knowledge

MDPI

Article

Machine Understanding of Harms: Theory and Implementation

Joseph Jebari [1,2] **and Ariel M. Greenberg** [3,*]

[1] Georgetown-Howard Center for Medical Humanities and Health Justice, Washington, DC 20001, USA; joseph.jebari@howard.edu
[2] Department of Philosophy, Howard University, Washington, DC 20059, USA
[3] Research & Exploratory Development Department, Johns Hopkins Applied Physics Laboratory, Laurel, MD 20723, USA
* Correspondence: ariel.greenberg@jhuapl.edu

Abstract

The deployment of autonomous systems in human environments demands sophisticated mechanisms for recognizing and preventing harm. This paper proposes an innovative discovery method for identifying harm-relevant features through the systematic analysis of thick harm verbs—semantically and pragmatically rich linguistic concepts like "puncture", "crush", or "poison" that encode both the mechanics and normative evaluations of specific harm types. By analyzing thick harm verbs to extract the information they encode, we can systematically identify the objects, properties, mechanisms, and contextual conditions that autonomous systems need to track to recognize and prevent harm. We demonstrate how this discovery method can be implemented with the support of large language models as analytical assistance tools, showing how human analysts can operationalize the framework with current technology. The resulting feature specifications discovered through this method provide foundations for constructing harm ontologies that bridge abstract ethical principles and concrete system requirements, addressing a critical gap in autonomous systems design while maintaining explanatory transparency essential for safe deployment in human environments.

Keywords: harm ontology; autonomous systems ethics; machine ethics; harm recognition; thick concepts; LLM-assisted analysis

1. Introduction

The rise of autonomous and semi-autonomous systems has ushered in transformative changes across diverse domains, from healthcare and transportation to domestic assistance and industrial applications [1]. These systems hold the potential to enhance efficiency, reduce human labor, and save lives, but their deployment also introduces profound ethical challenges [2–4]. However, they also raise pressing concerns about their capacity to respond appropriately to harm or risk of harm, whether through unintended physical injury, emotional distress, or disruptions to social and economic wellbeing. Current ethical frameworks for autonomous systems emphasize high-level principles such as beneficence, justice, and autonomy, but these principles prove insufficient to guide the complex, real-time decision-making that harmful situations require [5]. The problem is not merely that abstract principles lack specificity, but that they fail to address the fundamental challenge of determining when particular actions or events actually constitute harm in the first place.

This problem is made clear when we consider the basic requirement that autonomous systems must avoid harming people. In order to avoid harm, an autonomous system must

Check for updates

Academic Editor: Constantin Bratianu

Received: 1 October 2025
Revised: 22 December 2025
Accepted: 25 December 2025
Published: 4 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

be able to distinguish harmful from harmless actions across diverse contexts—a capacity that depends on detailed knowledge of the mechanisms through which harm occurs, the conditions that enable harmful interactions, and the interests that explain why particular changes constitute setbacks to human welfare. Such knowledge proves both highly contingent and context-specific. For example, exposing an individual to peanut oil constitutes a harmless action in most circumstances, but this same exposure risks severe harm when the individual has a peanut allergy. Tracking these relational and contextual features such that autonomous systems can effectively reason about harm demands a level of granularity and specificity that falls outside the scope of existing top-down ethical frameworks.

This implementation challenge reveals a critical gap in autonomous systems design: the absence of actionable harm ontologies that bridge the gulf between abstract moral principles and concrete recognition capabilities. Existing approaches to harm reasoning prove inadequate in different but related ways. Abstract theoretical frameworks provide essential normative foundations but offer insufficient mechanistic detail for practical implementation [1,3,4,6]. They specify that autonomous systems should avoid causing harm but provide little guidance about the specific environmental features, object properties, or contextual conditions that systems must monitor to prevent harmful interactions. The resulting gap between principle and practice creates an urgent need for structured knowledge representations that preserve normative content while providing the detailed specifications required for autonomous systems to recognize harmful situations across diverse contexts. Such harm ontologies must capture not merely that certain actions cause harm, but precisely how harmful interactions unfold, under what conditions they occur, and why they constitute setbacks to human welfare.

In this paper, we propose an innovative approach to constructing a harm ontology for autonomous systems centered on the analysis of thick harm verbs. Thick harm verbs—semantically and pragmatically rich terms like *puncture*, *crush*, or *poison*—encapsulate both the mechanics and normative evaluations of harm in specific contexts. By analyzing these verbs, we can extract a natural language ontology that reflects the concrete features and conditions that people are typically responsive to when reasoning about harms. This ontology serves as a bridge between high-level ethical principles and the detailed, context-sensitive reasoning required for autonomous systems to operate safely and ethically.

Building on Greenberg's schema for harms-sensitive reasoning, we propose that analyzing thick harm verbs offers a systematic path toward constructing an actionable harm ontology for autonomous systems [7,8]. Greenberg's framework establishes key requirements for harm representation, including the need to model: (1) the specific attributes of objects and agents that enable or constrain harmful interactions, (2) the vulnerability profiles of potential victims, and (3) the social and normative contexts that modulate whether particular interactions constitute harm. Our analysis extends this work by showing how thick harm verbs naturally encode precisely these features while remaining grounded in ordinary moral reasoning.

This paper makes three interconnected conceptual and methodological contributions. First, we develop a conceptual framework and discovery method for identifying harm-relevant features through local analyses of individual thick harm verbs. Each analysis addresses four fundamental dimensions: the causal mechanisms through which harm occurs, the material prerequisites that enable harmful interactions, the interests whose violation explains why particular changes constitute setbacks to interests, and the contextual conditions that modulate when actions or events constitute harm. By analyzing thick harm verbs to extract this encoded information, we generate mechanistically explicit accounts with well-defined scope that avoid the limitations of both abstract theorizing and black-box learning. Second, we demonstrate how human analysts can use large language

models (LLMs) as analytical assistance tools to generate candidate feature specifications for validation and refinement. Third, we outline how individual analyses can be integrated by identifying shared structural features, producing integrated harm ontologies that maintain explanatory transparency while achieving broad coverage across harm domains. Together, these contributions provide a concrete methodology for discovering what features belong in harm ontologies without requiring either top-down theoretical derivation or comprehensive empirical datasets.

2. Background

The transition from high-level ethical principles to actionable harm recognition represents one of the most pressing challenges in autonomous systems design. Current frameworks operate at such a level of abstraction that they provide minimal guidance for the specific recognition and response capabilities that autonomous systems require to function safely in human environments. To understand why thick harm verbs provide a solution to this implementation challenge, we must first examine why existing approaches prove inadequate and how semantic analysis can bridge the gap between ethical theory and practical system design.

2.1. From Abstract Principles to Concrete Implementation

For autonomous systems to be used safely, they must avoid doing harm. The problem, however, is that distinguishing harmful actions and events from harmless ones requires detailed and specific knowledge about the different ways in which people can be affected by interaction. Since this kind of knowledge is highly contingent, varied, and context-specific, it does not lend itself to being modeled in a top-down theory-driven way [1,3,4,6]. This suggests that what is needed instead is a bottom-up approach that infers the structure of the knowledge domain from a large set of instances. However, this approach also cannot work, for three reasons: First, there is no obvious dataset that can be used to do this. Even if one could find a dataset that contained instances of harms (e.g., videos of people being injured or harmed online) there is no guarantee that the dataset is comprehensive in the sense that it has sufficient examples of the full range of harms that an autonomous system needs to be responsive to (for instance, there may be no videos of pinpricks even though being pricked by pin-like objects is a fairly common way to be harmed).

Moreover, because harming can occur at different timescales and through different modalities (e.g., physical harms vs. emotional harms vs. financial harms), this kind of approach is likely to limit the generality of the resulting model. Second, whether a harm has occurred typically depends on facts about (1) the causal structure of the situation and (2) the interests of the person involved, neither of which are easy to extract using the kinds of associative learning algorithms that are standardly deployed in AI research. Third, because of the moral significance of harms, it is important that an autonomous system not only avoid doing harm but also do so for the right reasons. Otherwise, there is no guarantee that the autonomous system with continue to not do harm in novel contexts. As such, a black-box approach to constructing a harm ontology is unacceptable, since the inferences it relies on to distinguish harms from non-harms matter too.

Our proposal is that analyzing thick harm verbs to extract the information they encode provides a way out of this methodological dilemma: The analysis of each verb is essentially a top-down process and thus brings with it the associated benefits. Each analysis yields a local model of a particular way of harming which specifies features relevant to that way of harming. Those features, in turn, can be inferentially to each other through the construction of harm ontology. Because each analysis makes the underlying inferential relations explicit, this avoids the black-box problem. However, because these models are local, there is no

assumption of cross-contextual generality. Rather, the models will apply to precisely the range of contexts in which the harm verb is well-defined. Once the terms are analyzed, the resulting local models can then be related to each other through a bottom-up process in which common feature in the models are identified and used to connect them to each other. This will allow the various models to be integrated into a general ontology of the harm domain and optimized to eliminate redundant information. The result is a bottom-up approach to determining which features need to be kept track of without relying on a priori assumptions about the complexity of the harm domain.

