The Neutrality of Things

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, June 2026 (v3)

ABSTRACT

The philosophy of technology has long been organized around a binary between instrumentalism, which holds that artifacts are tools whose moral status derives from their use, and substantivism, which holds that artifacts carry embedded values and shape moral possibilities in ways that exceed individual intent. This paper argues that the binary misplaces moral accountability for foreseeable technological harm. The binary forces a choice between locating accountability in the artifact or in the user, while leaving the design decision outside the analytical frame. The paper defends a narrowed neutrality thesis, grounded in the distinction between intrinsic and relational properties and in a three-layer analysis of intent at the build, deployment, and end-use levels; sketches an illustrative pre-industrial genealogy of maker accountability and intermediary responsibility across three legal traditions; and proposes a distributed responsibility framework with four actors (artifact, designer, user, intermediary) and a collapse threshold under which specific knowledge of foreseeable harm plus affirmative choice concentrates responsibility on a single actor. The framework is offered as an analytical tool for the current legal and policy debates over platform-mediated harm to adolescent populations, with a boundary condition identifying where adaptive AI interaction begins to require a different framework than the one developed here.

Keywords: philosophy of technology; ethics of design; design accountability; technological mediation; affordances; product liability; distributed responsibility; adolescent harm

I. INTRODUCTION

"Who Put That There?"

A father watches his child walk through a hallway and strike his hand against the edge of a table. The child winces. The father, half-joking, asks: "Well, who put that there?"

The joke contains, in miniature, the entire problem this paper addresses. The table is not at fault; it is an object, incapable of intent. The child bears some responsibility; awareness of one's environment is a reasonable expectation proportional to developmental capacity. But the father's sarcastic question points to a third actor that neither the child nor the table can account for: the person who decided to place a hard-cornered object at hip height in a corridor where people move quickly and do not look down. That is a placement decision with a foreseeable outcome.

Nobody writes philosophy papers about table placement. The stakes are a bruised hand and some colorful language. But scale this exact structure to a social media platform deployed to three billion users, and the same analysis collapses.

The object (the app, the screen, the device) absorbs blame from one direction. The user (the teenager, the parent, the educator) absorbs blame from the other direction. The placement decision, which is to say the design decision, the architectural choice, the deliberate engineering of the interaction conditions, falls into a gap that the public discourse has not yet figured out how to name.

This paper names it. The central claim is that moral accountability for foreseeable technological harm attaches primarily to the design decisions that structure the interaction between artifacts and users, rather than to the artifact alone or to the user alone. The artifact is morally neutral in the limited sense developed in Section III: it lacks moral agency. The user bears proportional responsibility calibrated to agency, capacity, information, and developmental status. The design decision, made by identifiable agents with knowledge of foreseeable outcomes, carries the moral weight that the binary framing has consistently misplaced.

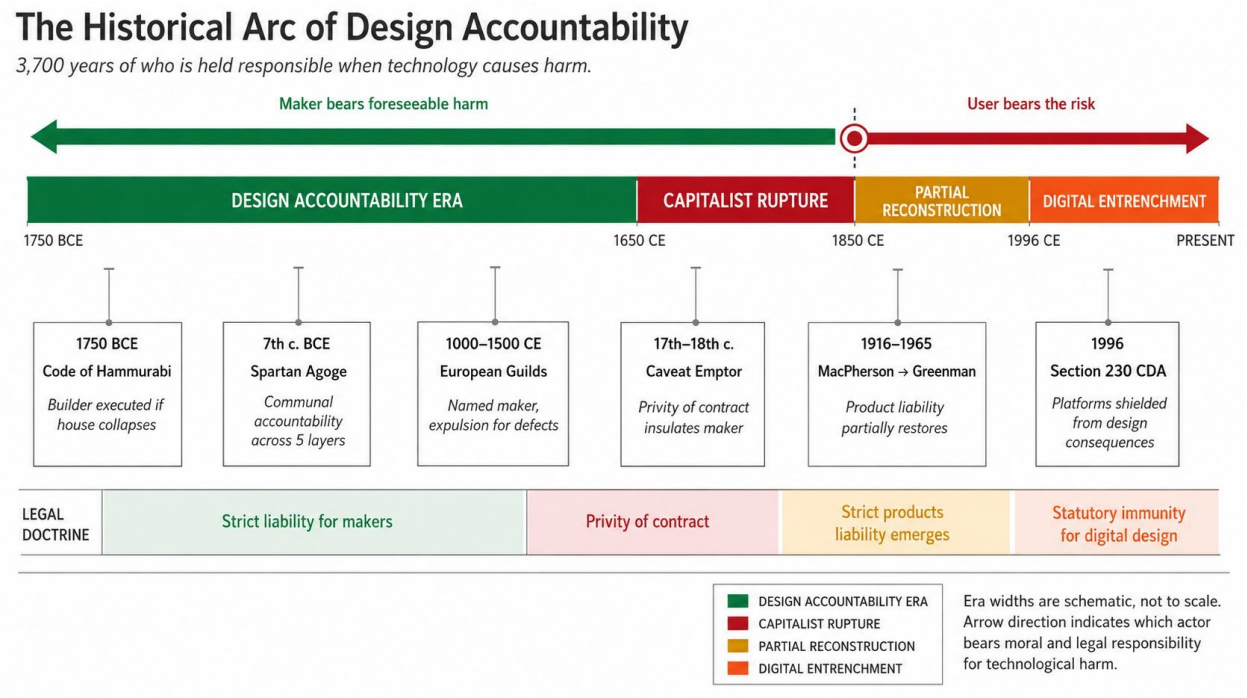

Figure 1: The historical arc of design accountability. From Hammurabi's builder liability code (c. 1750 BCE) through the Spartan agoge (7th c. BCE) and medieval European guilds (1000-1500 CE), several pre-industrial legal and institutional traditions located accountability for foreseeable harm with makers, builders, or communal intermediary systems. The capitalist rupture (red, 17th-18th c.) inverted the accountability vector through caveat emptor and privity of contract. Partial reconstruction occurred during the 20th century (amber) through product liability law (*MacPherson v. Buick*, 1916; *Greenman v. Yuba Power Products*, 1963). Section 230 of the Communications Decency Act (orange, 1996) extended the insulation into the digital age, shielding not merely artifacts but the active design choices that determine how artifacts interface with users.

A. Practical Stakes and Positioning

The framework developed here is not an abstract exercise. The current legal moment is structured by litigation that turns on exactly the question this paper addresses. The federal multidistrict litigation MDL No. 3047, *In re Social Media Adolescent Addiction/Personal Injury Products Liability Litigation* (N.D. Cal., before Judge Yvonne Gonzalez Rogers), and the parallel California state coordinated proceeding JCCP 5255, *Social Media Cases* (L.A. Super. Ct., before Judge Carolyn B. Kuhl), together consolidate thousands of cases advancing theories that target platform conduct as product

design rather than as third-party content distribution. In March 2026 a Los Angeles jury in the first bellwether trial of JCCP 5255, *K.G.M. v. Meta Platforms, Inc.*, found Meta and Google's YouTube negligent for platform design and failure to warn, awarding $6 million in combined compensatory and punitive damages. Snap and TikTok settled with the K.G.M. plaintiff before trial. Federal MDL school-district bellwether trials are scheduled for late 2026. Taken together, these developments suggest that design-defect theories are no longer merely testing the boundary of Section 230 but are beginning to survive long enough to reach juries and settlement pressure. The framework this paper develops is offered as an analytical tool for those debates rather than as a replacement for them.

