A Formal Language of Ethics and Causation

By

Vu Le

* * * * * * * * *

Submitted in partial fulfillment

of the requirements for

Honors in the Department of Computer Science

UNION COLLEGE

June, 2024

Abstract

VU LE A Formal Language of Ethics and Causation. Department of Computer Science, June, 2024.

ADVISOR: TJ Schlueter and Marianna Bergamaschi Ganapini

It is no longer an exaggeration to claim that Artificial Intelligence carries a lot of risks. With its increasing application in critical social contexts such as healthcare and self-driving car, it is the more difficult for us to keep them under control. So what can we do about this? Certainly, chasing after its uses and cleaning up its messes are no longer feasible. Theoretical debates about its risky consequences are not enough. So, instead of us humans handling the risks, what about we let the AI handle the risks themselves? This paper explores a possibility of a moral AI agent that can deliberate on its own and make ethical decisions in real-life scenarios. The overarching goal of this project is to design a logical formalism for rigorously encoding ethical situations and reasoning about them. Since causal reasoning is an integral part of ethical judgments, a formalism of ethics must also incorporate causation. Extending from prior works, this research presents and implements a formal framework of ethics and causation, which can give satisfying justification for decisions in ethical dilemmas.

Contents

List of Figures

List of Tables

1 Introduction

It is no longer an exaggeration to claim that Artificial Intelligence carries a lot of risks [28, 31]. With its increasing application in critical social contexts such as healthcare and self-driving car, it is the more difficult for us to keep them under control. So what can we do about this? Certainly, chasing after its uses and cleaning up its messes are no longer feasible. Theoretical debates about its risky consequences are not enough [54]. So, instead of us humans handling the risks, what about we let the AI handle the risks themselves? This is the goal of computational ethics: to build an AI that is capable of reasoning and making ethical decisions by itself.

Interests in Computational Ethics have only been growing in the recent decade. While there is a wide variety of approaches, there has been skepticism as to how ethics are to be computed [30]. The overall goal of this thesis is to explore the possibility of a formal language for rigorously encoding and reasoning about ethical scenarios. One requirement for such a formal language is the concept of causation. This is because causal reasoning plays an integral part in our ethical judgments. Blame and responsibility attribution, for example, requires tracing back the causal chain of actions to determine the person responsible. However, little attention has been paid to causation in ethical modeling. And if it does, the underlying theoretical framework of causation is still too simplistic to handle more complex cases. In this project, I introduce a framework of ethics and causation, which incorporate the causal theory of Necessary Element of Sufficient Set (NESS), developed by Wright and commonly used in legal practices.

The paper is structured as follows. I first examine the general challenges in computational ethics, specifically in the rule-based, logical approach adopted in my research (Section 3). From these pitfalls of ethical modeling arise certain standards/requirement for an ideal ethical formalism. One of which is causal reasoning, which I motivate in section 4. Next, I shift my attention to the issue of causation. In section 5, I consider influential theories of causation and their potential application in ethical modeling, before introducing the NESS theory of causation. Finally, I present a framework for causal and ethical reasoning proposed in a line of prior research as well as my own contributions [47, 11].

2 Related Work

Works in computational ethics are diverse in approaches, both computationally and philosophically. The general implementation paradigm can be classified into bottom-up, top-down, or hybrid approaches [60, 65]. Different ethical frameworks have been proposed and codified, such as Kantian deontology [10, 11, 44], the Doctrine of Double Effect [25, 41, 13], reflective equilibrium [3, 4], utilitarianism [1, 2, 11, 6]. This

section briefly reviews some of the prior works, and at the same time situating my approach in this area of research.

Bottom-up approaches is founded upon learning-based paradigms, such as artificial neural networks, reinforcement learning, and other machine learning techniques. These approaches assume that ethics can be learned with enough dataset and training. Examples are Delphi and Moral Machine experiment [29, 8]. Data are collected from human decisions in ethically-charged scenarios, which are then used to train the AI to respond to similar ethical situations. Though these experiments might be valuable to understanding people's behavior patterns in making ethical decisions, they cannot be used as moral prescriptions guiding the AI agent's action in the real world. This is because the training dataset are inherently filled with human biases, thus defeating the purpose of AI ethics. Furthermore, its ethical judgments lacks explainability, losing many nuances in ethical reasoning. Most importantly, they make fatal assumption of how ethics are to be computed [59].

My research approach aligns with the top-down tradition, which focuses on representation of ethical principles and automated reasoning. This methodology assumes that we have sufficient knowledge about the world as well as all the principles needed to make an ethical judgment. While this is a substantial assumption, it is a trade-off for making ethical decisions more explainable thanks to its explicit reasoning mechanism. Implementation techniques often involve logic programming [17]. Specifically, this project advances the ethical framework that has been proposed in a line of prior research [13, 12, 11, 47]. The goal of this framework is to design a logical formalism for representing ethical principles and faithfully encoding ethical scenarios.

3 Challenges in Ethical Modeling

The task of ethical modeling demands a rigorous description of the causal world. From the point of a practical application, ethical reasoning, as a feature-module, is to be used as a supplement for an AI agent which is capable of taking input from the environment, generating possible actions, and evaluating the consequences according to its causal understanding of the environment. In this vision, the ethics module should be able to make use of these information relevant to construct an ethical scenario, from which an ethical value can be calculated by some moral principles (of choice) and then returned to the agent's decision procedure.

But applications aside, it is almost impossible to perform ethical reasoning without an understanding of the causal world. For otherwise, one would not be able to properly trace back to the rightness or wrongness of an action, even if the utility of its consequence has been readily calculated. The same goes for deontolog-

ical theories. Without some underlying assumptions of causation, one would be forever stuck within the realm of intentions (or maxims), of which the good and bad could never be translated to actions in the real world.

An understanding of causation, it seems, is crucial for some aspects of ethical reasoning. In modeling ethical scenarios, however, it is imperative. In this section, I first lay out the worries and challenges of ethical modeling for rule-based approaches as well as for computational ethics in general. In addressing these concerns, I wish to motivate the goals, values, and responsibility in this work of ethical modeling, and a need for causal reasoning within ethical models. Exploring the problems of causation through the lens of moral dilemmas, I believe, will be fruitful not only to formalizing ethics, but also to the enterprise of causal modeling as a whole.

3.1 Problems of ethical formalization

Works in computational ethics mainly employ thought experiments to test their formalization of ethical theories. A thought experiment is usually in the form of an ethical scenario, such as the trolley problem and its many variations [20]. These ethical scenarios, though simple as they are purposefully designed, are non-trivial to formalize.