One benefit of this approach is that the features identified in this way can be expected to cohere features because ordinary speakers have developed and refined the concepts that thick harm verbs encode through repeated language use. When we analyze these verbs, we make explicit what competent speakers already know implicitly when they distinguish harmful actions from harmless ones. The validity of discovered features thus derives from the coherence already present in ordinary moral reasoning, not from any external validation mechanism. This grounding in natural language coherence explains why the method can generate reliable specifications without requiring comprehensive empirical datasets or formal theoretical derivations.

Furthermore, this approach is readily operationalizable using current LLM capabilities. Contemporary LLMs demonstrate substantial competence in semantic analysis tasks, making it feasible to prompt them systematically for the detailed mechanistic and contextual information that thick harm verbs encode. For instance, when prompted to analyze "puncture" using constrained decoding techniques, the LLM generated outputs that capture the causal mechanisms (concentrated tip pressure, surface breach, tract formation), material prerequisites (sharp objects, penetrable surfaces), and contextual conditions that determine when puncturing constitutes harm versus therapeutic intervention.

This allows for a piecemeal approach to discovering harm-relevant features that can build out dimensions of harm as needed. For instance, in the case of robots, one might be primarily interested in avoiding physical harms and thus focus on physical harm verbs whereas one might focus on psychological harm verbs if one is building a chatbot. This approach makes it possible to discover features for each of these domains separately and integrate them later.

2.2. Methodological Note

Importantly, this paper aims to make a conceptual and methodological contribution rather than an empirical one. We propose a discovery method for identifying harm-relevant features through the analysis of thick harm verbs and demonstrate that this method can be operationalized using current technology. The method treats thick harm verbs as encoding information about the objects, properties, mechanisms, and conditions that matter for harm—information that competent language users implicitly understand but that must be made explicit for autonomous systems implementation. Our contribution is the systematic methodology for extracting this information, not the specific features discovered through any particular application of the method.

The LLM implementations presented in this paper were created using ChatGPT-5 Pro and are intended to serve as analytical assistance tools for human analysts rather than as a method for autonomously generating an ontology. By showing that an LLM can help analysts extract mechanistic specifications from "crush" or identify thickness orderings for harm verb chains, we establish that the proposed analytical approach is implementable with existing capabilities. The value of using LLMs is that they are able to rapidly generate candidate specifications of the structure of thick harm verbs—a cognitively demanding task that would be difficult for analysts to complete in a systematic and reliable

fashion using traditional methods. However, these outputs require human validation and refinement—the LLMs function as analytical assistants rather than authoritative sources.

These prototype analyses function as existence proofs that the theoretical framework translates into practical methodology. A full empirical evaluation of ontologies constructed using this discovery method would require separate research examining how effectively autonomous systems employing these ontologies recognize and respond to harmful situations across diverse contexts. Such validation lies beyond the scope of this paper, which aims to establish the conceptual foundations and demonstrate operationalizability for what we argue represents a promising new approach to the implementation challenge in autonomous systems ethics.

3. Thick Concepts and Thick Harm Verbs

The analysis we propose draws its theoretical foundation from established philosophical work on thick concepts and adapts these insights for the specific requirements of autonomous systems design. To understand how thick harm verbs can serve as resources for discovering harm-relevant features, we must first establish their relationship to the broader philosophical tradition of thick concept analysis and clarify how our approach extends that tradition to address practical implementation challenges (discussed in the Philosophical Foundations subsection below).

Thick harm verbs are semantically and pragmatically rich verbs that describe particular ways of harming. The distinction between thick and thin is graded rather than binary. Thus, harm verbs can be more or less thick. Although there is no general metric for comparing the thickness, it is possible to construct local orderings by means of the "is-a-way-of" relation. For instance, the verb to bludgeon is intuitively thicker than the verb to hit. In this case, we can make sense of this intuition by means of the "is-a-way-of" relation: bludgeoning someone *is a way of* hitting someone. By definition, the verb *to harm* is the thinnest harm verb because all harm verbs describe ways of harming. Note that to satisfy this condition, it is often necessary to characterize the harming reflexively, i.e., in terms of harming oneself. This means, for instance, that for passive harm verbs like to fall the correct construction is that "falling is a way of harming oneself" rather than "falling is a way of being harmed".

It is important to distinguish between harm verbs and their corresponding nouns as they mean different things (e.g., *to burn* versus *a burn*). Typically, the verb characterizes the way the harm comes about whereas the noun refers to the injury produced in being harmed. Thick harm verbs often have both an agent-oriented construction and an object-oriented construction. The agent-oriented construction has an agent as the subject of the sentence. The agent is typically harming something with something. For example, "Tina burned Warren's skin with fire". The object-oriented construction has an object as the subject of the sentence. The object is typically the thing with which the harm was caused. For example, "The fire burned Warren's skin".

We operationalized these theoretical thickness relationships through systematic LLM-assisted analysis. By prompting LLMs with structured queries about entailment relationships, analysts can identify the "is-a-way-of" hierarchies that establish thickness orderings. For example, our prototype analyses confirmed that "bludgeon $\succ$ beat $\succ$ hit $\succ$ harm" forms a valid thickness chain, where each arrow represents increased specificity through additional mechanistic constraints. This demonstrates how the theoretical framework can be made operational through existing AI capabilities as analytical assistance tools.

Philosophical Foundations

Our approach derives its theoretical foundation from two complementary developments in contemporary philosophy: Bernard Williams' [9] analysis of thick ethical concepts

and Nancy Cartwright's [10] related account of thick causal concepts. Williams' analysis emerged from his critique of the tendency within moral philosophy to reduce complex evaluative language to abstract theoretical principles, while Cartwright's account takes up parallel reductionist pressures within the philosophy of science. Both frameworks share a common insight about the irreducible unity of descriptive and evaluative (or causal) content in semantically rich concepts. Together, they provide the conceptual foundation for understanding how thick harm verbs encode the specific mechanisms and normative features essential for practical harm reasoning.

Williams' account of thick ethical concepts challenged the dominant analytical approach of decomposing moral language into separate descriptive and evaluative components. According to Williams, terms such as "treachery", "brutality", and "courage" resist this kind of decomposition because they express what Williams characterized as a union of fact and value that cannot be separated without loss of content [9]. This irreducibility has important implications for practical reasoning. According to Williams, thick concepts organize our understanding of situations by directing our attention toward features that are simultaneously factual and normatively significant. When someone recognizes potential treachery in a situation, they are responding to specific patterns of commitment and behavior that have inherent moral significance. In this way, thick ethical concepts do not first describe behaviors neutrally and then add an evaluative judgment, rather they identify those behaviors through an evaluatively structured representation of what constitutes the behavior or disposition.

Cartwright's analysis of thick causal concepts extends Williams's insights to scientific reasoning about causation. Just as Williams argued that thick ethical concepts cannot be reduced to separate descriptive and evaluative components, Cartwright maintains that thick causal concepts such as "compress", "attract", "smother", "enrich", and "clog" cannot be decomposed into a thin causal relationship plus additional non-causal information [10]. When we say that a soil fumigant "poisons" eelworms, we are not combining the abstract claim that the fumigant causes some effect with a separate specification of the mechanical process involved. Rather, "poisoning" represents a unified understanding of a particular type of causal mechanism that includes both the causal relationship and the specific conditions under which that relationship obtains.

Williams's and Cartwright's analyses provide the foundation for extending thick concept analysis to harm reasoning. Thick harm verbs combine factual and evaluative content and encode specific mechanistic information that resists reduction to abstract causal claims. This dual inheritance means that thick harm verbs provide both the mechanical specificity required for practical intervention and the evaluative structure needed to explain why intervention matters.

Nevertheless, it is important to emphasize that our understanding of thick harm verbs departs from standard treatments of thick ethical concepts in the philosophical literature in a theoretically significant way. Most discussions of thick ethical concepts assume that evaluative content is constitutive of their meaning—that terms like "treachery" or "courage" cannot be understood without grasping their inherent ethical significance [11]. This assumption has generated theoretical debates about whether evaluative and descriptive content are truly inseparable and whether implied evaluations belong to the semantics rather than the pragmatics of the term. By contrast, our understanding of thick harm verbs is somewhat broader and includes terms that describe ways of harming only under some conditions, meaning that such terms do not refer to harmful events or actions by virtue of their meaning alone. The reason for this breadth is that it serves our specific theoretical purposes that extend beyond traditional concerns about semantic unity.

The verb "to hit" demonstrates why this broader conception proves necessary for our purposes. It is clear that hitting constitutes harm when directed at a person but not when directed at a pillow. This contextual variation means that harmfulness emerges from pragmatic features of usage rather than semantic content alone. Yet "hit" remains a thick harm verb according to our definition because hitting represents a specific way of causing harm under appropriate conditions. The key insight here is that the contextual modulation of harm significance does not undermine the theoretical value of thick harm verbs for autonomous systems design—rather it adds to it. By using thick harm verbs to delineate the contextual conditions that modulate when a type of action or event constitutes a harm, we can extract precisely the kind of conditional information that autonomous systems require to distinguish harmful from harmless instances of the same action type. In this way, contextually variable harm verbs provide a valuable pathway from linguistic analysis to the specification of monitoring and assessment capabilities that harm-responsive machines require to operate safely in human environments.