The contribution is positioned within the substantivist and postphenomenological tradition rather than against it. Verbeek's *What Things Do* (2005) and *Moralizing Technology* (2011) established that morality is a hybrid affair, neither exclusively located in things nor exclusively in humans, and the present framework develops a more granular four-actor decomposition of that hybrid. Verbeek and Kudina (2018) on the Collingridge dilemma and the mediated value of privacy demonstrates how technological mediation works through the design of specific features, which is the territory on which the present paper's argument operates. The paper makes three primary contributions: a defended version of the neutrality thesis that survives the affordance-theoretic critique (Section III); an illustrative pre-industrial genealogy of maker accountability (Section V); and a distributed responsibility framework with a collapse threshold (Section VII).

II. The Binary That Failed

The philosophy of technology has long been organized around a binary between instrumentalism and substantivism. Instrumentalists hold that technologies are tools whose moral status derives from the uses to which they are put. Substantivists hold that technologies carry embedded values, mediate human action and perception, and shape moral and political possibilities in ways that exceed the intentions of any individual user. The binary is venerable and productive, and, this paper argues, inadequate for the work that the philosophy of technology is now being asked to do. The inadequacy is not in either pole's substantive insights, which are largely correct as far as they go. The inadequacy is in the binary itself, which frames the question of moral accountability as a choice between locating it in the artifact or locating it in the user, while leaving the design decision (the actual locus of moral weight) outside the analytical frame.

A. The Instrumentalist Position

The strongest contemporary statement of instrumentalism is Pitt's. His argument is that technologies, considered as physical objects, do not possess intentions or moral properties; they possess physical properties, and the moral properties that attach to technological situations attach to the human agents who design, deploy, and use them. The classic Pitt example is the loaded firearm: the firearm has the physical property of being able to discharge a projectile, but the firearm has no intention to harm and no responsibility for any harm that might result. The responsibility attaches to the agents who designed, manufactured, distributed, possessed, and discharged it.

Pitt's instrumentalism is not the strawman that some substantivist critiques present. It is a careful position that draws a precise distinction between the physical properties of artifacts and the moral properties of agents. The neutrality this paper defends is a refinement of Pitt's position rather than a rejection of it.

B. The Substantivist Critique

The substantivist tradition runs through Heidegger's analysis of technology as a mode of revealing, Ellul's account of technique as an autonomous social force, the Frankfurt School's critique of technological rationality, Feenberg's critical theory of technology, Verbeek's postphenomenological mediation theory, and the more recent affordance-theoretic work

of Heyndels (2023). The unifying claim is that technologies are not neutral instruments but carry embedded values and shape moral and political possibilities.

Several lines within this tradition bear directly on the framework developed here. Verbeek's mediation theory, developed in *What Things Do* (2005) and extended in *Moralizing Technology* (2011), holds that morality is a hybrid affair, distributed across humans and the technologies that co-shape their actions and decisions. The present paper accepts this distributed-morality claim and contributes a more granular four-actor decomposition (Section VII) that specifies how the hybrid distributes across designer, user, intermediary, and the artifact's substrate, with a collapse threshold for cases where one actor's specific knowledge of foreseeable harm shifts the proportional weight.

Langdon Winner's "Do Artifacts Have Politics?" (1980) is the canonical statement that technological artifacts can be inherently political. Winner distinguishes two senses: artifacts can be designed in ways that produce political effects (his example is Robert Moses's low overpasses on the Long Island parkways, which Winner argues were designed to exclude bus traffic carrying lower-income, predominantly Black New Yorkers to the public beaches), and artifacts can be such that their adoption requires particular political arrangements (his example is nuclear power, which requires centralized institutional structures to function safely). Winner's first category aligns closely with the design-decision framework developed here: the political effects flow from design decisions made by identifiable agents, and accountability attaches at the design layer.

Bruno Latour's "Where Are the Missing Masses?" (1992) and Madeleine Akrich's "The De-Scription of Technical Objects" (1992), both in *Shaping Technology / Building Society* (Bijker and Law, eds., MIT Press), argue that mundane artifacts embed scripts that delegate action and prescribe user behavior. The present paper agrees that artifacts carry the consequences of delegation; it argues that the moral weight of the delegation attaches to the designer who delegated, not to the artifact that carries out the delegated action. The analytical move in Section III preserves the Latour-Akrich insight while correcting the location of moral weight. Helen Nissenbaum's contextual integrity (2010) and the value-sensitive design tradition developed by Friedman and Hendry (2019) both locate moral weight at the design layer in compatible ways. Verbeek and Kudina (2018) on the Collingridge dilemma and the mediated value of privacy demonstrates how technological mediation works through the design of specific features, which is the territory on which the present argument operates.

The affordance-theoretic critique of instrumentalism, pressed in particular by Heyndels (2023), is the most direct challenge to the neutrality position this paper defends, and Section III engages it directly.

C. Where Both Positions Fail

The instrumentalist position is correct that artifacts lack moral agency. The substantivist position is correct that artifacts shape action through affordances, scripts, and embedded design history. The binary that organizes the field treats these as competing claims, but the work of Section III in this paper is to show that they are compatible claims at different levels of analysis. The artifact, considered as a physical object with intrinsic properties, lacks moral agency. The artifact, considered as a participant in a relational structure with users and intermediaries, carries the moral weight that the design decision distributed into that relational structure. Both claims survive the analytical move; what does not survive is the binary framing that forced them into competition.

The instrumentalism-substantivism debate has not adequately foregrounded the design decision as the primary locus of assignable moral accountability. The instrumentalist tradition has tended to treat the design decision as analytically subordinate to the use decision, on the view that what matters morally is what the user does with the artifact. The substantivist tradition has tended to treat the design decision as analytically subordinate to the artifact's properties, on the view that what matters morally is what the artifact carries embedded within it. Both moves obscure the design

decision as the morally weighted event that produces both the artifact and the use context. The framework developed in this paper foregrounds the design decision and treats it as the primary locus, while preserving the insights of both traditions about the artifact's physical properties and the user's intentional acts.

The substantivist objection to neutrality, in its strongest form, is not that artifacts have intentions (which would be a categorical confusion) but that artifacts carry the consequences of design decisions in ways that the instrumentalist position has not adequately theorized. The present paper accepts this objection and offers, in Section III, an analytical framework that handles the consequence-carrying without treating the artifact as the locus of moral weight. The neutrality claim defended here is therefore narrower than the version Pitt advances and is compatible with the substantivist insight that artifacts shape action. What it denies is that moral weight inheres in the artifact rather than in the relational structure that the design decision establishes.

III. Three Layers of Intent and the Defended Neutrality Claim

The neutrality claim defended here is narrower than the one Pitt advances and narrower than the one Heyndels critiques. Stating it precisely requires two distinctions: between intrinsic and relational properties of artifacts, and between three analytically separable layers of intent that can attach to any technology. Once both distinctions are in place, the affordance-theoretic objection to neutrality becomes answerable without abandoning the core claim.

A. Intrinsic Properties, Relational Properties, and the Light Bulb in a Room of Blind People

Consider a thought experiment. A room contains the brightest light bulb ever manufactured, and five people, all of whom are functionally blind. The bulb is on. Photons stream from the filament at the rated lumen output. Is the bulb bright?

The answer depends on what kind of property "brightness" is. As a physical measure, brightness is intrinsic to the bulb; the lumen output is a fact about the filament and the voltage, existing whether anyone sees it. As an experienced phenomenon, brightness is relational; it requires a perceiver with functional vision to instantiate. The blind people in the room cannot instantiate the perceptual property of brightness. The lumens exist; the experience of brightness does not.

Bring a sighted person into the same room. The photons have not changed. The bulb has not changed. But the experiential property of brightness now instantiates, and the sighted person's eyes hurt. The pain is not a property of the bulb. The pain is a property of the interaction between a particular kind of perceiver and the photons. A creature without functional vision experiences no pain from the same photon stream. The bulb does not change; it has no intent toward any of these creatures. The harm some perceivers experience in proximity to the bulb is not a property of the bulb but a property of the interaction between the bulb's emitted photons and the receiving entity's particular vulnerabilities.