A common pitfall in this enterprise of computational ethics, especially in the rule-based representation tradition, is the gross oversimplification of the problem formalization, as rightly noted by Sarmiento et al. [47]. The problem formalization, when not taken with care, not only will fail to preserve the nuances of each thought experiment, but also make the process of ethical reasoning appear ad-hoc and as if an obvious extension of the formalization itself. This defeats the purpose of a thought experiment, which aims to expose the implicit considerations underlying our day-to-day moral judgments by provoking the "irk" in our moral intuitions.

Thus, the value of ethics modeling does not lie in the mere conclusion of rightness or wrongness with respect to some moral theories, but rather in the number of moral propositions derivable from the same formal structure of reasoning that capture the many (conflicting) intuitions of ours. Needless to say, though there might be no right or wrong answers for an ethical dilemma, there is still much value in trying to study and formalize the structure of reasoning behind our moral intuitions.

Let us first artificially construct a formalization for one variation of the trolley problem, so as to demonstrate an (extremely) bad example of ethical modeling. The shared circumstance of the problem is as follows: a trolley is running on a track, headed towards the five people who would be killed upon collision. In the bystander case (Figure 1a), you have the option to pull the switch, diverting the trolley onto a side

track and thus killing a bystander who happens to be nearby.

$$running(trolley, main\_track) \leftarrow watch. \tag{1}$$

$$running(trolley, side\_track) \leftarrow pull\_switch. \tag{2}$$

$$dead(5) \leftarrow running(trolley, main\_track). \tag{3}$$

$$dead(1) \leftarrow running(trolley, side\_track). \tag{4}$$

$$intentional\_killing \leftarrow pull\_switch. \tag{5}$$

$$intentional\_saving \leftarrow pull\_switch. \tag{6}$$

$$wrong(utilitarianism) \leftarrow dead(5). \tag{7}$$

$$wrong(kantian) \leftarrow intentional\_killing. \tag{8}$$

$$1\{watch; pull\_switch\}1. \tag{9}$$

Rules 1, 2 describe the consequence of the agent's two possible actions: either interfering (by pulling the switch) or not (by watching things be). If pulled the switch, the trolley will run on the side track; otherwise, it remains on the main track. Rules 3, 4 are deadly consequences of the running trolley, five deaths if on the main track and only one death if on the side track. If death is of the mass, it is wrong by virtue of utilitarianism (rule 7); whereas in Kantian view, intention makes up the intrinsic wrongness of the act (rule 8). Within our context, the intention of pulling the switch, as rule 5 dictates, is to kill the one person on the side track, while saving the other five (rule 6). Lastly, rule 9 is called a "choice" rule. Here, the agent must choose exactly one act; and for each act, we have a separate model calculating all the consequences according to the aforementioned rules. As a result, this logic program yields two models, one describing the outcomes of the watching act, which is wrong by utilitarianism, and one describing the outcomes of the switch-pulling act, which is wrong by Kantian.

This formalization seems reasonable enough. After all, it agrees with all of our intuitions, by producing the "right" answer for each of the chosen theories. Our immediate reaction to this particular formalism, I hope, is that of indifference and unimpressiveness - ok, sure, where do we go next? It is hard to conceive, with this program *alone*, how an AI agent could find itself in just any arbitrary situations and articulate with such precision that its act of pulling the switch essentially amounts to intentional killing. But when given these rules *a priori* that readily describe the situation itself, the moral consequences are immediately obvious, if not trivial. This is because the reasoning structure is already embedded in the crafted rules themselves - the hard work has already been done by the programmers, the AI only needs to "automate"

it. The gravest issue, however, is that it introduces a risk of programmers unknowingly injecting their own biases into the system: on what grounds, one could ask, do we attribute an intention, an unobserved object, to an action, an observable object?

Without a common language of formalism, we may be subsequently led further down the hellhole of moral relativism, where the ethics of each AI is whatever rules of choice it has been coded with; one could imagine a future, in which (humankind has long gone extinct) the robots would be arguing with each other to eternity, by merely outputting propositions without ever reaching a point of dialectic reconciliation. It is in this sense that the formalization may appear ad-hoc, subjective, and lack applicability in the general situation.

Granted, this problem is not exclusive to ethics, but to the field of knowledge representation and reasoning in general. Three fundamental challenges, amongst many, in this area of research [16]:

- What knowledge does a system need to have in advance, as opposed to what can be acquired by observations?
- What is the language for representing and reasoning with the background knowledge and observations?
- What kind of semantics governs the updating of knowledge, given new and possibly conflicting observations?

These questions remain open. But we may, with good reasons, remain positive in this enterprise: by asking these questions in a more confined context of ethical modeling, we will be gradually chipping away the more general questions. At the very least, it can be acknowledged that the design of problem formalization is not to be taken for granted, and that there are ”good” formalism and ”bad” formalism. Although it may not be directly obvious or explicable how a problem formalization has been grossly oversimplified, I believe that by engaging deeply in the ethical literature, one could arrive at certain standards for what makes a formalism good.[1]

3.2 Some standards for an ideal ethical formalism

Before moving forward and potentially risking oneself of frivolous assumptions, I propose taking a brief rest to lay out the following 3 questions, which I dare not claim to be fundamental to the bigger business, but merely mildly interesting for my task at hand:

[1] But formally speaking, what makes the ”good” of a good formalism? - jestly one asks. Beware of falling in a vicious regress! for our very purpose is to formalize the good in a formalism.

1. How does the choice of problem representation affect the moral conclusions?

2. What are the fundamental ontology to sufficiently formalize an ethical problem?

3. Does there exist a formal/universal language of ethics? Can ethics be even formalized at all?

We are, first of all, concerned with mitigating subjective biases unintentionally polluted into ethical models. The biggest challenge for computational ethics, I believe, is the worry that there is "no ground truth about what ethical principles to code" [43]. While it is true that there are conflicting norms and many competing moral theories, it does not necessarily amount to subjectivism, which is the view that all ethics is merely a matter of opinions;[2] for if it is, *as a matter of fact*, merely opinion, there should not have been any genuine ethical disagreements in the first place. If ethical models are contaminated with our own implicit biases, it will defeat its own purposes, losing the values of ethical reasoning and tempting people more towards subjectivism. Thus, question 1 asks us to take caution in formalizing ethical problems.