This methodological choice reflects a deeper point about the approach to linguistic analysis we adopt in what follows. The practical requirements of designing harm-responsive autonomous systems generate different theoretical priorities than those that have traditionally motivated analysis of thick ethical concepts. When philosophers analyze thick ethical concepts, they typically aim to resolve questions about the nature of moral language and the relationship between descriptive and evaluative content. However, when we analyze thick harm verbs for the purpose of discovering harm-relevant features, our primary concern is not establishing the correct semantic analysis of these terms but rather extracting actionable information that can guide system design. That is, our approach treats the analysis of thick harm verbs as a means to identify the environmental features, object properties, and contextual conditions that determine when particular actions or events constitute harm.

Thus, the goal of the analytical framework we develop in what follows is not to get the linguistic analysis of these words right per se, but rather to provide a structured method for discovering features that autonomous systems must track in order to effectively identify and respond to harmful situations. This instrumental use of linguistic analysis means that many of the theoretical questions and concerns that exercise linguists and philosophers—such as whether an implication is a feature of the semantics or pragmatics of a term—can be largely left to one side. Instead, what matters for our purposes is whether the analysis of thick harm verbs yields insights about the conditions under which harmful interactions occur and the mechanisms through which harm is produced.

4. How to Model Thick Harm Verbs

As we have just been discussing, the epistemic value of thick harm verbs for autonomous systems research derives from their capacity to encode information about the concrete mechanisms and conditions of harm that are unavailable when harm reasoning relies exclusively on abstract moral principles or semantically thin concepts. Consider the difference between researchers knowing that a proposed action "causes harm" versus knowing that the action "punctures", "crushes", or "traumatizes" a human agent. The abstract characterization provides no actionable guidance for system designers about the specific mechanisms through which harm occurs, the environmental conditions that enable harmful interactions, or the particular vulnerabilities that make individuals susceptible to different forms of damage. This absence of mechanistic detail creates what we might call the implementation problem for autonomous systems ethics: the inability to translate high-level normative principles into specific monitoring and intervention strategies [12].

Thick harm verbs encode information that helps solve this implementation problem, in part, because they almost always constitute thick causal concepts in Cartwright's

sense—they encode detailed information about specific causal processes as part of their semantic and pragmatic structure. The theoretical significance of this detailed encoding becomes clear when we recognize how much more information thick harm verbs carry compared to abstract causal claims. When we say that sharp glass "punctures" skin, we are specifying a precise causal mechanism that includes information about the objects involved (sharp, pointed implements against membrane-like surfaces), the process by which harm occurs (penetration leading to structural breach), the conditions under which this process operates (sufficient force applied to overcome material resistance), and the interests that explain why this constitutes harm (setback to interest in bodily integrity). This informational density has crucial implications for autonomous systems design: it means that systematic analysis of thick harm verbs can extract specifications for monitoring requirements, intervention strategies, and contextual assessments without requiring additional theoretical machinery to bridge the gap between abstract principles and concrete implementation.

More specifically, we can use the analysis of thick harm verbs to systematically extract four interconnected types of information that researchers require to discover harm-relevant features for autonomous systems: causal mechanisms, material prerequisites, underlying interests, and contextual conditions. These can be elicited by addressing four corresponding fundamental questions about harming: the how, the what, the why, and the when of harm occurrence. Each of these questions targets a different dimension of the information encoded in thick harm verbs, and together they provide a comprehensive framework for extracting the knowledge required for harm-responsive autonomous systems.

4.1. The How: Mechanisms and Processes

The first question concerns the particular causal mechanism or process through which harm is brought about. This mechanistic dimension emerges from the fact that thick harm verbs are always thick causal concepts that encode specific forms of causal interaction. The practical significance of this mechanistic encoding becomes apparent when we recognize that it makes accessible what we might call process-specific causal knowledge—knowledge that specifies not merely that one thing causes another, but precisely how that causal relationship obtains under determinate conditions.

Consider how this process-specific encoding operates across different harm types. The verb "to crush" describes harm through compression forces that exceed material tolerance limits, thus revealing information about a specific dimension of physical interaction that harm-responsive systems must monitor to mitigate harm. The causal mechanism here involves the application of force across a surface area that results in structural deformation beyond the material's elastic recovery capacity. This mechanistic specification entails that autonomous systems must track not merely the presence of objects that could exert force, but also the magnitude and distribution of applied forces relative to the material properties of potential targets. Alternatively, the verb "to poison" specifies harms that involve exposing biological systems to toxic substances at concentrations sufficient to disrupt normal physiological processes, thus indicating requirements for tracking chemical compositions and biological response patterns. The causal mechanism here operates through molecular interactions that interfere with biological function, which means that effective harm prevention requires monitoring not only the presence of potentially toxic substances but also their concentrations, delivery mechanisms, and the physiological vulnerabilities of potential victims.

LLMs are remarkably effective at identifying these mechanistic details when prompted appropriately. Our prototype analysis of "to crush" demonstrates how LLM-assisted analysis can systematically identify this mechanistic information. For example, when prompted to identify the primary mechanism of crushing, the analysis revealed: "compression (often

with entrapment)". This mechanistic analysis directly specifies the features autonomous systems must monitor (Figure 1).

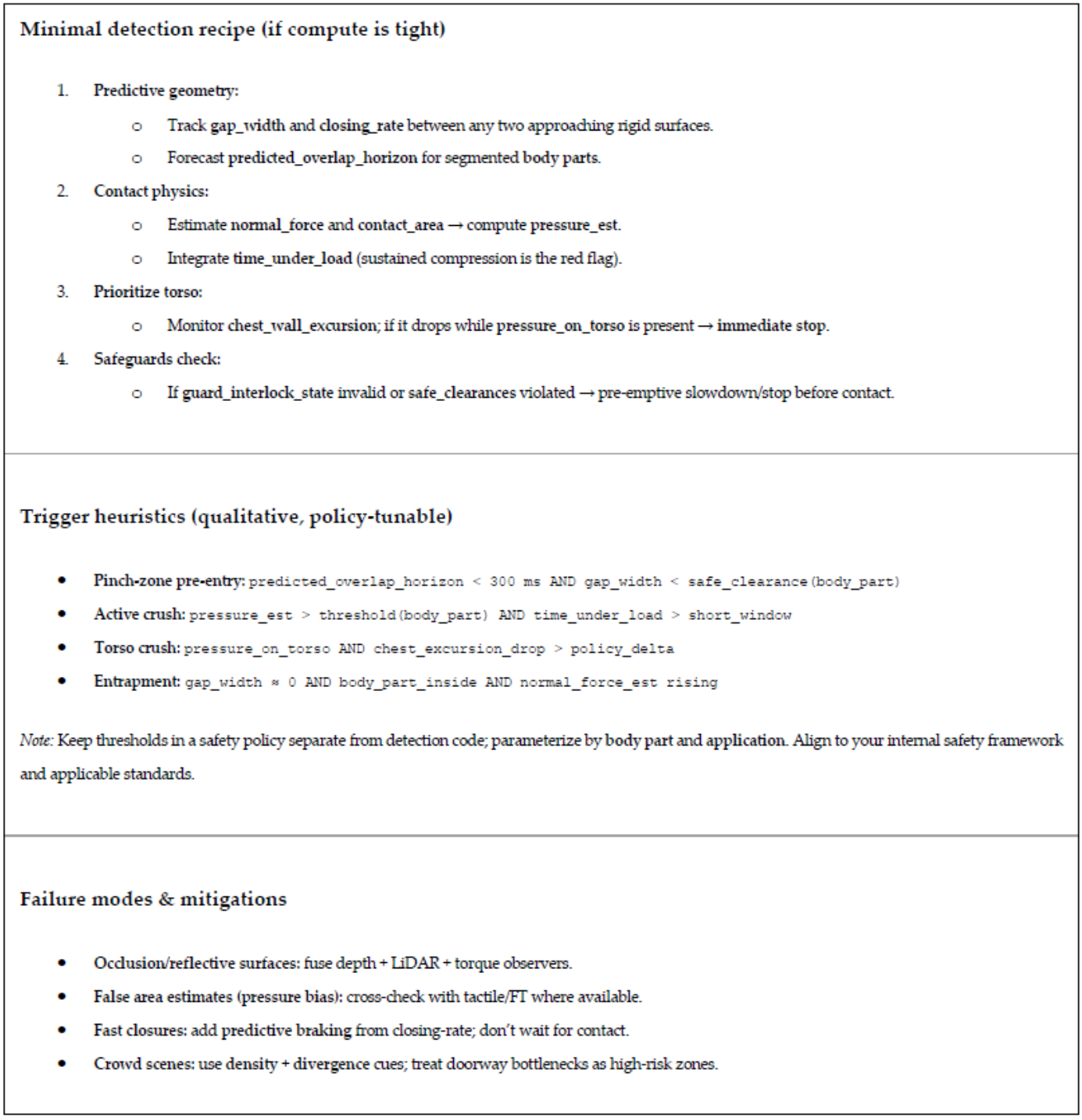


Figure 1. Crush-risk monitor specification.