This is the structure of every neutrality claim worth defending. Artifacts have intrinsic physical properties: mass, shape, material composition, photon emission, chemical reactivity. These properties exist independent of observers. Artifacts also have what is sometimes called moral weight, but moral weight is not among the intrinsic properties. It is a relational property that requires intentional agents to instantiate. When this paper calls artifacts morally neutral, it means neutral with respect to moral agency, not neutral with respect to consequences, affordances, or embedded design history.

B. The Gibsonian Affordance and Why It Does Not Defeat Neutrality

The most persistent objection to instrumentalist neutrality, developed by Heyndels (2023) and rooted in a tradition running back through Verbeek, Latour, and Norman, is the affordance argument. Artifacts have affordances. A speed

bump affords slowing down; a handle affords gripping; a handgun affords lethal action against soft tissue; a glowing screen affords passive consumption. These are not neutral facts about artifacts. They are structuring facts that channel possible action and therefore, the argument goes, carry moral weight independent of any user's intent.

The objection is correct about affordances but incorrect about what affordances are. Gibson, who introduced the concept in *The Ecological Approach to Visual Perception* (1979), was explicit on this point. Affordances are not properties of the environment alone. They are properties of the animal-environment system. Gibson's canonical example is a horizontal surface at knee height: for a human-sized creature, that surface affords sitting; for a microbe, the same surface is a continent and affords nothing of the kind; for a creature without legs, it affords no sitting either. The affordance is real and determinable, but it is determined by the joint properties of the surface and the perceiver, not by the surface alone.

This is doing the same philosophical work as the light bulb in the room of blind people. Critics of neutrality who claim that affordances make artifacts morally non-neutral are correct that affordances are real and that they shape action. They are incorrect when they slip from "real and shaping" to "intrinsic to the artifact." Gibson's own framework forbids that slip. Once affordances are understood as relational rather than intrinsic, the neutrality claim becomes coherent under the affordance objection. The artifact alone has no affordance and no moral weight. The artifact placed into a particular context, with particular users, with particular foreseeable interactions, instantiates both affordance and moral weight together. The moral weight tracks the relation, which tracks the intentional agents who structured the relation.

C. Three Layers of Intent

When critics say "this technology is not neutral," the claim usually collapses three separable questions. Sometimes the speaker means the designer intended harm; sometimes the deployer chose harmful deployment conditions; sometimes users routinely produce harm with the artifact. These are different claims about different intentional agents, and they can align, diverge, or invert across the same artifact.

The intent of the build is the question of what the artifact's designer was trying to make. What problem was the build meant to solve? What outcome was it meant to produce? The build intent is settled at the moment of design and fixed once the artifact exists.

The intent of the deployment is the question of how, where, and under what conditions an artifact is placed into the world. A given artifact can be deployed for the purpose its build intended, for an adjacent purpose, or for a purpose the build did not anticipate. Deployment is a second intentional act, often performed by a different agent than the designer.

The intent of the end use is the question of what an actual user does with an artifact once it is in their hands. End use can align with build and deployment intent, diverge from both, or invert one or both.

These three layers are analytically independent. Four cases illustrate the range.

In *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984), the Supreme Court held that the Betamax video recorder did not generate contributory infringement liability despite some users employing it to record copyrighted broadcast content. The build intent was time-shifting; the end-use intent for some users was unauthorized duplication. The Court located liability at the use layer because the technology had "substantial noninfringing uses." *Id.* at 442. The doctrine that emerged is the legal expression of the build-versus-use distinction: a technology designed for a lawful purpose is not transformed into a contraband artifact by the existence of users who put it to unlawful purposes.

Nuclear technology presents a more complex case across all three layers and across time. Szilard conceived the nuclear chain reaction in 1933 and patented it in 1934, anticipating both peaceful energy and weapons applications from the outset. Hahn and Strassmann's experimental confirmation of fission in December 1938 made both possibilities concrete within weeks. The Manhattan Project produced the first sustained chain reaction at Chicago Pile-1 in December

1942, two and a half years before Trinity in July 1945. Civilian power deployment began at EBR-I in December 1951 and commercially at Calder Hall in 1956. As of this writing, eighty-one years have elapsed since the last detonation of a nuclear weapon against a human population. The enriched uranium and plutonium that exist in current stockpiles and reactors have identical physical properties regardless of which deployment trajectory they trace. The moral weight of nuclear technology has redistributed across the deployment layer dramatically over time, while the intrinsic physical properties of the material have remained constant. No framework that locates moral weight in the artifact itself can explain this redistribution coherently; a framework that locates moral weight at the intent layers can.

Insulin offers a case where build and deployment intent align cleanly toward benefit, with occasional end-use divergence. The build intent is the prevention of diabetic ketoacidosis; the deployment intent is distribution to diabetic patients; the end-use intent is overwhelmingly medical self-administration. A small number of cases diverge: the nurse Charles Cullen, convicted in 2006 of murdering at least twenty-nine patients, used insulin among other medications as a homicide agent. The molecule does not change between licit medical use and homicidal misuse.[1]

Engagement-optimized infinite scroll inverts the Betamax pattern. The build intent is the engagement loop itself. Aza Raskin, who invented the pattern at Humanized in 2006, has stated publicly and repeatedly that he engineered the feature to exploit the variable-reward dopamine response and that he regrets having done so. The deployment intent of major platforms adopting the pattern was to maximize time-on-platform. The end-use intent of users was rarely extended compulsive consumption; that outcome was the operational mechanism of the design. All three intent layers align toward the harmful outcome, and the harm is not a downstream surprise but the engineered function of the product.

D. The Neutrality Claim, Formalized

The four cases support a formal statement. When this paper calls artifacts morally neutral, it means neutral with respect to moral agency, not neutral with respect to consequences, affordances, or embedded design history. Affordances are real and relational, following Gibson (1979). Moral agency, however, requires intent, knowledge, and culpability, all of which inhere in the intentional agents who operate at the build, deployment, and end-use layers. Moral weight attaches at the relational level, in the interaction between an artifact's intrinsic physical properties and the intentional acts of designers, deployers, and users. It does not attach to the artifact as a physical object considered apart from those relations.

IV. Design Decisions as the Locus of Moral Agency

The instrumentalism-substantivism binary fails by treating moral accountability as a choice between two poles: the artifact or the user. The framework developed in Section III introduces a third position. The locus of moral agency for foreseeable technological harm is the design decision: a choice made by an identifiable agent or organization, with knowledge or reasonable ability to acquire knowledge of probable outcomes, about how an artifact will be constructed, configured, deployed, or interfaced with users.

The design decision is not a single category but a class of related choices that share the structural feature of preceding and shaping the user's encounter with the artifact. The class includes individual designer choices that materially shape foreseeable interaction; organizational design policies governing the range of individual choices and the conditions under which the artifact is released; interface features that structure the user's encounter at the moment

[1]The insulin case is more complex at the deployment layer than the clean build-and-deployment alignment narrative suggests. Pricing decisions, patent evergreening, and supply-chain choices by pharmaceutical companies have produced foreseeable patient harm at the deployment layer, including deaths from rationing. The deployment layer is not a single decision but a continuing series of choices, each of which is analyzable under the framework developed in Section VII.

of interaction (the placement of a button, the timing of a notification, the persistence of a feed, the absence of a stopping cue); deployment conditions about where, when, to whom, and under what supervisory or institutional conditions the artifact enters the world; and metric-selection processes that determine what gets measured and optimized in iterative improvement.

The breadth of the definition has an analytical cost: it does not pre-specify which design decisions are morally weightier than others, and the proportionality analysis developed in Section VII must do work that a narrower definition would partially preempt. The cost is worth paying because artificial narrowing always tracks the political and economic interests of actors who would benefit from being excluded from accountability.