One solution is to modularly separate facts (description of the problem) from values (norms or prescriptions). This is because there is a substantial gap between statement of facts (what *is* the case) and statement of values (what *ought* to be the case) [46]. One cannot simply derive *ought* from *is*. Yet it is not easy to distinguish evaluative statements from purely descriptive statements. This can potentially lead to the naturalistic fallacy of reasoning, making a big leap from the is to the ought without sufficient justification, as Hume remarked [62]:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surprized to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, it is necessary that it should be observed and explained; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.

The best one can do is to take precaution and make explicit the fact-value distinction in a formalism. The scenarios description should be about strictly factual matters and contain no evaluative terms at all. The gap between facts and values is mediated by the moral theories. This not only minimizes unintentional infection of values on the modeler's part, but also allows for multiple ethical judgments to be drawn, using different

[2]It is, admittedly, a little too charitable to legitimately use the word "view," for a true subjectivist would say that subjectivism, in virtue of being an ethical assertion itself, is after all a mere opinion.

moral theories, from a single ethical scenarios. Many ethical frameworks, though different in approaches, share an explicitly modularized separation between situations' description and normative evaluation [7, 11, 14].

Any rules must also be adequately grounded in well-developed theories to minimize leap-of-faith assumptions; for example, rule 5 and 6 should be backed up by a theory of action and intention. Additionally, the rationale (traces of inference) must be outputted along with the conclusion. Not only will it be explainable, but also provide valuable insights to the algorithmic "movements" of inference underlying each moral theory. A mere judgment without rationale, as we have seen in the Delphi Experiment, is certainly more challenging for us to make value out of it [29].

One is thus invited to ponder the possibility of a bias-free ethical language. This daydreaming must, however, be counterbalanced by the fact that one's moral intuitions are susceptible to particular presentation of moral scenarios [42]. Even if normative terms are completely pruned out from purely factual matters, it is possible that a particular factual representation of a scenario can affect one's ethical judgments. How does it affect? If our moral intuitions are highly sensitive to a factual presentation, then it may be helpful to study how it get tweaked around by manipulating the representation in a particular way. Let us then consider the role of moral intuitions in an ethical formalism. In the presence of many conflicting intuitions, should we simply disregard them as remnants of our deeply-rooted prejudices [58], or treat them seriously as sense-data for uncovering the underlying principles [32]?

I offer no arguments for this question; rather, I shall assume the following view: every intuition derives its appeal from one's cognition of certain morally relevant features, and there exists an algorithmic process (of rationalization) to reconstruct the intuition. Different intuitions are elicited by focusing on certain morally relevant features of a given scenario, while possibly excluding others. The abstraction of features can be induced by a particular presentation of the scenario, which may promote some features to be more salient. We may or may not be conscious of the features, and may have implicitly assigned different moral weights to each. Intuitions, however conflicting, each exhibit an underlying rational structure; one can thus approximate any intuition at will by zooming in its underlying feature-constituents and assimilating its reconstructive movement step by step.

Therefore, a formalism, if it is not to be biased, must be independent of any particular intuitions, but at the same time, establish a fundamental structure such that any moral intuitions can be rationally reconstructed from its machinery. By extension, the formalism should not be too rigidly dependent on any one moral theory; it should instead aspire to be meta-ethically neutral and expressive enough that any moral theories can be faithfully formalized using its language, and all the reasoning characteristic of each theory can be reflectively derived therein.

We are now ready to investigate such a structure. Question 2 is concerned with the fundamental concepts underpinning all moral intuitions and reasoning. They include descriptive features of a situation that is relevant to our moral deliberation, such as agency, intention, action, number of moral patients involved, act/omission, etc. More normatively-charged concepts such as treated as a means, treated as an end, doing/allowing harm, etc. must be explicitly derivable. For example, in the Doctrine of Double Effect, the critical distinction is whether a bad effect is used as a means or merely foreseen side effect for some good end, which can be analyzed using counterfactual reasoning [25]. In their works, Govindarajulu and Bringsjord also introduced the **Deontic Cognitive Event Calculus**, which formalizes important intensional notions including knowledge, belief, and desire.

The ideal formalism, therefore, must outline all the fundamental concepts, enforcing an exhaustively descriptive representation of any ethical scenario, and able to capture all morally significant considerations employed in our ethical reasoning. In proceeding forward, however, I must restate our assumption that there exists some form of homogeneous structure circumscribing all of our moral reasoning. This is a substantial assumption, which requires substantial justification to completely dispel the worries of question 3. But answering this question needs not be a prerequisite to our task of modeling ethics; rather, I believe attempts in designing such an ethical language can help illuminate this question, potentially having implications to a number of meta-ethical problems.

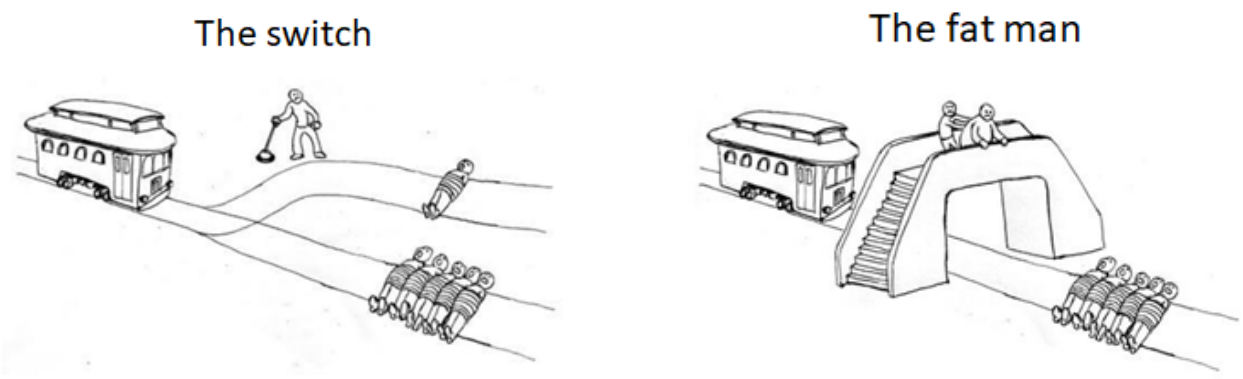
In sum, I believe these worries are not only pertinent to logical, rule-based approaches, but also to the larger field of computational ethics. Indeed, there are concerns as to the legitimacy of the overarching methodologies in this research area [30]. Fundamentally, they are questions of whether ethics can be computable at all, and if so, how should they be computed?[3]

4 Ethics and Causation

After laying out the general challenges of ethical modeling, I shall argue for the following: in order to model ethical scenarios descriptively, we first need a rigorous formalism for modeling the causal world. At the outset, this is because an understanding of causation has many implications to our ethical judgments [48, 37]. In this section, I examine potential applications of the concept of causes in ethical modeling and highlight issues open to debate, thus requiring cautious philosophical positioning before deploying the formalism.