Such detailed mechanistic specifications enable the design of targeted sensor arrays and computer vision algorithms that can recognize crush-risk scenarios before harm occurs. By cataloging the mechanistic information discovered through thick harm verb analysis, researchers can identify the features and processes that autonomous systems must track to recognize potential harm-producing actions and situations before these processes result in actual injury. That is, by analyzing thick harm verbs to extract information about how

different types of harm occur, we can design systems that can effectively prevent those harms from occurring across a range of contexts.

4.2. The What: Objects, Agents, and Material Requirements

The second question addresses the types and properties of objects and agents that constitute the material prerequisites for particular forms of harm. Because thick harm verbs are a type of thick causal concept, they typically specify not only causal mechanisms but also the kinds of entities that can participate in those mechanisms. Causal mechanisms can only operate when the appropriate material conditions are satisfied. For instance, there can be no crushing without objects capable of exerting compressive force and materials capable of undergoing compression, and there can be no drowning without a fluid in which an air-breathing entity can be submerged.

The conditions are reflected by constraints on the range of objects and agents that can satisfy the verb's semantic structure. "Puncturing", for instance, requires both sharp objects capable of penetration and penetrable surfaces that exhibit membrane-like properties, whereas "poisoning" requires toxic substances and biological systems capable of being disrupted by those substances. The constraint operates through compatibility requirements: the verb can only be correctly applied when its semantic structure can be satisfied by entities with the appropriate properties. This means that puncturing cannot occur without pointed geometries capable of penetrating membrane-like surfaces, and poisoning cannot occur without chemical agents that can interfere with biological processes.

Importantly, these material requirements not only include simple object properties but also extend to relational attributes that emerge from object interactions in specific contexts. This extension proves necessary because harm production typically depends not merely on the intrinsic properties of individual objects, but on how those properties interact under particular conditions. Membrane-like properties, for instance, involve not just material composition but also thickness, flexibility, and resistance to penetration under specific force applications. Sharp objects require not only pointed geometries but also sufficient hardness and structural integrity to maintain their penetrative capacity during interaction with resistant surfaces. "Burning" necessitates heat sources and thermally vulnerable materials, but the specific vulnerability depends on factors such as thermal conductivity, specific heat capacity, and combustion thresholds. These relational attributes create what we might call conditional compatibility requirements—requirements that are only satisfied when objects with particular properties interact under appropriate conditions.

Moreover, LLM analysis proves particularly valuable for identifying these material requirements systematically. For the analysis of "to puncture", the system helped identify that the process requires "artificial tips: needles, nails, tacks, awls, barbs, syringes, projectiles with sharp points" as well as "biological tips: fangs, spines, stingers, thorns" paired with penetrable surfaces exhibiting membrane-like properties.

In this way, researchers can determine which features autonomous systems must detect and analyze to assess harm potential, moving beyond requiring explicit harm labels for every possible object interaction toward principled recognition of harmful interaction possibilities. By modeling which combinations of object properties allow particular forms of harm, researchers can design recognition systems that identify those combinations rather than requiring exhaustive catalogs of harmful objects. Thus, instead of programming autonomous systems to, for example, avoid all sharp objects, we can program them to recognize the relational properties that make sharp objects dangerous in particular contexts—their capacity to puncture membrane-like surfaces when applied with sufficient force.

4.3. The Why: Interests and Normative Significance

The third question, and perhaps the most significant from a theoretical point of view, concerns the underlying interests that explain why particular actions or events constitute harm rather than mere change or inconvenience. It is the interference with an agent's interests that explains why an action or event is bad for an agent and thus why it constitutes a harm [13]. As such, this question constitutes the normative core of harm reasoning: answering it tells us why certain changes in the world count as making someone worse off rather than merely making them different. Without an account of why particular changes constitute harm, autonomous systems would lack the conceptual resources necessary to distinguish between harmful and merely inconvenient interactions.

The key to resolving this problem lies in recognizing that harm concepts are essentially relational—they specify not merely that a change has occurred, but that the change constitutes a setback to something that matters to the individual. We can leverage this relational structure to systematically identify the relevant interests. Many thick harm verbs can describe actions or events that do not involve harm when applied to entities that lack the relevant interests. The sentence "the pencil punctured the balloon" does not describe a harm since balloons are not entities capable of having interests (although they can be damaged). By examining what makes the difference between harmful and non-harmful applications of the same verb, we can determine the specific interest that explains why the action is bad for agents when it affects them.

This analytical strategy reveals that every thick harm verb necessarily implies at least one interest that explains why the described action or event makes an individual worse off, when it does. To identify this interest, we ask: what explains why this action or event is bad for the agent harmed? The answer to this question reveals the interest whose interference constitutes the harm. "Puncturing" is harmful to persons because creating a hole in living tissue undermines their interest in bodily integrity—this interest in maintaining structural wholeness explains why such damage constitutes harm rather than mere a neutral change (body art considered in a following section). "Imprisoning" is harmful because restricting movement undermines an individual's interest in autonomy—this interest in freedom of action explains why such constraint constitutes harm rather than mere inconvenience or spatial displacement.

This relationship between thick harm verbs and underlying interests provides researchers with a discovery mechanism for identifying the range of interests that autonomous systems must track to implement comprehensive harm prevention. The methodological advantage here is that this discovery process operates bottom-up rather than top-down: rather than beginning with predetermined lists of interests derived from philosophical theory, the analysis of thick harm verbs reveals which interests are actually implicated in human reasoning about harm across diverse contexts. This has crucial implications for the construction of harm ontologies because it ensures that the resulting frameworks capture the full range of interests that people actually have rather than the interests that philosophers typically focus on.

The discovery process operates through systematic analysis that reveals interest patterns embedded in ordinary moral reasoning. Each thick harm verb analysis generates evidence about which interests explain why particular actions constitute harms for agents. Analyzing "traumatize" reveals interests in psychological wellbeing and emotional stability that explain why overwhelming experiences are bad for individuals. Analyzing "deceive" indicates interests in epistemic reliability and trust that explain why false information harms agents who receive it. Analyzing "humiliate" points to interests in dignity and social standing that explain why public degradation harms individuals beyond mere embarrassment. This emergent ontology of interests provides autonomous systems with the normative

foundations necessary to recognize and respond to the full range of ways that individuals can be made worse off through interaction with their environment. Rather than relying on abstract principles that may fail to capture the concrete structure of human welfare concerns, the analysis of thick harm verbs generates an empirically grounded understanding of which interests actually matter for harm prevention in human social contexts.

Notably, this interest-discovery approach proved highly effective in our prototype analyses (Figure 2). When analyzing "puncture", the LLM systematically identified that puncturing "harms by concentrated tip pressure that breaches a surface and drives a narrow tract through tissues, causing direct structural damage" which specifically undermines bodily integrity interests. The analysis further revealed secondary interests including pain management, functional capacity, and infection resistance. This multi-layered interest analysis can provide autonomous systems with the evaluative framework needed to prioritize different types of interventions—understanding that puncture risks affect not only immediate structural integrity but also longer-term functional and health interests.