Within the class of foreseeable harms, a gradient of culpability emerges. At one end are design decisions made with specific, documented knowledge of probable harm to identifiable user populations; these satisfy the collapse threshold developed in Section VII. At the other end are decisions made with general awareness that some users might experience some adverse outcomes, without specific knowledge tied to specific design choices; these remain within the distributed-responsibility model with the designer's contribution proportional to the design decision's actual structural role in the harm.

A. The Concept of Incrementalism

Many of the design decisions that produce significant aggregate harm at platform scale are individually small, individually justified, and individually defensible. The persistent push notification was justified as user engagement. The autoplay feature was justified as user convenience. The algorithmic amplification of high-engagement content was justified as user preference satisfaction. The infinite scroll was justified as friction reduction. Each choice, evaluated in isolation, can be defended on terms that any competent engineering review would accept.

The aggregate effect is a different question. The persistent push notification combined with autoplay combined with algorithmic amplification combined with infinite scroll combined with the absence of stopping cues produces a system that, considered as a whole, structures the user's interaction in ways that no individual designer of any of the components specifically intended. The phenomenon is sometimes called emergent design or composition harm. The relevant analytical question is whether incremental design decisions that produce aggregate harm relocate the moral weight to the level of the organization that approved the composition, or whether the weight remains with the individual designers who made the component decisions.

The framework's answer is that the moral weight follows the agency, capacity, and information of the deciding actor. Individual designers of components often lack the information and capacity to assess aggregate effects, and the agency to refuse the composition. Organizational leadership and product management actors have both the agency and the institutional position to assess aggregate effects, and in many cases the information from internal research and user feedback that documents the aggregate outcomes. The moral weight of composition harms therefore concentrates at the organizational design layer where the composition decisions are made, even where individual component decisions are individually defensible. This is one of the patterns the collapse threshold in Section VII captures: specific organizational knowledge of aggregate harm, combined with affirmative organizational choice to continue the composition, concentrates responsibility at the organizational level despite the dispersion of component decisions across individual designers.

V. A PRE-INDUSTRIAL GENEALOGY OF MAKER ACCOUNTABILITY

This section turns from analytical structure to historical grounding. The question is whether the framework developed above describes a recent philosophical innovation or recovers a pattern that pre-industrial legal and institutional traditions

already implemented. The cases below are offered as illustrative genealogy rather than universal proof. They do not establish that every pre-industrial civilization handled technological harm in the same way; many did not. What they establish is that maker accountability and intermediary responsibility were operative and codified across distinct legal traditions at widely separated times, which forecloses the argument that locating moral weight at the design decision is a category error imposed by modern values on technologies whose ancient predecessors were governed entirely by user-side caveat.

A. Hammurabi and the Builder's Liability (c. 1750 BCE)

The Code of Hammurabi contains the oldest extant legal codification of maker liability for foreseeable construction harm. Sections 229 through 233 attach criminal and civil liability directly to the builder for collapse of structures the builder erected. Section 229 specifies that if a builder constructs a house and the construction is not sound, with the result that the house collapses and kills the owner, the builder shall be put to death. Section 230 extends the principle to the death of the owner's son. The standard is one of foreseeable harm flowing from inadequate construction; the code does not require proof of subjective malice.

The code's structure is doctrinally striking. The harm originates in a building; the building is an artifact; the artifact is neither punished nor the subject of inquiry. The user of the artifact, the owner who lived in the house, is also not the locus of accountability. The locus is the builder, the agent who made the construction decisions. In the analytical vocabulary developed in Section III, the Code of Hammurabi codifies accountability at the build layer with strict-liability force.

The historical context matters. Babylonian construction in the eighteenth century BCE relied on mud brick and timber, with builders operating in an information asymmetry vis-a-vis their clients. The client could not inspect the wall thicknesses or foundation depth once the structure was complete. The code's response was to attach severe consequences to foreseeable failures, creating a structural incentive for sound design choices despite the impossibility of contemporaneous user-side verification. This is the same regulatory pattern that modern strict products liability implements in response to similar information asymmetries.

B. The Spartan Agoge as Communal Intermediary Responsibility (7th c. BCE)

The Spartan agoge provides a different and complementary case. The agoge is not best understood as direct evidence for maker accountability in the construction-liability sense; it is better understood as evidence that responsibility for foreseeable harm to citizens was distributed across multiple intermediary actors, including parents, the paidonomos, the Ephors, and the broader citizen body. The agoge is therefore evidence for the distributed-responsibility model developed in Section VII rather than for the simpler maker-liability model that Hammurabi codifies. The two cases are making related but distinct genealogical claims; the genealogy as a whole gains strength from showing that both patterns were operative in pre-industrial systems.

The agoge's accountability structure is documented in Xenophon's *Constitution of the Lacedaemonians*, Plutarch's *Life of Lycurgus*, and Aristotle's *Politics*, with synthesis in Cartledge's modern scholarship. The system removed boys from parental households at age seven and placed them under state supervision through approximately age twenty. The harm pattern relevant here is the foreseeable physical and psychological injury that the training inflicted on participants, including the systematic flogging at the altar of Artemis Orthia and the krypteia, the secret police function. These were not accidents of an otherwise benign system; they were designed features.

What distinguishes the Spartan case is that no single designer can be identified as the locus of responsibility for the agoge's harms. The system was attributed to the semi-legendary Lycurgus, but actual operation was distributed across

the offices named above, with parents retaining residual responsibility before age seven and the broader citizen body bearing accountability for sustaining or reforming the system across generations. The moral weight depended on each actor's agency, information, and capacity to act.

C. Medieval European Guilds (1000-1500 CE)

The European craft guilds that operated between approximately 1000 and 1500 CE provide the closest pre-industrial analogue to the modern accountability framework developed in this paper. Guild structure attached identifiable maker accountability to physical artifacts through three institutional mechanisms: required maker's marks on finished goods, internal inspection and quality enforcement, and expulsion for defective production. Ogilvie's (2011) study documents the consistency of these mechanisms across Germany, England, France, and Italy.

The maker's mark transformed anonymous artifacts into traceable products of identifiable makers. A goldsmith's plate, a weaver's bolt of cloth, a cooper's barrel, and a swordsmith's blade each carried a punch or stamp identifying the responsible guild member. The mark functioned as the medieval equivalent of a manufacturer's identifier. Defective goods could be traced back to their maker, and complaints could be lodged with the guild's wardens. The inspection function ran in two directions: wardens inspected raw materials and finished goods at point of sale. Substandard production triggered guild discipline ranging from fines and confiscation to expulsion, which in the medieval economy was equivalent to disqualification from the trade.

The genealogy supports a bounded historical claim. Across three distinct legal and institutional traditions, separated by millennia and continents, accountability for foreseeable harm flowing from designed artifacts attached either to identifiable makers (Hammurabi, guilds) or to identifiable intermediary actors at multiple institutional levels (Sparta). In none of these cases did accountability attach exclusively to users on a caveat-emptor model. In none of these cases was the artifact itself treated as the locus of moral weight.

What the genealogy does not show is equally important. It does not show that all pre-industrial civilizations governed technological harm in this way, or that the design-accountability and intermediary-responsibility patterns were uniformly applied across all artifact classes within the civilizations discussed. Roman law, in particular, developed an extensive doctrine of *culpa* and *damnum injuria datum* that operated quite differently from the Babylonian strict-liability model, distributing responsibility through fault-based assessment rather than through the categorical maker liability that Hammurabi codified. Pre-industrial polities in the Americas, Asia, and Africa developed their own approaches that this paper does not survey. The genealogy is therefore not a proof of historical inevitability. It is a demonstration that the analytical framework developed in earlier sections describes a pattern that was available to and operative in pre-industrial legal systems, which forecloses the rhetorical move of treating the framework as a recent or culturally narrow imposition on ancient practice.