In blaming or praising someone for an outcome, it seems that the agent must have causally contributed to that outcome. However, although causal responsibility is used to explain moral responsibility, it does

[3]How *ought* ethics to be computed? Is this an ethical statement? If so, how could the statement itself be computed?



(a) Pulling the lever

(b) Pushing the fat man

Figure 1: Two variations of the trolley problem

not necessarily imply moral responsibility. Moral responsibility for an outcome requires, in addition to causal contribution, certain expectations or duties for an agent, within their causal powers, to bring about or prevent that outcome [49]. For example, suppose that I was coerced by an evil hypnotist to set up a bomb, killing thousands of civilians in the city. My action was a part of the causal chain of events leading to the catastrophe, but certainly I was not deemed to be morally responsible for it. For my act was not free, but predetermined by a force beyond my causal power. Thus, while intricately related, moral responsibility must be distinguished from the notion of causal responsibility. It is critical for an ethical formalism to model these two concepts separately.

The killing/letting die distinction is invoked to explain the moral difference between active euthanasia (giving terminal patients lethal doses of drug) and passive euthanasia (withholding medical treatment critical to sustaining the patient's life). If there is no moral difference between killing and letting die, then active euthanasia is as morally permissible as passive euthanasia, and failure to prevent millions of children's deaths from starvation is as impermissible as donating them poisoned food. This moral difference could be grounded in causal terms, one way of which is with a related distinction of action/omission [48, 61]. Roughly put, killing is an action, hence cause of death, whereas letting die is omission, hence not considered as causes of death. However, whether omission can be proper cause is a highly debated issue in the metaphysics of causation. For example, the death of a house plant is resulted from my failure to water it, so I am morally responsible for its death. And given that moral responsibility arguably implies causal responsibility, I must have had a causal role in relation to the death of the houseplant. But theoretical objections aside, common sense suggests that there is enough difference between action and omission that they should be modeled separately. These notions can be treated as agential volition (the making of decision), which allows for capturing the agent's causal responsibility in the absence of events [12].

Causal reasoning is employed, not only in consequentialism, but also in deontological theories such as Kantian and Doctrine of Double Effect. For example, the *means-end* reasoning, central to Kant's second formulation of the Categorical Imperative, can be formalized in the causal agency model, an extension of the structural equation framework [10]. In their formulation, a person is treated as a means if affected by my action (or its consequence) in some way, and treated as an end if benefited from the intended goals of my action. An action is impermissible if some affected person is treated as a means without at the same time being treated as an end. A closely related distinction in the Doctrine of Double Effect is the *means-side effect*, which explains the moral difference of causing harm as an unintended and foreseeable side effect in promoting some good end versus causing harm as a means to bring about the same good end [36]. The means-side effect, similarly, can be rendered in causal terms [25].

A common mechanics in these formalisation is their counterfactual analysis on the concept of means. Here I will focus on a simplistic counterfactual account as well as its application in the Doctrine of Double Effect.

Definition 1 (Counterfactual account of means-side effect) *Let A be action, P be the person involved in the action, E be the outcome. Let c be the following counterfactual: "Had A occurred without P, E would have not occurred". P is a means to E if and only if c holds true.*

At first glance, this account of counterfactual is enough to explain the difference in the pair of trolley cases: pulling the lever and pushing the fat man (Figure 1). The fat man is a means for the survival of the five, because had it not been for the fat man, the five would have not survived. Contrarily, the one person on the side track is not a means, because had it not been for him, the group of five would still have survived regardless, for the trolley would have already been diverted to the side track without any casualties. The death of the one person is merely a side effect.

Note however that the distinction on this account does not necessarily amount to a normative evaluation, for it lacks intentionality of the action; the person is descriptively a means, but might not be *treated* as a means. Still, it yields counter-intuitive intuitions. Consider an extension of the fat man case, in which there are two fat men on the bridge, each of them is sufficient to stop the trolley when crashed, and I have equal urges of pushing either of them to save five; call this the two fat man variation. In actuality, I pushed the fat man 1. Was the fat man 1 a means to save the five people? No, because without the fat man 1, I would have pushed the fat man 2 and the outcome would not have changed.[4] The same would go for the fat man 2 had I actually pushed him. Thus, neither of them are a means for the survival of the five.

[4]To reduce the number of people of involved, the original fat man variation can be minimally modified such that I and the fat man are both... fat, and we have equally altruistic urge to push each other and save the five people. Thus, if I do not push him, he will push me. A side effect of this variation is that it can potentially distract us from the original problem, as saving of myself is now also worthy of consideration.

However, if we group up both the fat man 1 and fat man 2 into the counterfactual, we would have 'Had it not been for either the fat man 1 or fat man 2, the five people would not have survived.' This is true. Each of the fat men is individually not a means, but the disjunction of their presences - (fat man 1 $\vee$ fat man 2) is a means to the outcome. It remains to be shown how far we can generalize this answer, which depends on the question whether disjunctive facts can be causes [50].

Having fully motivated the significance of causation in ethics, I will now turn my attention wholeheartedly to the concept of causation as well as its applications in contemporary causal modeling.

5 Actual Causation

Philosophers distinguish two notions of causation: *type causation* (or general, law-like causality) and *token causation* (or singular, actual causation) [21]. Type causation is concerned with the general relation between types of events, such as "Dropping objects cause them to be on the ground." In contrast, token causation relates two particular events, for instance, with the statement "The fact that I dropped the ball caused it to be on the ground."

Reasoning about actual causation is less useful to making future predictions, than to determining the causal relation of events that happened in the past. As such, it plays a critical role in the law, especially in attributing blame and responsibility. There are cases in which all the causal laws of relevance are known, yet the problem of identifying the "causes" remain obscure. For example, a drunken driver was driving into an intersection, the car's brakes were broken, while another car was carelessly running through the red light, resulting in a fatal accident for the drunken driver. The question is who (or what) is to blame? Was it the driver's drunkenness, or the faulty brakes, or the traffic light violation of the other car, or some combination of the mentioned factors? Type causation and actual causation is deeply intertwined. The primary focus of this section (as well as the whole project) is the notion of actual causation. I first review briefly some prominent approaches in actual causation, before introducing the account of Necessary Element of Sufficient Set (NESS-causation) [64].