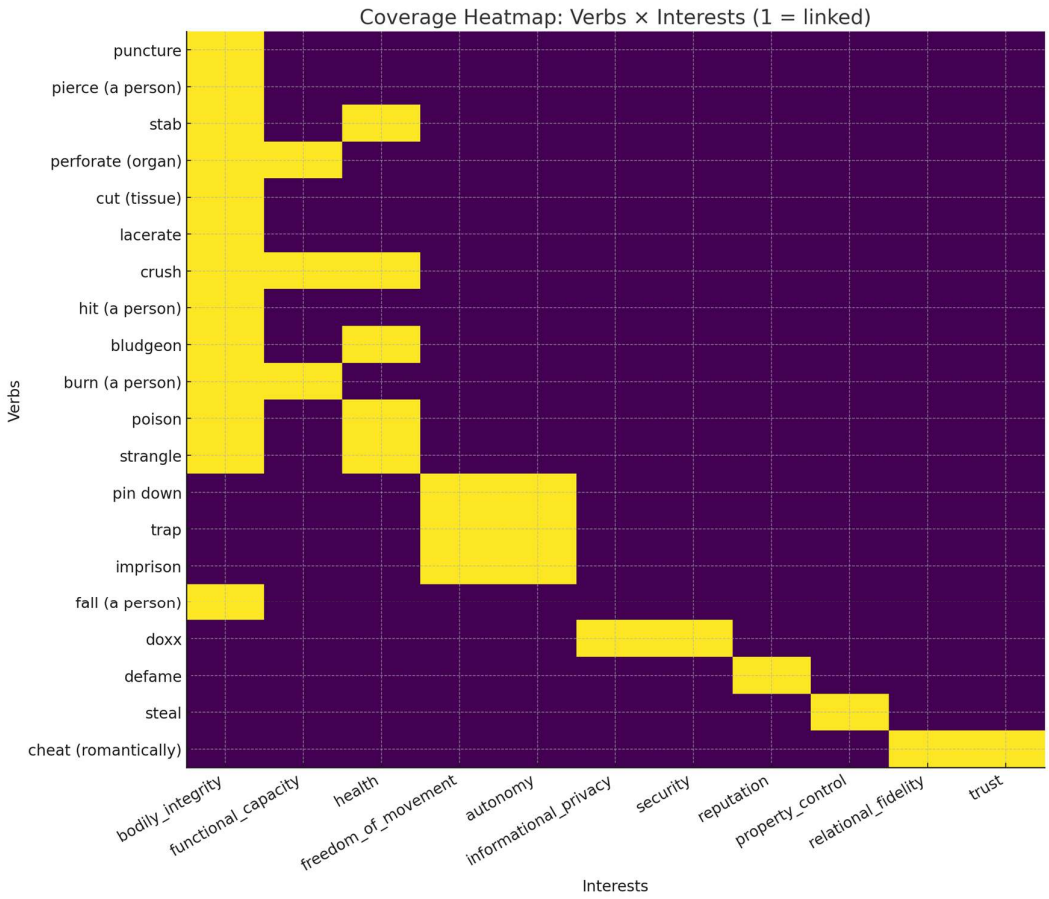


Figure 2. Harm verbs and their discovered interests. Yellow indicates which interest bears on which verb.

4.4. The When: Contextual Conditions

The fourth question addresses the additional contextual and constitutive conditions that modulate when particular actions or events constitute harm. This question emerges as a consequence of the answers to the first three questions because the mechanistic, material, and interest-based requirements identified through those analyses create conditional structures that determine when harm actually occurs. This conditionality reveals why harm attribution requires more than simply identifying the presence of harmful mechanisms

and vulnerable objects—it requires understanding the contextual factors that determine whether the interaction between mechanism and vulnerability actually constitutes harm.

These modulating conditions operate through pragmatic constraints that govern when and under what conditions a given event or action type constitutes harm. Thick harm verbs inherit their harm-constituting properties not merely from their semantic content, but from the interaction between that content and the pragmatic conditions under which the verbs are applied. Consider how this operates in paradigmatic cases. The act of restraining typically constitutes harm through autonomy violation because it satisfies the mechanistic requirements (restricting movement), the material requirements (involving agents capable of autonomous action), and the interest requirements (undermining freedom). However, the same physical action ceases to be harmful if the person has consented to be restrained because consent alters the normative context in which the action occurs—it transforms the action from an autonomy violation into one that is consistent with it.

The inference here is crucial for autonomous systems design: it shows that harm recognition cannot operate merely through the identification of harmful mechanisms and vulnerable targets, but must also incorporate assessments of the normative contexts that determine whether particular interactions constitute harm. This has significant implications for how autonomous systems must be designed because it means that effective harm prevention requires not only perceptual capabilities for detecting harmful mechanisms and vulnerable objects, but also reasoning capabilities for assessing contextual factors such as consent, intention, and social roles.

Modeling these contextual dependencies provides an additional way to expand our understanding of how harm unfolds in dynamic contexts, as not all thick harm verbs respond to contextual modulation in the same way. Some verbs necessarily describe a harm and thus resist contextual modulation entirely, while others can be defeated by specific conditions like consent, and still others vary in their harm-constituting properties depending on their target objects. This variation suggests that effective autonomous harm reasoning requires a systematic understanding of how different types of thick harm verbs respond to contextual factors. To capture these differences, we propose a taxonomic classification that organizes thick harm verbs according to their contextual sensitivity patterns.

Three Classes of Harm Verb

Thick harm verbs can be systematically classified into three distinct categories based on how modulating factors affect their harm-constituting properties, each with different implications for autonomous systems design (Figure 3).

Absolute harm verbs can only be used to refer to harmful events or actions. The verbs “poison” and “bludgeon” are absolute harm verbs since part of what it means to be poisoned or bludgeoned is that one has been harmed. Social harms are often described using absolute harm verbs—the verbs “cheat” (romantically) and “steal” can only be used to refer to harmful actions. Importantly, these absolute social harms often have constitutive conditions that modulate whether the action actually constitutes the harm. In the case of both cheating and stealing, whether someone’s action constitutes cheating or stealing depends on whether the relevant party gave that person permission to perform the underlying action.

Defeasible harm verbs refer to harm unless some further condition is satisfied (the defeater). The verb “pin down” is defeasible because when we say that someone is pinned down, we take this to be harming the person. However, if we discover that the person chose to be pinned down, then we cease to think that they are being harmed. Defeasible harm verbs differ from absolute harm verbs in that they continue to describe the situation even if the harm component has been defeated. Unlike cheating or stealing, the fact that

someone consented to being pinned down does not make it any less true that they are pinned down—it makes it untrue that they are being harmed by being pinned down.

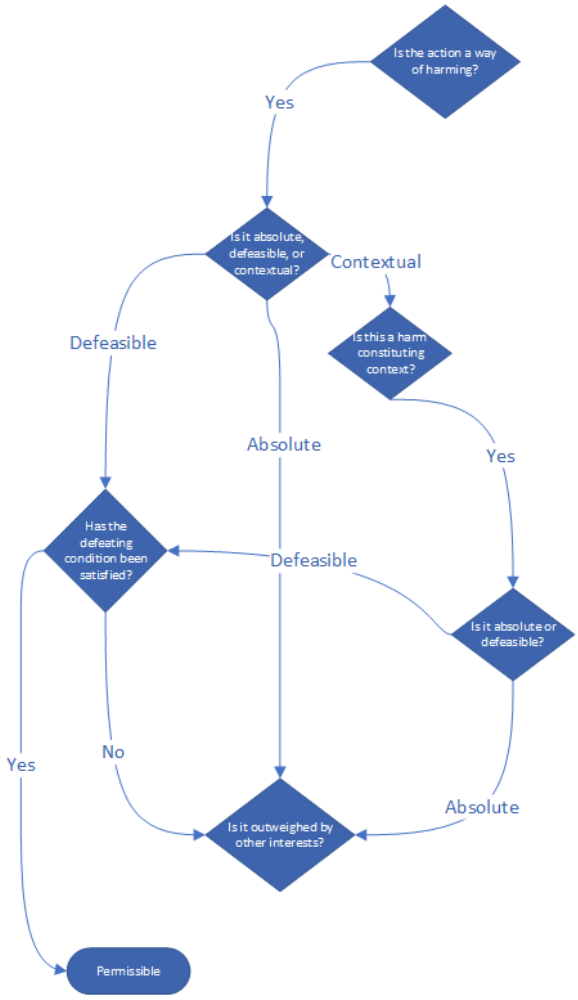


Figure 3. Flowchart for classifying thick harm verbs.

Contextual harm verbs describe actions or events that only constitute harms under specific circumstances. These are typically verbs that can have either a person or a thing as their grammatical object but only constitute harm if the object is a person (such as "puncture" or "hit"). The verb "pin down", when its object is a person, is not a contextual harm verb because it implies a restriction on agency. When we focus on the harming contexts of contextual harm verbs, we can draw a further distinction between contextual absolute harms and contextual defeasible harms. Contextual absolute harm verbs describe actions or events that are always harms when the scope is restricted to persons. The verb "burn" is contextual since we can burn all sorts of things without constituting harm, but it is plausible that anytime a person is burned this constitutes harm. By contrast, contextual defeasible harm verbs describe actions or events that we take to constitute harm when the object is a person, unless certain conditions are met. The verb "fall" is plausibly of this type—when we say that someone fell, we take this to imply that they were harmed unless some further condition is met, such as that they fell into something adequately cushioned.

4.5. Therapeutic Interventions and Contextual Complexity

It can sometimes be difficult to determine whether a verb describes a contextual absolute harm or contextual defeasible harm because many contextual harm verbs describe actions or events that play a role in therapeutic interventions. As a result, it can often be

unclear whether the harm is defeated when the act is performed in a therapeutic context or outweighed by the benefits of the therapeutic use. Because we puncture skin to administer vaccines, we might think that this is a defeasible harm verb that refers to harm unless performed for medical purposes. However, the more plausible interpretation is that the benefits of the vaccine far outweigh the very small cost of being punctured by the needle. Similarly, the fact that someone chose to pierce their ears does not seem to change the fact that this constituted harm. Rather, the more plausible reading is that whatever interest was served by the action outweighed or overrode this concern.

The verb "cut" illustrates this complexity particularly well. Cutting most parts of the body constitutes a contextual absolute harm, except when what is being cut are hard keratin cornified epidermal appendages of the integumentary system: hair (other than eyelashes) or nails. Cutting someone's hair is plausibly a contextual defeasible harm, since it is harm unless that person gave permission (and even then, it may still constitute harm if the haircut was especially bad).