The relevance of this for the paper's argument is structural. The capitalist rupture treated in Section VI represents not the historical default of how technological harm has been governed but a comparatively recent deviation from a longer pattern in which maker accountability and intermediary responsibility were available institutional tools. Recovering those tools, through the distributed-responsibility framework developed in Section VII, is therefore not a radical reorganization of moral responsibility but a return to patterns that earlier civilizations recognized as appropriate responses to the structural features of designed artifacts.

VI. The Capitalist Rupture and Its Legal Architecture

If the pre-industrial pattern attached accountability to makers and intermediaries for foreseeable harm, the modern pattern is structured very differently. The shift was the product of specific legal and economic choices made between the

seventeenth and twentieth centuries that systematically relocated accountability away from designers and toward users.

A. Caveat Emptor and Privity of Contract

The doctrine of caveat emptor emerged in English common law during the seventeenth and eighteenth centuries as commercial transactions scaled beyond local face-to-face arrangements. The doctrine shifted the burden of investigation from seller to buyer. Where the guild model presumed that the maker bore accountability for the soundness of the goods and that buyers could rely on the maker's mark and guild oversight, the caveat-emptor model presumed that buyers were responsible for ascertaining quality at the point of sale, with no continuing obligation flowing from the seller to the buyer once the transaction was complete.

The doctrine was not invented to address technological harm in any modern sense. It addressed the practical problem of governing commercial transactions in an expanding market where personal acquaintance between buyer and seller could no longer be assumed. As Emroch (1970) documents, the doctrine's spread tracked the rise of commercial markets that outran the local guild-and-warden infrastructure described in Section V. The functional consequence, however, was the inversion of the accountability vector that pre-industrial legal traditions had established. The maker who had been accountable under Hammurabi and the guild system now had no obligation to the buyer beyond the bare terms of the immediate sale. The information asymmetry that the guild's inspection function had previously addressed was now reframed as a problem for the buyer to solve.

The privity doctrine, which crystallized in English law during the nineteenth century and propagated to American jurisdictions through the latter half of that century, further insulated designers from end-users by limiting contractual obligations to the parties immediately in contractual relationship. A manufacturer who sold to a wholesaler, who sold to a retailer, who sold to a consumer, owed nothing to the consumer because the manufacturer was not in direct contractual privity with the consumer. The doctrine functioned as a legal firewall between the design decision and the eventual injured party.

The most famous early application of privity to a manufactured-product injury case is *Winterbottom v. Wright* (1842), in which an English court held that a coach driver injured by the failure of a mail coach could not recover from the coach's manufacturer because the manufacturer had contracted with the postmaster general, not with the driver. The court's reasoning was explicit: extending liability beyond contractual privity would, in the court's view, produce an indefinitely large class of potential plaintiffs and intolerable exposure for manufacturers. The decision became the leading nineteenth-century authority for the proposition that designers and manufacturers owed nothing to remote users injured by foreseeable defects in the products they had designed. Privity, combined with caveat emptor, completed the inversion. The designer was now legally insulated from the end-user along two dimensions simultaneously: there was no continuing obligation flowing from the sale, and there was no path for the end-user to reach the original designer through legal action.

B. Product Liability as Partial Reconstruction

The reconstruction of design accountability began in the early twentieth century. In *MacPherson v. Buick Motor Co.*, 217 N.Y. 382 (1916), Judge Cardozo held that an automobile manufacturer owed a duty of care to the eventual user despite the absence of direct contractual privity, where the product was such that it would cause harm if negligently made. The reasoning was that foreseeability of harm and the manufacturer's superior position with respect to design and inspection were sufficient to ground a duty running from manufacturer to foreseeable end-user.

The second stage was the emergence of strict products liability. In *Greenman v. Yuba Power Products, Inc.*, 59 Cal. 2d 57 (1963), Justice Traynor held that a manufacturer who places a product on the market knowing it will be used

without inspection is strictly liable in tort for injuries caused by defects. The doctrine was codified in section 402A of the Restatement (Second) of Torts (1965) and propagated through state courts during the 1960s and 1970s. The reconstruction did not restore the pre-industrial pattern in full; it operates primarily through private civil litigation rather than through institutional inspection and enforcement, which means injured users must initiate proceedings and bear costs, often against well-resourced corporate defendants.

C. Section 230 and the Digital Caveat Emptor

Section 230 of the Communications Decency Act, enacted in 1996, established that providers of interactive computer services would not be treated as publishers of information provided by other content providers. The provision was a deliberate legislative choice made in the context of mid-1990s concerns about the development of the early internet, and it reflected the policy judgment that imposing publisher-level liability would have destroyed the emerging industry. As an original-intent matter, Section 230 was designed to address the publisher-versus-distributor liability question that had arisen in cases such as *Stratton Oakmont, Inc. v. Prodigy Services Co.* (1995).

The doctrine's subsequent application extended far beyond the original publisher-liability question. Courts have repeatedly applied Section 230 to insulate platforms from claims arising not from third-party content but from platform design decisions. The reasoning typically characterizes the challenged platform conduct as editorial activity and therefore within the statutory shield. The cumulative effect has been to provide platform designers with a level of legal insulation that resembles, in functional terms, the privity barrier that *MacPherson* and *Greenman* spent the twentieth century dismantling for tangible products.

The legal landscape is now moving. JCCP 5255 in California, the coordinated state-court proceeding consolidating thousands of social-media addiction claims, and the parallel federal MDL No. 3047 in the Northern District of California, both advance theories that target platform conduct as product design rather than as publication of third-party content. The Ninth Circuit has continued to develop the doctrinal contour of when design-based claims fall inside or outside the Section 230 shield, most recently in *Calise v. Meta Platforms, Inc.*, 103 F.4th 732 (9th Cir. 2024), and *Estate of Bride v. Yolo Technologies, Inc.*, 112 F.4th 1168 (9th Cir. 2024). In March 2026 a Los Angeles jury in the first JCCP 5255 bellwether, *K.G.M. v. Meta Platforms, Inc.*, found Meta and Google's YouTube liable for negligent platform design and failure to warn, awarding $6 million in combined compensatory and punitive damages. The doctrinal question is whether the design-defect framing can survive Section 230 scrutiny by characterizing harm as flowing from the platform's product design (the algorithmic amplification system, engagement-optimization features, absence of break prompts, persistent push notifications) rather than from the content the platform hosts. The stronger formulation of the Section 230 critique is therefore not that it categorically shields all platform design decisions but that it has functioned as a design-accountability shield wherever courts characterize challenged platform conduct as publishing activity, and that recent verdicts suggest design-defect theories are beginning to survive that characterization.

D. The Cumulative Pattern

Caveat emptor, privity, the partial reconstruction of strict products liability, and the Section 230 carve-out describe an arc rather than a single event. The arc moved the legal location of accountability away from designers and toward users between the seventeenth and nineteenth centuries; partially reversed direction during the twentieth century for tangible products through the products-liability doctrine; and then extended the older insulation pattern into the digital domain through the Section 230 carve-out beginning in 1996. The current legal moment is one of unstable equilibrium, with design-defect litigation against platform architecture pushing toward a partial reconstruction analogous to what *MacPherson* and *Greenman* accomplished for tangible products, and platform defendants pushing back through Section 230 and First Amendment defenses. The framework developed in this paper takes no position on the outcome of that

legal contest. It does take the position that the underlying moral structure (which Section VII develops in distributed-responsibility terms) is intact regardless of which way the legal doctrine resolves, and that the accountability pattern the doctrine ultimately implements ought to track the moral structure rather than the inverse.

VII. The Distributed Responsibility Framework

The legal architecture traced in Section VI produced a doctrinal vocabulary that often treats accountability as a binary question: either the designer is fully liable or the user assumed the risk; either Section 230 shields the platform entirely or the platform is fully exposed. This binary framing is doctrinally awkward, philosophically inadequate, and historically aberrant. The framework developed in this section returns to the distributed pattern that Section V documented.

