between AI safety and AI welfare? Philosophical Studies 182, 2005–2033 (2025). https://doi.org/10.1007/s11098-025-02302-2
43. Mahi, E., Strand, M.: Layered agency as a basis for allocating accountability for AI. AI & Society (2026). https://doi.org/10.1007/s00146-026-03224-1
44. Mamak, K.: AI personhood, criminal law, and punishment. In: Hacker, P. (ed.) Oxford Intersections: AI in Society. Oxford University Press (2025). https://www.researchgate.net/publication/391406617_AI_Personhood_Criminal_Law_and_Punishment
45. Mamak, K.: AI agency and criminal responsibility: A category mistake. Inquiry (2026a). https://doi.org/10.1080/0020174X.2026.2720904
46. Mamak, K.: In defense of artificial suffering. Philosophical Studies (2026b). https://doi.org/10.1007/s11098-026-02493-2

47. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
48. Moret, A.: AI welfare risks. Philosophical Studies (2025). https://doi.org/10.1007/s11098-025-02343-7
49. Morrison, A.D., Mota, R., Wilhelm, W.J., Jr.: Relationships, authority, and reasons: A second-personal account of corporate moral agency. Business Ethics Quarterly 32(2), 322–347 (2022). https://doi.org/10.1017/beq.2021.18
50. Mosakas, K.: Artificial consciousness and moral personhood. In: Hacker, P. (ed.) Oxford Intersections: AI in Society. Oxford University Press (2025). https://www.researchgate.net/publication/391525181_Artificial_Consciousness_and_Moral_Personhood
51. Mulligan, C.: Revenge against robots. South Carolina Law Review 69, 579–608 (2018). https://dx.doi.org/10.2139/ssrn.3016048
52. Nussbaum, M.C.: Anger and Forgiveness: Resentment, Generosity, Justice. Oxford University Press (2016).
53. Poama, A.: Punishment without pain: Outline for a non-afflictive definition of legal punishment. Philosophy and Public Issues 5(1), 97–134 (2015).
54. Ross, W.D.: The Right and the Good. Edited by P. Stratton-Lake. Clarendon Press (2002).
55. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology 34, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
56. Sparrow, R.: Killer robots. Journal of Applied Philosophy 24(1), 62–77 (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x
57. Strawson, G.: The impossibility of moral responsibility. Philosophical Studies 75(1–2), 5–24 (1994). http://www.jstor.org/stable/4320507
58. Thulin, E.W., Bicchieri, C.: “I’m so angry I could help you”: Moral outrage as a driver of victim compensation. Social Philosophy and Policy 32(2), 146–160 (2016). https://doi.org/10.1017/S0265052516000145
59. Tigard, D.W.: Artificial moral responsibility: How we can and cannot hold machines responsible. Cambridge Quarterly of Healthcare Ethics 30(3), 435–447 (2021). https://doi.org/10.1017/S0963180120000985
60. Vacek, D.: Responsible artificial intelligence? AI & Society 41, 843–855 (2026). https://doi.org/10.1007/s00146-025-02604-3
61. Véliz, C.: Moral zombies: Why algorithms are not moral agents. AI & Society 36, 487–497 (2021). https://doi.org/10.1007/s00146-021-01189-x
62. Vierkant, T.: Does accountability require agency? Comment on responsibility and accountability in the algorithmic society. Philosophy & Technology 39, Article 13 (2026). https://doi.org/10.1007/s13347-025-01014-z
63. Zwitter, A.: The embodied ethics alignment problem of AI. Discover Artificial Intelligence 6, Article 852 (2026). https://doi.org/10.1007/s44163-026-01905-w