The practical significance of this taxonomic analysis is that it enables discovery of the conditional information that can provide autonomous systems with the specifications needed to distinguish harmful from harmless instances of the same action type. By delineating the contextual conditions that modulate when actions or events constitute harm, thick harm verb analysis enables autonomous systems to implement sophisticated harm recognition capabilities that account for consent, interests, vulnerability, and circumstance. This represents a significant advance over rigid categorical prohibitions that would prove both overinclusive (prohibiting harmless actions) and underinclusive (failing to recognize novel forms of harm) in complex real-world environments where context determines whether particular actions constitute harm.

However, the analytical framework we have developed thus far only operates at the level of individual harm concepts, generating a local model for each thick harm verb. Each analysis produces detailed understanding of particular harm types—the mechanisms through which "crushing" operates, the interests violated by "imprisonment", the contextual factors that determine when "restraining" constitutes harm. Yet autonomous systems operating in complex environments cannot rely on isolated understanding of individual harm types. They require comprehensive frameworks that capture the structural relationships between different forms of harm and enable coherent responses across diverse situations. The challenge is therefore to transform the collection of local analyses into an integrated theoretical framework that preserves the mechanistic precision of individual thick harm verb models while establishing the conceptual architecture necessary for autonomous harm reasoning across multiple domains.

We thus need a method of integrating these discoveries into unified frameworks that organize these individual analyses into a coherent ontology for autonomous systems. Individual thick harm verb analyses generate detailed local models—precise specifications of how "crushing" operates, what interests "imprisonment" violates, when "restraining" constitutes harm. However, autonomous systems deployed in complex environments cannot operate with disconnected collections of harm-specific knowledge. They require unified ontological frameworks that organize these local models according to shared structural principles, enabling coherent harm reasoning across diverse situations and consistent responses when different harm types interact or conflict.

4.6. Tabular Summary of Thick Harm Verbs Modeling

The analysis described above around the *how*, the *what*, the *why*, and the *when* of harm occurrence is given in the Tables 1 and 2 for our example collection of thick harm verbs. The associated monitoring components for each are listed in the last column.

Table 1. Physical harm verbs—How/What/When/Why + Monitoring.

Verb	How (Mechanism/Process)	What (Objects/Agents/Materials)	When (Contextual Conditions)	Why (Interest)	Monitoring Components
puncture	Penetration by concentrated tip pressure; surface breach; narrow tract formation.	Sharp/pointed tip (needle, thorn, fang); penetrable living tissue; sufficient normal force.	Harm when patient is a person/part (CA). Therapeutic uses typically outweigh, not defeat, the harm.	Bodily integrity (structural discontinuity); secondary risks: bleeding, infection.	Penetration Monitor
pierce (a person)	Create a hole through tissue (often full-through).	Sharp implement/projectile; compliant tissue/membrane.	Harm for persons/parts (CA); ornamental/medical contexts usually outweigh rather than defeat.	Bodily integrity.	Penetration Monitor
stab	Forceful blade penetration, often deep and targeted.	Knife/blade wielded by an agent; vulnerable body region.	Always harm if correctly applied (ABS); if no injury, likely misclassification ("poke").	Bodily integrity; health (high lethality).	Penetration Monitor
perforate (organ)	Full-thickness breach of organ wall.	Instrument/projectile/disease; hollow organ or membrane.	Harm for persons/parts (CA).	Bodily integrity; functional capacity (loss of organ function).	Penetration Monitor
cut (tissue)	Shear/incision producing an open wound.	Sharp edge (knife, glass).	Harm for persons/parts (CA); surgery typically outweighs, not defeats, harm.	Bodily integrity.	Shear/Incision Monitor
lacerate	Irregular tearing by shear/tension; jagged wound.	Jagged edge, machinery, high-energy impact.	Harm for persons/parts (CA).	Bodily integrity; functional impact/scarring.	Shear/Incision Monitor
crush	Compression beyond tolerance; structural collapse; perfusion/ respiratory compromise.	Heavy object/roller/vehicle/crowd; opposing surfaces; time-under-load.	Harm for persons/parts (CA).	Bodily integrity; functional capacity; health (e.g., crush syndrome).	Compression Monitor; Blunt Impact Monitor
hit (a person)	Impact transferring impulse to body.	Striker (hand/object/projectile); target body region.	Default harm (CD); defeated by consent/safety contexts (sport, pads, stunt).	Bodily integrity.	Blunt Impact Monitor
bludgeon	Repeated heavy blunt-force strikes.	Blunt, heavy instrument; attacker.	Always harm (ABS).	Bodily integrity; health (severe trauma).	Blunt Impact Monitor; Compression Monitor
burn (a person)	Thermal/chemical/electrical/radiant energy damages tissue.	Heat source/caustic/voltage/UV; susceptible tissue.	Harm for persons/parts (CA).	Bodily integrity; functional capacity.	Thermal Exposure Monitor
poison	Toxin disrupts physiological processes at sufficient dose/route.	Toxic substance; ingestion/inhalation/injection; vulnerable physiology.	Always harm (ABS).	Health; bodily integrity.	Toxic Exposure Monitor
strangle	Asphyxiation via airway/blood-flow restriction at neck.	Hands/ligature/entrapment; sustained pressure.	Always harm (ABS).	Health; bodily integrity (oxygenation).	Airway/Asphyxia Monitor
pin down	Restraint limiting degrees of freedom; immobilization.	Human restrainers or devices; body contact; leverage.	Default harm (DEF); defeated by consent/clinical necessity while verb remains true.	Freedom of movement; autonomy.	Spatial Freedom Monitor
trap	Environmental constraint blocks exits/paths.	Barriers/enclosures/locks; path-planning dead-ends.	Default harm (DEF); can be defeated by consent/safe game contexts.	Freedom of movement; autonomy.	Spatial Freedom Monitor
imprison	Institutional confinement preventing exit.	Custody, enclosure, authority; enforced access control.	Always harm if correctly applied (ABS); permission/authority issues are constitutive.	Autonomy; freedom of movement.	Spatial Freedom Monitor
fall (a person)	Unintended descent; impact on landing.	Gravity; loss of support; hard surface.	Default harm (CD); defeated by soft landing/harness/controlled fall.	Bodily integrity.	Fall Risk Monitor; Blunt Impact Monitor

Table 2. Social/Informational harm verbs—How/What/When/Why + Monitoring.

Verb	How (Mechanism/Process)	What (Objects/Agents/Materials)	When (Contextual Conditions)	Why (Interest)	Monitoring Components
doxx	Exposure of private identifiers to the public.	PII store; publishing channel; audience access.	Always harm (ABS); if info already public, act reclassifies (not doxx).	Informational privacy; security.	Info Privacy Monitor
defame	Publicize false harmful statements.	Text/speech/media; audience; subject identifiable.	Always harm (ABS); falsity is constitutive (truth negates act type).	Reputation/ social standing.	Reputation Monitor
steal	Remove property without owner's consent/control.	Property, taker, owner; access to item; lack of permission.	Always harm (ABS); permission reclassifies to borrow (constitutive).	Property control/resource autonomy.	Property Control Monitor
cheat (romantically)	Violate exclusivity agreement in a relationship.	Partners; exclusivity norm/contract; opportunity.	Always harm (ABS); permission negates act type (constitutive).	Relational fidelity; trust.	Relational Norm Monitor

5. Connecting and Expanding Local Harm Models

The transformation from individual thick harm verb analyses to comprehensive harm ontologies requires systematic identification of the structural relationships that connect different forms of harm. This integration addresses a fundamental challenge in autonomous systems ethics: transforming isolated understanding of particular harm types into comprehensive frameworks that enable responsiveness to harm across diverse contexts. The process operates through identification of shared conceptual structures that connect different forms of harm, ultimately producing both immediate guidance for specific decisions and theoretical foundations for addressing novel cases.

The integration process begins by identifying structural relationships that exist between different thick harm verb analyses through systematic mapping of shared features. These shared elements function as natural organizational principles that transform collections of isolated analyses into coherent conceptual networks. When researchers analyze "puncture", they discover mechanisms involving piercing actions, conditions requiring sharp objects and penetrable surfaces, and interests concerning bodily integrity. This analysis connects directly to related verbs through overlapping structural features: "stab" and "pierce" share the penetration mechanism, "lacerate" and "cut" involve similar interests in physical integrity, and "perforate" combines both mechanistic and interest-based connections. However, the significance of these connections extends beyond mere categorization, as they reveal the underlying conceptual architecture that organizes human reasoning about harm.

The systematic identification of these connections serves a dual function that both ensures comprehensive coverage and enables principled extension of ontological frameworks. First, the networking process operates as a quality control mechanism by revealing gaps in coverage within identified conceptual domains. For example, when analysis of "pin down" reveals that constraining physical autonomy represents the central interest violation, this discovery naturally prompts the examination of related concepts that affect the same interest domain, leading researchers to identify "trap", "imprison", and related constraint-based harm verbs. Second, this generative capacity enables the principled expansion of harm ontologies beyond the initially selected terms through systematic exploration of interest-based and mechanistic connections. The resulting frameworks produce increasingly comprehensive coverage within each identified domain rather than remaining limited to arbitrary initial selections, ensuring that autonomous systems can be designed with robust capabilities for recognizing and responding to the full range of harmful interactions within their operational contexts.