A. The Four-Actor Model

Responsibility for foreseeable harm flowing from a designed artifact distributes across four actors, with the proportional weight depending on the agency, knowledge, capacity, and institutional position of each.

The artifact bears no responsibility in the moral sense developed in Section III; it has no intent, no knowledge, no capacity for action. The four-actor model includes the artifact as a category not because the artifact bears responsibility but because the artifact's physical properties are the substrate that constrains and enables the other three actors' choices.

The designer is the agent whose decisions about construction, configuration, and intended use shape the field of foreseeable interactions. The designer's responsibility is heavy when the designer possesses superior knowledge about the artifact's effects, when the design decisions structure the user's possible actions in significant ways, and when the designer has the capacity to make different choices.

The user is the agent who interacts with the artifact in the field of possibilities the design decisions established. The user's responsibility tracks agency, capacity, information, and developmental status. An adult user with full information about an artifact's effects, full capacity to use it differently, and no developmental constraints on judgment bears greater responsibility than a child user lacking those features.

The intermediary is the agent positioned between designer and user, with some capacity to mediate the interaction, some access to information about both sides, and some institutional responsibility for the relationship. Parents, schools, employers, regulators, and platforms in their content-distribution function are all intermediaries in different contexts.

The three-layer intent structure developed in Section III maps onto the four-actor model as follows. The designer operates primarily at the build layer, with some involvement at the deployment layer where build and deployment decisions are made by the same organization. The user operates at the end-use layer. The intermediary operates principally at the deployment layer (parents deploying an artifact to a child, schools deploying an artifact to a classroom, regulators conditioning the deployment of an artifact to a market) and may have some residual involvement at the end-use layer through ongoing supervisory functions. The artifact operates at none of the three layers in the morally weighted sense because it has no intent; it is the substrate through which the three layers express themselves.

The model is not a mechanical algorithm. It is an analytical structure that resists collapse into binary blame. When a harm occurs, the question is not whether the designer is responsible or the user is responsible but what proportion of responsibility attaches to each of the four actors given their relative agency, capacity, information, and institutional position.

B. Proportionality

Responsibility under the four-actor model is proportional rather than categorical. The proportionality runs along five dimensions that interact rather than aggregate cleanly.

Agency: how much capacity did each actor have to make different choices? An infant has minimal agency; an adolescent has substantial but developmentally limited agency; a fully competent adult has high agency.

Capacity: what resources did each actor have to act on available choices? A parent with twenty other obligations and limited technological expertise has lower capacity than a regulator with statutory authority and a staff of investigators.

Information: what did each actor know or have reason to know about the foreseeable outcomes? A designer with internal research documenting harm has high information; a user with no exposure to such research and no reason to suspect harm has low information.

Developmental status: for human actors, what is the actor's stage of cognitive and emotional development? A six-year-old, a thirteen-year-old, and a thirty-year-old face structurally different proportionality calculations even when the other dimensions are roughly similar. The empirical literature on adolescent brain development (Crone and Konijn, 2018) establishes that adolescents are not adults in miniature and that responsibility assigned to adolescent users must reflect their developmental status as a distinct factor.

Institutional position: what institutional or structural role did each actor occupy with respect to the harm? A regulator's position is different from a platform's position is different from a parent's position is different from a child's position, even where the other dimensions are similar.

These five dimensions are not weights to be summed; they are analytical lenses through which responsibility is assessed. The result is a description of the distribution of responsibility across the four actors, which then bears on questions of legal liability, policy design, and moral judgment.

C. The Collapse Threshold

The four-actor model resists binary collapse in most cases. There is, however, a particular structural condition under which the distributed responsibility model concentrates heavily on a single actor: when one actor possesses specific knowledge of foreseeable harm and proceeds with affirmative choice anyway, the proportionality calculation shifts toward concentration on that actor, even when other actors retain some residual responsibility.

This is the collapse threshold. The structural logic is that the distributed model presumes a baseline of reasonable diligence and reasonable information across actors; when one actor possesses information that materially exceeds the baseline and acts in ways that produce the foreseeable harm anyway, the moral architecture cannot continue to treat that actor's responsibility as proportionally similar to actors who lacked the same information.

The threshold raises a genuine epistemological problem that the framework identifies but does not fully solve: how specific must the knowledge be? Internal research showing a statistical correlation between platform use and adolescent self-harm ideation is specific. Anecdotal reports from customer service teams are less specific. A product manager's vague sense that "something is wrong" is less specific still. The framework's working answer is that the knowledge condition is satisfied when the relevant actor possesses information that a reasonable person in the actor's position would recognize as predicting harm to a defined population at a magnitude that materially exceeds normal background risk, with the assessment performed against the information available at the time of the relevant choice rather than retrospectively. This is admittedly imperfect; the operationalization of the standard is one of the directions in which the framework will need further development.

The People v. Crumbley convictions (Michigan, 2024), in which parents were convicted of involuntary manslaughter

for failing to act on specific knowledge of their son's foreseeable harmful intent in advance of a 2021 school shooting, are structurally analogous though doctrinally distinct. Structurally, both situations involve an actor with specific knowledge of foreseeable harm and an affirmative failure to act in ways that would have mitigated the harm. The structural similarity supports the analytical claim that specific knowledge plus affirmative inaction satisfies a moral threshold above the baseline of distributed responsibility. Doctrinally, the cases differ in nearly every respect: Crumbley involved criminal parental negligence with individualized, immediate knowledge of an imminent threat; the platform cases involve corporate civil liability over years of population-level statistical evidence. The structural moral logic survives the doctrinal differences; the structural moral logic does not collapse them.

The collapse threshold is offered as an analytical tool, not as a legal rule. Its function is to identify a particular pattern within the distributed responsibility model and to inform the framing of policy debates and legislative reasoning, not to direct any specific doctrinal outcome.

D. The Misattribution Cycle

The four-actor model and the collapse threshold together explain why the public discourse around technological harm so often misattributes responsibility. The misattribution operates in two directions.

In the first direction, the harm is attributed to the artifact itself. The screen is bad. The phone is addictive. The app is dangerous. This is the substantivist move that Section II critiqued. It is rhetorically common because the artifact is the visible, tangible element of the harmful interaction, and because attributing responsibility to a thing is easier than working through the responsibility distribution across the human actors who produced and deployed it. The artifact, however, has no intent and no agency. The artifact cannot be held responsible in any meaningful moral sense. The substantivist attribution produces policy responses that miss the locus of the actual harm (the design decision) and that often produce results that further insulate the actual responsible actors.

In the second direction, the harm is attributed to the user, especially when the user is a child or adolescent. The parent should have monitored. The teenager should have shown more self-control. The teacher should have intervened. This is the user-blame move that emerges from the caveat-emptor and privity-shielded structure described in Section VI. The user-blame attribution is rhetorically attractive because it preserves the moral universe in which individual responsibility is the default mode of accountability, and because attributing responsibility to identifiable individuals is psychologically simpler than attributing responsibility to corporate decisions made by anonymous designers in distant cities. The user-blame attribution produces policy responses that focus on user education, parental monitoring, and individual restraint, while leaving the design decisions that structured the harmful interaction untouched.

Both misattributions share a common structural defect: they treat the design decision as if it were not an event in the world that produced the harmful interaction. The four-actor model corrects the misattribution by insisting that the design decision is the event that structured the interaction, that the designer is the actor who made that decision, and that responsibility distributes across the four actors with proportional weight rather than concentrating at either of the two poles where the misattributions land.