5.1 Counterfactual reasoning

A common line of causal reasoning used the law is called the *but-for* test. A is a cause for B, if, but for A, B would not have occurred. The test fails for some cases of causation. For example, the story of Suzy and Billy goes as follows. Suzy and Billy both throw rocks at a bottle. Both have perfect aim and are exerting enough force to shatter the bottle. However, Suzy's rock got to the bottle first, shattering it before Billy

managed to. Was Suzy the cause for the broken bottle? Applying the but-for test, had Suzy not thrown, the bottle would have been shattered anyway due to Billy's throw.

This is an example of preemptive causation, one of many problematic cases of *over-determined* causation as defined by Wright [63]:

> cases in which a factor other than the specified act would have been sufficient to produce the injury in the absence of the specified act, but its effects either (1) were preempted by the more immediately operative effects of the specified act or (2) combined with or duplicated those of the specified act to jointly produce the injury.

The but-for test captures notion of causal efficacy, appealing to the intuition to single out the event that makes a *difference*. A cause is necessary for the effect to occur, or the effect depends on the occurrence of the cause. This understanding of causal necessity generates the paradox of *Conditio Sine Qua Non* in cases of overdetermined causation [51]. When there are multiple causes relevant to the scenario, each is sufficient for the effect, but neither is necessary.

The *counterfactual* reasoning structure has its roots in Hume's analysis of causation [55]:

> we may define cause to be an object followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed

This little quote is actually packed with two different understandings of causation: regularity theory of causation and counterfactual theory of causation. I will return to the former later. The latter, formulated after the "in other words" part, has been developed into more sophisticated forms in contemporary counterfactual theories, which aim preserve the notion of dependence/necessity in causal relations. First elaborated by Lewis with possible world semantics [35], it is eventually formalized in the dominant HP approach of actual causality, with the idea of structural equation models (SEM) [26]. This allows for a more nuanced analysis of causal dependence - *de facto dependence* [40]:

> E de facto depends on C just in case had C not occurred, and had other suitably chosen factors been held fixed, then E would not have occurred.

The key for this analysis is interventionism. Given a structure equation model of a situation, in order to analyze the causal relation between two variables C and E, in addition to the value (occurrence or non-occurrence) of C, we would have to "intervene" with other variables in model, and then observe downstream the value of E. If there is such an intervention (a forced assignment of other variables' value), then there is a sense in which E can be said to depend on C.

This method has proved extremely useful in experimental sciences: tweaking around the context in order to find the causal influence of an event/object. However, in daily life situations, this analysis can get awkward real quick, due to the fact that intervention sort of violates the natural order of events. Going back to the rock-throwing example, Suzy, rather than Billy caused the bottle to break, due to the following counterfactual "if Suzy had not thrown, and Billy's rock had still somehow failed to strike the bottle, then the bottle would not have broken" [40]. Indeed, if analysis of causal dependence *depends* on finding a suitable intervention, then one is still left to find such an intervention that is reasonable to the her ordinary intuition.

5.2 NESS account of causation

The regularity theory of causation rests on the idea of a constant conjunction between cause and effect. Cause and effect are mere instantiation of a regularity, which can be observed and habitually inferred in the natural association of events. In contrast to the counterfactual understanding, there is no necessary relation between a cause and effect, nor any substantial causal power/efficacy underlying successive sequence of events.

Indeed, the Humean idea of mere association of events, though less strict and metaphysically demanding, is not too helpful in determining causes and effects. Mill refined this regularity account by introducing the notion of jointly sufficiency of a set of factors [5]. According to Mill, a cause is *nomologically* sufficient for its effect. That is to say, there is a natural, law-like regularity that guarantees the cause to be followed by the effect, or that the effect is brought about as soon as the cause occurs. There is no need for further analysis as in the counter-factual account, which demands additional evaluation of the truth of a counter-factual 'but for A, would B have been?'. Here, the notion of necessity is subordinated to sufficiency.

Thus, the problem of identifying causes and effects is reducible to the problem of identifying the law-like regularities that allow an event to function as a "cause" and another as an "effect." What is the structure of such law-like regularities? Each event can be said to have a causal field, in which the event is to be brought about, as an effect [5]. The causal field is organized into disjunction of clusters of conditions. Conditions can be positive, indicating a presence, or negative, indicating an absence of some factor. Each cluster is a conjunction of conditions that are sufficient to bring about the effect. This means that as soon as all the conditions in the cluster are satisfied, the effect immediately follows. However, because there is a plurality clusters, no single one is necessary for the outcome. Each cluster also has to be *minimally* sufficient, which rules out any redundant conditions. Hence, with respect to the cluster, each conjunct is necessary; in other words, if the conjunct is removed, the cluster is no longer sufficient for the effect.

A regularity will have the following general form:

$$(C_{1,1} \wedge \cdots C_{1,n}) \vee \cdots \vee (C_{k,1} \wedge \cdots \wedge C_{k,m}) \longleftrightarrow E.$$

The left hand side is the entire causal field. The causal field as a single collection is necessary for E, hence the $\leftarrow$ direction. The $\rightarrow$ direction refers to the sufficiency of each cluster of conditions. Each $C_{i,j}$ is thus an *insufficient but non-redundant part of an unnecessary but sufficient condition* (INUS condition) for E.

Wright developed a similar account for identifying *actual* causes which can be applied in legal contexts. The NESS test is described as follows [64]:

> a condition c was a cause of a consequence e if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e.

There are 3 requirements for a condition c to be a cause for some event e: (1) c is a member of some sufficient set S, (2) all the conditions in set S were actually instantiated, and (3) S is minimally sufficient for the effect. The notion of minimal sufficiency of a set already implies the necessity for each of its element. Instead of being *strongly* necessary for the outcome itself, here a cause only needs to be necessary for a set of conditions which would bring about the outcome.

Wright distinguishes between causal sufficiency and mere lawful sufficiency. Causal sufficiency requires, in addition to the jointly sufficiency of the conditions, an actual instantiation of all said conditions which would then instantiate the effect. This additional criterion is critical for discerning actual causation from general causation. Suzy's and Billy's rocks are two conditions that are both *lawfully* sufficient for the shattering of the bottle. But only one condition was fully instantiated. Suzy's rock actually came into contact with the bottle, whereas Billy's did not.