The practical implementation of this integration follows a two-stage optimization process that transforms redundant collections of harm analyses into efficient monitoring frameworks for autonomous systems. In the feature extraction stage, researchers identify the mechanisms, conditions, and interests implied by an individual thick harm verb that enable harmful interactions to occur. For instance, analyzing "puncture" reveals the need to monitor sharp objects and penetrable surfaces, while "crush" indicates the need to track compressive forces and material tolerance limits, and "poison" implies the need to monitor toxic substances and biological vulnerabilities. Nevertheless, this initial extraction typically generates overlapping and redundant feature sets across different harm analyses, creating monitoring requirements that would overwhelm autonomous systems with excessive perceptual demands and computational overhead.

The optimization stage addresses this redundancy through systematic identification of common features that appear across multiple analyses, thereby reducing computational complexity without compromising harm detection capabilities. Movement-constraining verbs, for instance, like "pin down", "trap", and "imprison" all require monitoring of

spatial constraints, movement restrictions, escape route availability, and barrier properties, despite their surface differences in application contexts. The optimization process recognizes these shared environmental requirements and consolidates them into unified model components related to spatial freedom and movement capacity. This consolidation creates increasingly general models of the harm domain that capture essential environmental properties through efficient sensing architectures, directly informing sensor selection, computer vision algorithms, and environmental modeling systems required for effective harm prevention in autonomous systems deployment.

We validated this integration approach through comprehensive prototype analyses that demonstrate systematic identification of shared structural features across harm domains. Our tripartite conceptual network analysis (Figure 4) revealed how verbs cluster around shared mechanisms—for example, "puncture", "pierce", "stab", and "perforate" all connect through penetration mechanisms to bodily integrity interests.

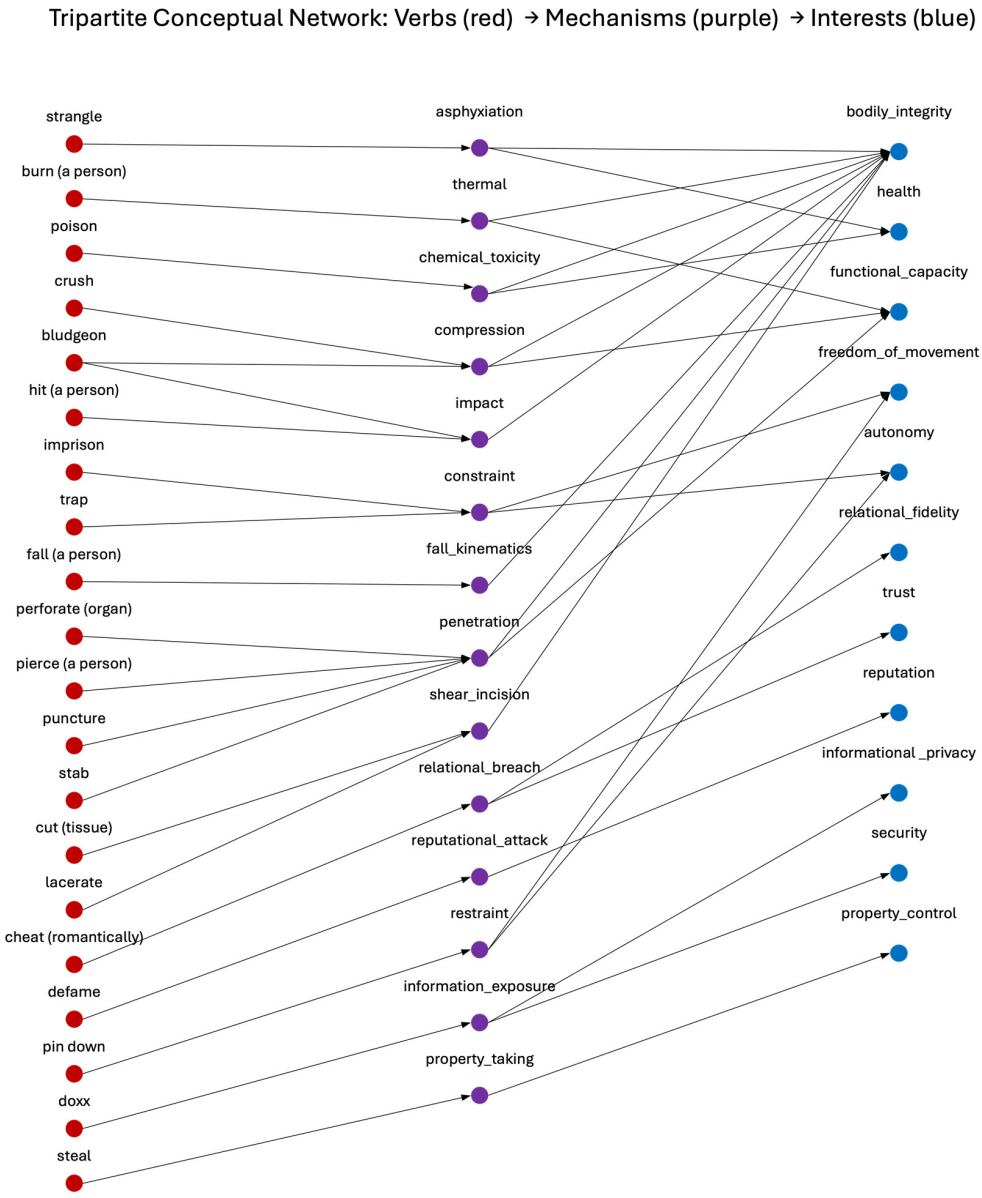


Figure 4. Tripartite conceptual network connecting verbs (red circles) to interests (blue circles) through mechanisms (purple circles). This network exposes how verbs cluster and which mechanisms bridge domains (e.g., puncture/pierce/stab via penetration to bodily_integrity; pin down/trap/imprison via restraint/constraint to freedom_of_movement/autonomy, etc.).

This structural mapping enabled the consolidation of overlapping monitoring requirements into unified components (Figure 5). For example, the Penetration Monitor covers all puncture-type verbs through shared features including sharp-object detection, surface-breach recognition, and tissue segmentation capabilities.

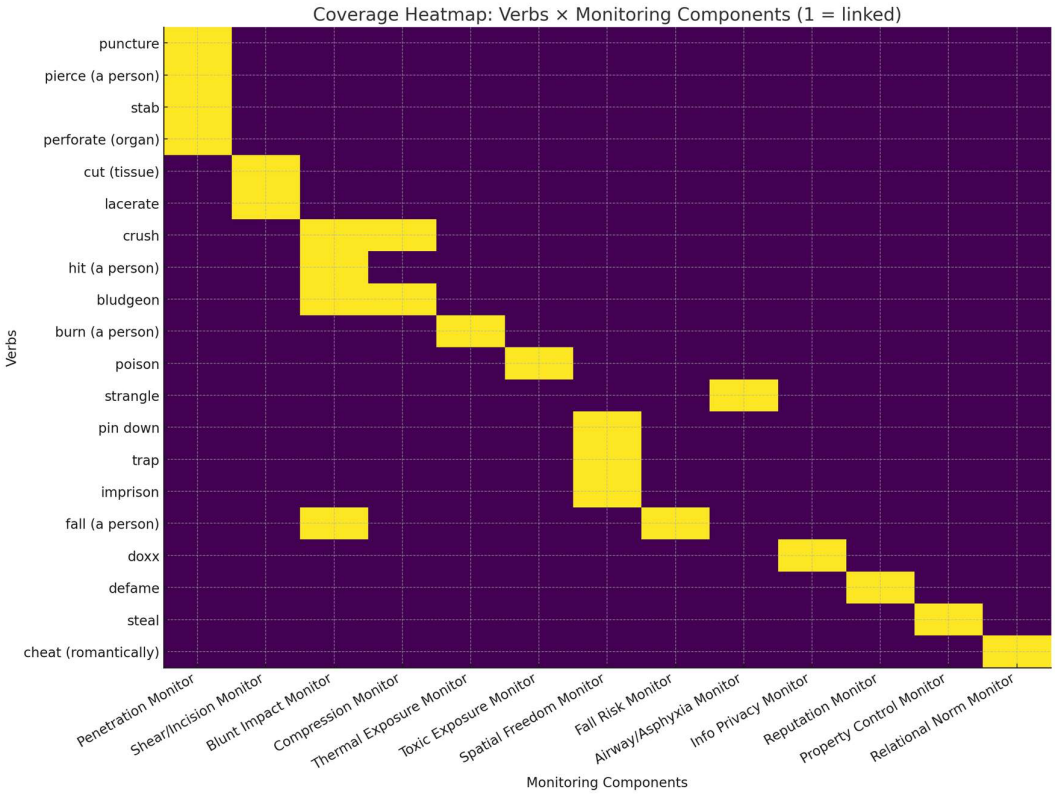


Figure 5. Harm verb-monitor coverage: This matrix shows how many verbs collapse onto the same monitoring modules, cutting perceptual load while maintaining coverage. Yellow indicates which monitoring components relate to which harm verb.