VIII. Digital Dementia as a Case Study in Public Misframing

The misattribution patterns developed in Section VII are visible in current public discourse around adolescent digital harm, with the "digital dementia" framing serving as a representative example. The term entered Anglophone discourse through Spitzer's 2012 book and propagated through subsequent media coverage. The recent 60 Minutes Australia segment is a representative example of how the framing reaches mass audiences, presenting adolescent cognitive and attentional difficulties as a neurological consequence of screen exposure with the rhetorical force of a comparison to

neurodegenerative disease.

The clinical content of the framing does not survive scrutiny. Dementia is a clinical category that describes progressive, generally irreversible neurodegenerative conditions characterized by specific pathological markers. The cognitive and attentional difficulties that adolescent heavy users of digital platforms experience are not characterized by these pathological markers and are largely reversible with changes in use patterns. The relevant empirical literature, including Crone and Konijn (2018) and Orben and Przybylski (2019), describes attention, sleep, and mood effects that are statistically detectable but not analogous to irreversible neurodegenerative process. The slip from "reversible cognitive effects" to "dementia" is a category error. The error is rhetorically productive because it activates the public association with Alzheimer's disease and similar conditions; the price of the productivity is that the framing misdirects attention from the actual causal structure of the harm.

The framing locates the harm in the screen, the phone, the app, or technology in the abstract, activating the substantivist intuition that the technology itself is the cause. The policy responses that emerge focus on time limits, screen bans, and individual restraint. Australia's recent under-sixteen social media restriction is one example of a policy response calibrated to the substantivist framing. Notably, however, the Australian regulatory text itself is more sophisticated than the public framing it responds to: the official definition explicitly references recommender algorithms, endless feeds and infinite scroll, likes and upvotes, and time-limited stories. The Australian framework, in other words, names platform design features as the regulatory target while implementing an age-based remedy. The substantivist remedy is calibrated to the misframing, but the regulatory text shows that policymakers are partially aware that the harm tracks design architecture rather than screens as such.

Two analytical tests sharpen the distinction. The *nature-documentary test* considers an adolescent who watches nature documentaries on a phone screen for two hours each night: the screen is engaged, hours accumulate, blue light is delivered to the retina, and none of the alleged digital-dementia symptoms develop. The test isolates the screen from the design architecture; the harm tracks the design architecture, not the screen. The *office-worker test*, in its narrower form, considers an adult office worker exposed to a high-resolution monitor for eight hours per day across decades. The office worker experiences eye strain, ergonomic problems, and sleep disruption from late-evening use, but not the specific cluster of effects that the digital dementia framing attributes to adolescent platform users. The test isolates time-on-screen from design-on-screen; if the harm were a property of screen exposure as such, the office worker would experience the same pattern. They do not.

The point of this section is not to dismiss the empirical literature on adolescent platform effects. The literature is real and growing, and recent litigation rests on evidence that the engineering choices the platforms made were associated with adolescent harm in ways the platforms' internal research had identified. The point is that the public translation of that literature into the digital dementia framing represents a structural misattribution that the framework developed here is well-positioned to diagnose. The framework directs attention to the design decisions that produced the harm-associated features, distributes responsibility across the four actors with the proportionality the framework specifies, and identifies the collapse threshold at which the designer's responsibility concentrates due to specific knowledge of harm and affirmative choice to proceed.

IX. The Boundary Condition

The framework developed in this paper applies to artifacts whose interaction with users is structured at the population level by design decisions made in advance of any individual user's encounter with the artifact. A speed bump, a handgun, a social media platform's recommendation algorithm, an infinite-scroll interface, and a video game's level design all share this structural feature: the design decision precedes the user interaction and shapes it across the entire population

of users in ways that are determinable from the design itself rather than from the particular history of each user. The four-actor model and the collapse threshold are tools for analyzing this class of artifacts.

A different class of artifacts is now emerging that does not share this structural feature. Adaptive AI systems that engage in individualized interaction with users produce harm patterns that are not fully determinable from design decisions made in advance, because the interaction itself is co-produced by the user and the system in ways that depend on the unique history of their particular exchanges. A large language model deployed as a conversational companion, a recommendation system that has learned a particular user's preferences across thousands of interactions, and an emotional-support AI that responds to a user's specific affective patterns all share the feature that the design decision does not fully structure the interaction in advance.

A. Why the Framework Stops

The four-actor model presumes that the design decision is an event in the world that structures the interaction between the artifact and a population of users. The proportionality calculation in Section VII depends on this presumption, because the distribution of responsibility across designer, user, intermediary, and (categorically) artifact depends on each actor's contribution to the structural conditions that produced the harm. The collapse threshold likewise depends on the presumption that specific knowledge of foreseeable harm can be tied to specific design decisions, with the corresponding responsibility concentration.

In adaptive AI interaction, these presumptions weaken. The "design decision" is no longer a discrete event but a distribution of decisions that includes the training procedure, the alignment work, the deployment configuration, the user's interaction history, and the system's adaptive response to that history. The "foreseeable harm" is no longer fully determinable from the design decisions because the harm pattern emerges from the particular interaction history of a particular user rather than from the design alone. The actors at the table for the proportionality calculation now include not just the four of the present model but additional categories that the framework does not adequately handle: the system itself in some emerging moral sense, the prior users whose interactions shaped the system's adaptive trajectory, and the institutional actors who deployed the system into the relational context where the harm occurred.

B. The Handoff to Relational Ethics

The relational and ontological questions that arise at this boundary, particularly whether adaptive AI interaction places systems outside the category of "thing" and into a relational moral status that resists both the agent and patient categories of Western ethics, are not deferred because they are unimportant. They are deferred because they constitute a distinct line of inquiry, one that Gunkel (2012, 2018) has been developing for over a decade and that requires its own paper to engage adequately. The contribution of the present paper is to identify *where* that line of inquiry begins, which is at the moment a technology's interaction with a user becomes individualized and relational rather than population-level and architectural.

Gunkel's *Machine Question* (2012) and *Robot Rights* (2018) have built the conceptual infrastructure for thinking about moral status under conditions where the traditional agent-patient categories no longer organize the phenomena. The relational turn that Gunkel develops, drawing on Levinas and other Continental sources, treats moral status not as a metaphysical property that an entity either has or lacks but as a relation that emerges in the encounter between entities. For artifacts whose interaction with users is individualized and relational, this framing offers analytical resources that the present paper's framework, designed for population-level design accountability, does not provide. The scope condition is offered not as a confession of weakness but as a methodological commitment. The framework's value is the cleanness with which it analyzes the cases within its scope; extending it into territory where its presumptions weaken

would produce worse analyses on both sides of the gradient.

X. Future Directions

The framework developed in this paper extends naturally into three areas that subsequent work will develop more fully.

A. Three Layers of Intent Applied to Dual-Use Technologies

The intent decomposition introduced in Section III supports a fuller analysis of dual-use technologies, medical innovations with potential for misuse, and content created within engagement-optimized platforms. The three intent layers can align, diverge, or invert in patterns that produce distinct moral analyses. *Sony Corp. v. Universal City Studios* (1984) illustrates the Betamax pattern where build intent (time-shifting) diverges from end-use intent (piracy in some cases) and the Court located liability at the use layer. Engagement-optimized infinite scroll inverts the pattern: the build intent is the engagement loop itself, and the harm is the operational mechanism rather than a downstream surprise. Nuclear technology presents a third pattern: dual-intent at conception, weapon at first deployment, civilian energy across subsequent decades, where the same physical material (enriched uranium) carries identical properties throughout while moral weight redistributes across deployments. Insulin and other dual-use pharmaceuticals present a fourth pattern, with clean alignment at the build and deployment layers and rare divergence at the use layer.