Arguably, the inherent flaw in the but-for test is that it conflates the notion of actual cause with lawful cause. Consider another case of preemptive causation. I poisoned you, but the poison would take two days to have effect. Meanwhile, a hitman from afar shot you in the head, killing you instantly. Now we evaluate the the following counterfactual 'Had the hitman not shot you, would you have not been dead?' The answer is no, because you would be dead anyways due to the poison. Indeed, the poison is lawfully sufficient for your death after two days. But in reality, it did not get the chance to be fully instantiated into effect. The evaluation of a counterfactual inherently invokes a general, lawful notion of causation, introducing other non-factual elements in the process. Thus explains the disturbance in the intuition when tested against cases in which there are a *would-be* cause and an *actual* cause.

Definition 2 (NESS-cause) *A condition c is a NESS-cause for an event e if and only if c is a member of some set s*

of conditions, such that:

1. *S is actual: all the conditions in S are instantiated*
2. *S is sufficient for e*
3. *S is minimal*

5.3 Application of NESS test: an example of duplicative causation

Consider the following case of duplicative causation. 5 people A, B, C, D, E participate in an election. President P is elected if 3 out of the 5 voted. As a matter of actuality, A, B, C, D voted - more than enough! Who is responsible for the election of P?

Let us first toy around with the but-for test, had A not voted, P would still have been elected anyways, for 3 people is enough. A's action made no difference and was of no significance.[5] Repeating this process for B, C, D, and it turns out none of them was responsible for the election of P.

Applying NESS test, we find the following four sufficient sets:

1. $\{A, B, C\}$
2. $\{A, C, D\}$
3. $\{A, B, D\}$
4. $\{B, C, D\}$

The sets are actual, that is, all the conditions in each set are satisfied. Each of them is minimally sufficient, because three voters are *just* enough. Since A, B, C, D are elements of *some* set, they are all considered as causes.

We can take a step further and quantify the weight of causal responsibility for each voter. The weighted contribution of a condition c would be the number of sufficient sets containing c, divided by the total number of sufficient sets. In this specific example, the causal weight of the voters are equal to each other and equal to $3/4$. Each of them belong to three sets, out of the total four sets.

The weight calculation has an interesting implication. If a weight of a condition c is exactly 1, it means that c is present in all the sufficient sets. In other words, had it not been for c, none of the sets would be sufficient, and thus the outcome would have not occurred. This collapses into the counterfactual reasoning.

[5] Oddly enough, this kind of reasoning bears resemblance to the "divide-and-conquer", totalitarian scheme. No single individual is of necessity, for the Oppressor can always easily replace them. But collectively, they are indispensable. The Oppressor depends on the entire class of the Oppressed.

Thus, a simplistic version of counterfactual theory can at least be derived from calculating weights embedded in the NESS acount of causation. How far we can derive a counterfactual from a regularity account of causation is an interesting problem that is worth further examination.

6 Causal Language

My work expands on the ACE framework (Action-Causality-Ethics) as developed in a previous line of research [11, 47]. Their approach aligns with a family of research that attempts to formalize causal reasoning in action languages [13, 12, 9, 39, 34, 14]. The entire ACE framework may be divided into four components:

- Causal Rules: knowledge of the causal world
- Event Axioms: general relation between events and states of the world
- Causal Axioms: causal inferences
- Ethical Rules: normative evaluation

This section aims to give a specification of the whole language framework, specifically on the causal dimension: Causal Rules, Event Axioms, and Causal Axioms.

6.1 Action Language Semantics

Action languages are a computational formalism used to model actions and their effects on the world. As such, the ontology of action languages consist of events and states of the world. States of the world serve as conditions for certain events to occur. In turn, the occurrence of events determine the evolution of the states of the world. Several different languages have been developed such as PDDL and action description languages $\mathcal{A}, \mathcal{B}, \mathcal{C}$, each with its own expressive power [27, 24]. These languages belong to the family of logical formalisms used for *commonsense reasoning*, such as Event Calculus, Situation Calculus, Temporal Action Logics, etc [38]. The formalism adopted in this approach aligns closely with the Event Calculus while adapting a number of fragments from other action languages.

The basic building blocks for representing the world are events, fluents, and timepoints. The set $\mathbb{E}$ are events or actions that may occur in the world at some time. The set $\mathbb{F}$ are fluents, representing time-varying properties of the world. The set $\mathbb{T}$ are timepoints, each representing an instant of time. Time in Event Calculus is linear, as opposed to branching like in Situation Calculus. In this formalism, time is assumed to be discrete, $\mathbb{T} = \{-1, 0, \cdots N\}$, where N represents the maximum timepoint, 0 is the initialization time, and -1 is the pre-initialization phase used for bookkeeping purposes.

A fluent is of binary value, either true or false at a specific timepoint. For a fluent $f \in \mathbb{F}$, the fluent literal l can be f or its negation $\neg f$. Let $Lit_{\mathbb{F}}$ be the set of fluent literals; $Lit_{\mathbb{F}} = \mathbb{F} \cup \{\neg f | f \in \mathbb{F}\}$. A complement of a literal l is denoted as $\bar{l}$. A collection of fluent literals constitute a state S at a specific timepoint. Let S_t denote a state of the world at the timepoint t. As such $S_{-1} = \emptyset$, and S_0 represent the initialized truth values of the fluent literals. A state is coherent and complete if every fluent has a determinate values. A partial state of the world will be referred as a coherent but incomplete state - a collection of certain fluent literals of interest.

Definition 3 (State [47]) *A set $L \subseteq Lit_{\mathbb{F}}$ is a state if and only if:*

- *S is coherent: $\forall l \in L, \bar{l} \notin L$.*
- *S is complete: $|L| = |\mathbb{F}|$*

Transition of states occur through events. An event is characterized by its preconditions and effects. Preconditions are a collection of fluent literals (or a partial state of the world) that must be satisfied in order for the event to take place. In turn, the happening of events in effect modifies the truth values of certain fluents. Precondition are written in Disjunctive Normal Form (DNF), which is useful for representing the cluster of INUS conditions. Effects are written in a conjunction of literals. A disjunction of effects can be interpreted as indeterministic effects of event, but within our scope, we will limit ourselves only to deterministic events.