6. Methodological Advantages

The thick harm verb approach generates local models of particular ways of harming that avoid the limitations of both purely top-down and purely bottom-up approaches. Each thick harm verb analysis yields a local model with well-defined scope—the model applies to the range of contexts in which the harm verb is appropriate. This locality prevents overgeneralization while maintaining sufficient specificity for practical applications. These local models address the black-box problem that plagues many machine learning approaches to ethical reasoning. Each analysis makes the underlying inferential relations explicit, showing exactly why certain actions count as harmful and which factors autonomous systems should track to avoid causing harm. When an autonomous system uses ontologies constructed in this way to reason about potential harms, it can provide clear explanations referencing the specific mechanisms, conditions, and interests involved.

The local models can be integrated into a comprehensive ontology through a bottom-up process that identifies common components across analyses. Shared mechanisms (like piercing), common conditions (like vulnerability), and overlapping interests (like bodily integrity) provide connection points between different local models. This integration process allows for optimization that eliminates redundant information while preserving the explanatory structure that makes the reasoning transparent. This approach enables incremental development of harm ontologies tailored to specific applications.

Robotics applications might focus initially on physical harm verbs, while chatbot development might prioritize psychological harm verbs. Each domain-specific ontology can be developed separately and integrated later as needed, providing practical flexibility while maintaining theoretical coherence. The resulting framework combines the benefits of both top-down and bottom-up strategies while avoiding their respective limitations. Like top-down approaches, it provides clear theoretical foundations and explicit reasoning principles. Like bottom-up approaches, it builds comprehensive coverage through systematic analysis of specific cases. Unlike either approach alone, it generates both practical guidance and principled reasoning that can adapt to novel situations.

The proposed framework can be operationalized using current LLM capabilities, particularly through constrained decoding techniques that help analysts generate outputs that conform to structured schemas. This implementation approach treats the current proposal as a concrete method for leveraging existing AI capabilities as analytical assistance tools rather than as future research. The availability of these capabilities transforms thick harm verb analysis from theoretical exercise into practical methodology for discovering features that can inform harm ontology construction. The implementation strategy builds on the recognition that LLMs already contain substantial semantic knowledge about harm concepts embedded in their training data. However, this knowledge remains largely implicit and unstructured, making it difficult to extract systematically without analytical methods. Thick harm verb analysis provides a method for extracting and organizing this knowledge into explicit feature specifications that preserve both the mechanical specificity and normative content needed for ethical reasoning.

7. Conclusions

This paper has developed a discovery method for identifying harm-relevant features through the systematic analysis of thick harm verbs that encode information about harm. By treating these semantically rich linguistic concepts as sources of information about objects, properties, mechanisms, and conditions that determine when actions constitute harm, we have shown how to bridge the gap between abstract ethical principles and practical implementation requirements. The approach discovers features through local analyses that can be integrated into comprehensive ontologies enabling sophisticated harm reasoning across diverse contexts.

The methodology makes three key contributions. First, it provides a systematic discovery method for identifying harm-relevant features without requiring comprehensive a priori theoretical frameworks or exhaustive empirical datasets. Second, it demonstrates how analyzing thick harm verbs to extract encoded information yields specifications about causal mechanisms, material prerequisites, underlying interests, and contextual conditions—the four dimensions of information that autonomous systems require for harm recognition. Third, it establishes that the discovery method can be operationalized using LLMs as analytical assistance tools, with human analysts applying the method to generate candidate feature specifications that domain experts can validate and refine.

The analytical framework we have developed extracts four types of essential information from thick harm verbs: causal mechanisms, material prerequisites, underlying interests, and contextual conditions. This comprehensive analysis discovers the detailed features that autonomous systems require to distinguish harmful from harmless actions across diverse contexts and to provide transparent explanations for their reasoning. The taxonomic classification of harm verbs according to their contextual sensitivity patterns further provides foundations for implementing sophisticated harm recognition that accounts for consent, intention, and circumstance.

Our prototype implementations demonstrate that this discovery method can be operationalized using current LLM capabilities as analytical assistance tools. Through systematic LLM-assisted analyses, we successfully generated candidate harm ontology components including thickness orderings, mechanistic specifications, material requirements, and interest mappings that can inform autonomous systems design after human validation and refinement. The resulting integrated feature specifications provide initial guidance for sensor selection and computer vision algorithm design, along with normative frameworks that can inform ethical reasoning about interventions and tradeoffs.

This work contributes to the broader project of developing autonomous systems that can operate safely and ethically in human environments. By providing a systematic discovery methodology for identifying features required for harm-sensitive reasoning, the thick harm verb approach offers a pathway from abstract moral principles to practical ethical systems. Future work should focus on expert validation of the discovery method's utility, systematic assessment of cross-analyst reliability when applying the method, integration with existing autonomous systems architectures, and extension to additional harm domains beyond the physical cases emphasized in this initial analysis. Finally, empirical testing should examine whether autonomous systems using ontologies constructed through this discovery method actually perform better at harm recognition and response compared to alternative approaches.

Author Contributions: Conceptualization, J.J. and A.M.G.; methodology, J.J.; writing—original draft preparation, J.J.; writing—review and editing, J.J. and A.M.G.; visualization, J.J. and A.M.G.; project administration, A.M.G. All authors have read and agreed to the published version of the manuscript.

Funding: J.J.: The author's position is supported through a grant from the Mellon Foundation through the Georgetown-Howard Center for Medical Humanities and Health Justice. A.M.G.: This work was supported by: JHU Discovery program, APL Janney program, JHU Institute for Assured Autonomy Existential Grant, APL National Security Analysis Department.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare that there are no competing interests relevant to the content of this article.

Generative AI Statement: During the preparation of this manuscript/study, the authors used ChatGPT-5 Pro for the purposes of prototype implementation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

References

1. Winfield, A. *Robotics: A Very Short Introduction*; OUP Oxford: Oxford, UK, 2012; 175p.
2. Awad, E.; Levine, S.; Anderson, M.; Anderson, S.L.; Conitzer, V.; Crockett, M.J.; Everett, J.A.C.; Evgeniou, T.; Gopnik, A.; Jamison, J.C.; et al. Computational ethics. *Trends Cogn. Sci.* **2022**, *26*, 388–405. [CrossRef] [PubMed]
3. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: New York, NY, USA, 2009; 275p.
4. Müller, V.C. Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy [Internet]*; Zalta, E.N., Nodelman, U., Eds.; Fall 2025; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2025. Available online: https://plato.stanford.edu/archives/fall2025/entries/ethics-ai/ (accessed on 15 December 2025).
5. Hutler, B.; Rieder, T.N.; Mathews, D.J.H.; Handelman, D.A.; Greenberg, A.M. Designing robots that do no harm: Understanding the challenges of Ethics for Robots. *AI Ethics* **2024**, *4*, 463–471. [CrossRef] [PubMed]
6. Mackworth, A.K. Agents, Bodies, Constraints, Dynamics, and Evolution. *AI Mag.* **2009**, *30*, 7. [CrossRef]

7. Greenberg, A.M. Deciding Machines: Moral-Scene Assessment for Intelligent Systems. In *Human-Machine Shared Contexts [Internet]*; Lawless, W.F., Mittu, R., Sofge, D.A., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 135–160. Available online: https://www.sciencedirect.com/science/article/pii/B9780128205433000067 (accessed on 30 December 2024).
8. Greenberg, A.M. A schema for harms-sensitive reasoning, and an approach to populate its ontology by human annotation. In *Putting AI in the Critical Loop [Internet]*; Dasgupta, P., Llinas, J., Gillespie, T., Fouse, S., Lawless, W., Mittu, R., Sofge, D., Eds.; Academic Press: Cambridge, MA, USA, 2024; pp. 265–278. Available online: https://www.sciencedirect.com/science/article/pii/B9780443159886000066 (accessed on 30 December 2024).
9. Williams, B. *Ethics and the Limits of Philosophy*; Harvard University Press: Cambridge, MA, USA, 1985; 297p.
10. Cartwright, N. Causation: One Word, Many Things. *Philos. Sci.* **2004**, *71*, 805–819. [CrossRef]
11. Väyrynen, P. Thick Ethical Concepts. In *The Stanford Encyclopedia of Philosophy [Internet]*; Zalta, E.N., Nodelman, U., Eds.; Spring 2025; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2025. Available online: https://plato.stanford.edu/archives/spr2025/entries/thick-ethical-concepts/ (accessed on 29 August 2025).
12. Greenberg, A.M. Antisera for Misconceptions Around Agentic Technology. In Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 5–7 May 2025; pp. 1213–1216. Available online: https://ieeexplore.ieee.org/document/11050537 (accessed on 15 December 2025).
13. Stanton-Ife, J. The Limits of Law. In *The Stanford Encyclopedia of Philosophy [Internet]*; Zalta, E.N., Ed.; Spring 2022; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2022. Available online: https://plato.stanford.edu/archives/spr2022/entries/law-limits/ (accessed on 15 December 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.