A forthcoming companion paper develops the three-layer model as a tool for evaluating cases where build, deployment, and end-use intent diverge significantly across the same artifact, with particular attention to weapons technology, pharmaceuticals, and security research. The three-layer model also responds to a longstanding affordance-theoretic objection to neutrality claims by clarifying that affordances are relational properties of artifact-user systems (following Gibson, 1979) rather than intrinsic properties of artifacts. The neutrality thesis defended in the present paper is therefore narrower than its instrumentalist precursors and more defensible against affordance-based critiques.

B. Recursive Design Accountability in Algorithmically Amplified Content

The four-actor model presented in this paper handles platform design accountability cleanly when content and platform are analytically distinct: the platform designer is accountable for the architectural choices that shape user interaction, while content creators retain their own moral standing as producers of speech, art, or commerce. This framing is sufficient for many cases but becomes inadequate when content creators understand the platform's amplification mechanisms and engineer their content to exploit those mechanisms.

A content creator producing rage-bait designed to trigger algorithmic amplification, an influencer engineering thumbnails and openings to maximize click-through, or an account producing content that exploits documented vulnerabilities in adolescent attention systems is not merely creating content that happens to spread. They are co-designing with the platform's architecture, leveraging the platform's design choices to produce harm at scale. In these cases, the moral weight does not concentrate on a single designer; it distributes between two designers (platform architecture and content engineering) with full weight at both layers simultaneously.

A forthcoming paper develops a recursive design accountability model that maps the three-layer intent stack onto both the platform and the content within it, identifying the conditions under which content creator and platform jointly satisfy the collapse threshold. The model also clarifies cases of asymmetric intent, where benign content gets weaponized by platform amplification against the creator's intent, and the inverse case where malicious content gets deamplified by platform safety measures.

C. Harm Blindness as a Constraint on Deployment Decisions

The collapse threshold identified in this paper, where distributed responsibility concentrates on a single actor upon documented foreknowledge of harm, points toward a broader application: any deployment decision that imposes harm on populations who were not represented at the decision table and have no mechanism for surfacing concerns. The Harm Blindness Framework develops this insight as a procedural rather than consequentialist constraint on design and deployment decisions.

The clearest test case is anti-personnel landmine deployment. Even within a just war framework, where the deploying force defends against an unjust aggressor, the moral analysis cannot terminate at the moment of deployment. Mines persist. Records get lost. The civilian population that steps on those mines decades later, including the deploying force's own future citizens, was not part of the deployment decision and had no feedback mechanism. The Ottawa Treaty (1997) functionally implements this intuition by banning anti-personnel mines on humanitarian grounds, but the underlying moral structure is procedural: the harm imposed on unrepresented future populations cannot be licensed by the deploying force's contemporary moral framing. A forthcoming paper extends this analysis to persistent infrastructure with multi-decade harm signatures, technologies released without sunset clauses or accountability provisions, and AI systems deployed at scale before their downstream effects can be assessed.

XI. Conclusion

The argument advanced in this paper has been that the instrumentalism-substantivism binary in the philosophy of technology has organized analytical attention around the wrong question. The right question is not whether moral accountability for foreseeable technological harm attaches to the artifact or to the user. The right question is how moral accountability distributes across designers, users, intermediaries, and (categorically) artifacts, with proportional weight tracking agency, capacity, information, developmental status, and institutional position, and with the design decision foregrounded as the primary locus where moral weight concentrates.

The framework supports this analytical move through three connected commitments. First, the neutrality of artifacts in the limited sense of moral agency is preserved, with the affordance-theoretic critique answered through the distinction between intrinsic and relational properties and through the Gibsonian reframing of affordance. Second, the historical genealogy of maker accountability across three pre-industrial traditions establishes that the framework recovers a pattern available to earlier legal and institutional systems rather than imposing a recent innovation on ancient practice. Third, the four-actor distributed responsibility model with its collapse threshold provides a tool for assigning proportional responsibility under the structural conditions that designed artifacts produce.

Return to the table in the hallway. The father's sarcastic question now has a defended philosophical answer. The person who decided to place a hard-cornered object at hip height in a corridor where people move quickly and do not look down made a design decision with a foreseeable outcome. The decision is the locus of moral weight. The table itself is morally neutral; the child bears developmentally proportional responsibility; the placement decision carries the weight that the binary framing had been unable to locate. The same analysis scales from the table in the hallway to the platform deployed to three billion users. What changes at scale is not the structure of the analysis but the magnitude of the foreseeable consequences and the institutional complexity of the actors involved. The framework defended in this paper is the analytical tool for performing the analysis at any scale within its scope.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Akrich, M. (1992). The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology / building society: Studies in sociotechnical change* (pp. 205–224). MIT Press.

Cartledge, P. (2002). *Sparta and Lakonia: A regional history 1300–362 BC* (2nd ed.). Routledge.

Communications Decency Act of 1996, 47 U.S.C. §230.

Crone, E. A., & Konijn, E. A. (2018). Media use and brain development during adolescence. *Nature Communications*, 9(1), 588. https://doi.org/10.1038/s41467-018-03126-x

Emroch, B. M. (1970). Implied warranty and strict liability: A study of the development of the doctrines. *University of Richmond Law Review*, 5(1), 25–48.

Feenberg, A. (1999). *Questioning technology*. Routledge.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Greenman v. Yuba Power Products, Inc., 59 Cal. 2d 57 (1963).

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Heyndels, S. (2023). The non-neutrality of technology: Pitt, Heidegger, and the philosophy of artifacts. *Philosophy & Technology*, 36(2), 1–19.

Calise v. Meta Platforms, Inc., 103 F.4th 732 (9th Cir. 2024).

Estate of Bride v. Yolo Technologies, Inc., 112 F.4th 1168 (9th Cir. 2024).

In re Social Media Adolescent Addiction/Personal Injury Products Liability Litigation, MDL No. 3047, 4:22-md-03047-YGR (N.D. Cal.) (consolidated proceedings, before Hon. Yvonne Gonzalez Rogers).

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology / building society: Studies in sociotechnical change* (pp. 225–258). MIT Press.

MacPherson v. Buick Motor Co., 217 N.Y. 382 (1916).

Mine Ban Treaty (Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction), opened for signature December 3, 1997, 2056 U.N.T.S. 211.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Ogilvie, S. (2011). *Institutions and European trade: Merchant guilds, 1000–1800*. Cambridge University Press.

Orben, A., & Przybylski, A. K. (2019). Screens, teens, and psychological well-being: Evidence from three time-use-diary studies. *Psychological Science*, 30(5), 682–696. https://doi.org/10.1177/0956797619830329

K.G.M. v. Meta Platforms, Inc., et al., JCCP No. 5255 (L.A. Super. Ct., verdict March 25, 2026) (bellwether trial in Social Media Cases, JCCP 5255, before Hon. Carolyn B. Kuhl).

People v. Crumbley (Jennifer), No. 2021-279030-FC (Oakland Cty. Cir. Ct. Mich. 2024).

People v. Crumbley (James), No. 2021-279031-FC (Oakland Cty. Cir. Ct. Mich. 2024).

Pitt, J. C. (2014). *Doing philosophy of technology: Essays in a pragmatist spirit*. Springer.

Plutarch. (1914). Life of Lycurgus. In *Plutarch's lives* (Vol. 1, B. Perrin, Trans.). Loeb Classical Library.

Restatement (Second) of Torts §402A (1965).

Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

Spitzer, M. (2012). *Digitale Demenz: Wie wir uns und unsere Kinder um den Verstand bringen*. Droemer.

Stratton Oakmont, Inc. v. Prodigy Services Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Pennsylvania State University Press.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Verbeek, P.-P., & Kudina, O. (2018). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, 44(2), 291–314. https://doi.org/10.1177/0162243918793711

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.

Winterbottom v. Wright, 152 Eng. Rep. 402 (Ex. 1842).

Xenophon. (1925). Constitution of the Lacedaemonians. In *Xenophon: Scripta minora* (E. C. Marchant, Trans.). Loeb Classical Library.