Definition 4 (Preconditions and Effects of Event) *For all event $e \in \mathbb{E}$:*

- *The precondition $\mathcal{P}$ is a propositional formula in DNF.*
  $\mathcal{P} = (l_{1,1} \wedge l_{1,2} \wedge \cdots l_{1,n}) \vee \cdots \vee (l_{k,1} \wedge l_{k,2} \wedge \cdots \wedge l_{k,m})$, *with* $l_{i,j} \in Lit_{\mathbb{F}}$.
- *The effect $\mathcal{E}$ is a propositional formula in a conjunction of literals.*
  $\mathcal{E} = l_1 \wedge l_2 \wedge \cdots \wedge l_n$, *with* $l_i \in Lit_{\mathbb{F}}$.
- *$pre: \mathbb{E} \to \mathcal{P}$ is a function describing the preconditions of events.*
- *$eff: \mathbb{E} \to \mathcal{E}$ is a function describing the effects of events.*

Let E_t denote all the events concurrently happening at timepoint t. E_t induces a state transition between S_t and S_{t+1}. E_{-1} represents the initialization steps for the fluents to have truth values at S_0. That is, suppose $S_0 = \{l_1, \neg l_2\}$, this means that l_1, l_2 had been initialized to be true and false respectively: $E_{-1} = \{init_{l_1}, init_{\neg l_2}\}$. E_{-1} is a placeholder for events in the past beyond the temporal scope of the formalization.

This formalism distinguishes between volitional actions and natural events. The set of volitional actions, denoted $\mathbb{A}$, are explicitly performed by an agent. The set $\mathbb{U}$ contain natural events that are automatically triggered as soon as all the preconditions hold. These two sets make up the entire set of events: $\mathbb{E} = \mathbb{A} \cup \mathbb{U}$. This division allows us to distinguish between different causal relations such as *causes* versus *enables* [12]. In brief, a precondition can be a *cause* with respect to natural events, but only an *enabling* factor with respect to volitional actions.

6.2 Answer Set Programming

Answer Set Programming (ASP) is a declarative logic programming paradigm widely used in symbolic Artificial Intelligence for knowledge representation and automated reasoning, especially in problems with incomplete information. A problem is first encoded as a logic program. The problem specification is then fed to an ASP solver to compute all the stable models (or consistent answer sets), from which the solutions can be extracted and interpreted.

ASP operates on a non-monotonic logic paradigm. In the context of moral reasoning, non-monotonic logic is particularly important. This is because moral reasoning is for the most part defeasible; for example, "All things considered, one should do X, *unless* the situation is such and such, then do Y instead." This kind of structure has been known to be very difficult to express in classical first-order logic [45]. Thus, ASP makes a good paradigm for implementing ethical reasoning [23].

This section aims to give a brief impression of answer set semantics; readers can refer to the cited source for a comprehensive overview [19]. Consider first a ground (variable-free) logic program P, consisting of the following rule:

$$A_1 :- B_1, B_2, \cdots B_m, not\ C_1, not\ C_2, \cdots not\ C_n$$

where A_i, B_i, C_i are literals. Each literal A_i, B_i, C_i can be of positive or negative form, that is, L or $\neg L$ for a given atom L. The symbol not is negation as failure (NAF), which means that for an atom L, $not\ L$ holds true if L cannot be demonstratively proven; in other words, $not\ L$ holds in the absence of L. For a rule r, the $A's$ make up the *head* of the rule - $head(r)$, the $B's$ are the positive parts of the body $body^+(r)$, and the $C's$ are the negative parts of the body $body^-(r)$. Intuitively, the rule can be read as "**if** all the $B's$ are true and all the $C's$ are not provably true, **then** A_1 is obtained." A rule without the body is called a *fact*. A rule without the head is called an *integrity constraint*.

Formally, let $Lit(P)$ be the set of all the literals in the program. A set $S \subseteq Lit(P)$ satisfies a normal rule r, denoted $S \vDash r$, if either $body^+(r) \nsubseteq S$ or $(head(r) \cup body^-(r)) \cap S \neq \emptyset$. The intuition is that S is "consistent"

with the rule r, if it trivially contains the head of r or vacuously the body of the rule is false. The body is false if $\exists b_i \notin S$ or $\exists c_i \in S$. S is considered an answer set of a normal program P, if S is a minimal set such that $S \models r, \forall r \in P$. Now we extend this definition to a logic program with variables. A rule R containing variables can be grounded by substituting all variables in the atoms with applicable values; this is known as the rule-instantiation process. Let $ground(R)$ be the set of all ground instances of the rule R. $S \models R$ if $S \models r \; \forall r \in ground(R)$.

All ASP programs in this project follows the semantics of an extended logic program [19]. That is, in addition to the standard rules mentioned above, it also offers disjunctive rules (where the rule's head has multiple atoms) and other high-level constructs such as choice rules, weak constraints, aggregates, etc.

Proposition 1 *Let* Π *be an extended disjunctive program, the set* S *be a minimal answer set of* Π*.*
Given a literal $\rho, \rho \in S$*, there exists a rule* $r \in \Pi$ *such that,* $head(r) = \rho, body^+(r) \subseteq S$*, and* $body^-(r) \cap S = \emptyset$

6.3 Causal Rules

Causal Rules module contain knowledge of the causal world. Basic predicates needed to model a scenario comprises a narrative of events (what happens when), world observations (what's true when), and the effect of events (what will be true after events happening).

Predicate	Meaning
$holds(F, T)$	Fluent F is true at time T
$holds(neg(F), T)$	Fluent F is false at time T
$performs(A, T)$	Action A is performed at time T
$action(A, GD, Eff)$	Precondition and effect of action A, $GD = pre(A), Eff = eff(A)$.
$auto(U, GD, Eff)$	Precondition and effect of the natural event U
$initiallyP(F)$	Fluent F is initially true
$initiallyN(F)$	Fluent F is initially false

Table 1: Some basic predicates

$fluent(F)$ is the domain predicate for each fluent $F \in \mathbb{F}$. To model a causal field, auxiliary predicates are needed to construct compound states, namely disjunctions and conjunctions. $conj(C)$ represents a conjunction C of fluent literals. $disj(D)$ represents a disjunction D of conjunctions. Predicate $in(S, M)$ indicates that M is a member of S; M as a literal will be a member of S as a conjunction; M as a conjunction will be a member of S as a disjunction. For example, a formula $\psi = (l_1 \wedge \neg l_2) \vee (l_3)$ will be encoded as:

$$disj(psi).$$

$$conj(c1).\ conj(c2).\ in(psi, c1).\ in(psi, c2).$$

$$in(c1, l1).\ in(c1, neg(l2)).\ in(c2, l3).$$

For each cluster of conditions we have a conjunction: c_1 designates the cluster $(l_1 \wedge \neg l_2)$, c_2 designates the cluster (l_3). Each literal is then nested in the corresponding cluster; for instance, $\neg l_2$ belongs in c_1. This particular representation stems from the need to reify the propositional formulas in our formalism. In the future, I aim to develop a more compact and convenient representation of propositional formulas. Ideally, users can construct a logical formula in any form, for which there exists an algorithm to convert into a DNF. The formula can be condensed into a string representation as it is, without having to put together all the auxiliary predicates.

Positive and negative fluents provide a convenient ways of speaking about negative causes. As in the INUS account, an absence of a condition (or negative condition) can serve as a necessary condition for the occurrence of an outcome. Formulating negative causes in terms of conditions (states of affair) help us avoid the difficulty of speaking in negative events (absence of events) [52]. For it is troubled with the intuition that something cannot come from nothing: how can an absence of an event give "rise" to some other event? Under this formalism, the absence of an event maintains a state of affairs that has been ongoing, which would then serve as a sufficient or necessary condition for an outcome. For example, failure to water the plan maintains the state of affairs that the houseplant is deprived of nutrition, which over time would decay to its own death.

6.4 Event Axioms

Event axioms describe the general relation between the occurrence of events and the truthfulness of fluents. These axioms allow the states of the world to evolve according to the causal rules described above.

The first group of rules describe the triggering of events based on its preconditions.

$$triggered(A, GD, T) :- action(A, GD, Eff), performs(A, T), holds(GD, T). \quad (10)$$

$$triggered(U, GD, T) :- auto(U, GD, Eff), holds(GD, T). \quad (11)$$

$$overtaken(E1, T) :- triggered(E1, \_, T), happens(E2, \_, T), priority(E2, E1), E1! = E1. \quad (12)$$

$$happens(E, GD, T) :- triggered(E, GD, T), not\ overtaken(E, T). \quad (13)$$

If it is a natural event U, as soon as the precondition GD (goal descriptor) holds, it will automatically be triggered (rule 11). On the other hand, if it is an action, it requires an additional fact that the agent explicitly performs the act (rule 10). Rule 12 provides a way to resolve conflict between concurrent events; if an event E_2 has priority over another event E_1, and if the two happen at the same time T, then the activation of E_1 will be discarded for that time (rule 13).

Next, we have the effect of events axioms.

$$initiated(E, F, T) :- apply(E, Eff, T), in(Eff, F), holds(neg(F), T), fluent(F). \quad (14)$$

$$terminated(E, F, T) :- apply(E, Eff, T), in(Eff, neg(F)), holds(F, T), fluent(F). \quad (15)$$

Predicate $apply(E, Eff, T)$ indicates that the event E has happened at time T and would bring about the effect conjunction Eff, where $eff(E) = Eff$. If the fluent F is present positively in the conjunction Eff, i.e. $F \in eff(E)$, it means that the happening of E would initiate the truthfulness of fluent F, from false to true (rule 14). Contrarily, if the fluent F is present negatively in the conjunction, i.e. $\neg F \in eff(E)$, its truthfulness would be terminated, from true to false, by the occurrence of E (rule 15).

Following the facts of $initiated$ and $terminated$, the truthfulness of fluents would actually be changed.

$$holds(F, 0) :- initiallyP(F), fluent(F). \quad (16)$$

$$holds(neg(F), 0) :- initiallyN(F), fluent(F). \quad (17)$$

$$holds(F, T + 1) :- initiated(E, F, T), fluent(F). \quad (18)$$

$$holds(neg(F), T + 1) :- terminated(E, F, T), fluent(F). \quad (19)$$

At time $T = 0$, truthfulness of fluents are managed by the initialization predicates $initiallyP$ and

$initiallyN$ (rule 17, 16). After that, if $initiated$ or $terminated$ by an event at time T, its value will be changed accordingly at time $T+1$ (rule 18, 19). If nothing happened however, the fluents' truthfulness should remain as it was, by the *commonsense law of inertia*:

$$holds(F, T+1) :- holds(F, T), fluent(F), time(T), maxTime(T2), T < T2,$$
$$not\ terminated(E, F, T) : event(E). \tag{20}$$
$$holds(neg(F), T+1) :- holds(neg(F), T), fluent(F), time(T), maxTime(T2), T < T2,$$
$$not\ initiated(E, F, T) : event(E). \tag{21}$$

Rule 20 says that a fluent, once true, will continue to be true unless explicitly terminated by some other events. Likewise, a fluent, once false, will continue to be false unless it is initiated by some other event (rule 21). In sum, the fluents' truth values under this formalism are influenced only by events, initialization states, or the law of inertia. In the standard Event Calculus however, a fluent can be affected by some other fluents, by being temporarily *released* by the law of inertia. This is also known as State Constraint rules [38]. Though useful in certain contexts, in order to incorporate it, we still need to determine the aspect of causality for this rule, which will be left for future development. Lastly, rule 22 enforces the coherency of state at any time point (Definition 3).

$$:- holds(F, T), holds(neg(F), T), time(T). \tag{22}$$

The last group of rules determine the truth values of the propositional formulas from its constituent fluent literals. A conjunction C holds true if all of its constituent fluent literals are true (rule 23). A disjunction D is true if at least one of its constituent conjunction is true (rule 24). Unlike primitive fluents, compound state (disjunction or conjunction) does not abide by the law of inertia.

$$holds(C, T) :- conj(C), time(T), holds(F, T) : in(C, F); not\ event(\_, \_, C). \tag{23}$$
$$holds(D, T) :- disj(D), holds(F, T), time(T), in(D, F), not\ event(\_, \_, D). \tag{24}$$

6.5 Causal Axioms

The event axioms emit traces of events in relation to states of the world, which can be used for higher-order causal inferences. Different causal theories can be embedded in this module, such as the counterfactual theory of causation or the NESS theory of causation. Sarmiento et al. in their paper has done much of the groundwork to formalize NESS causation in action languages [47]. This section reviews the formalized rules of NESS-cause as well as presents my own contribution of rules formalizing the weighted responsibility of a NESS cause.

6.5.1 Formal NESS causation

We first consider the direct causal relation between an event $e \in \mathbb{E}$ and a propositional formula ψ, which would be a precondition for some other event e' such that $pre(e') = \psi$.

Proposition 2 (Direct NESS-cause of a literal) *Given an event e happening at time t, $e \in \mathbb{E}_t$, and a fluent literal $l \in Lit_F$ true at time $t', t < t'$.*
e is a direct NESS-cause for l if and only if:

1. *l is $initiated$ or $terminated$ by e at t*
2. *$\forall t_1, t < t_1 \leq t'$, we have $holds(l, t_1)$.*

An event e is a direct NESS-cause for a literal l, if e is the last occurrence of events that made l true. In other words, the truth value of the fluent underlying the literal l was not changed ($initiated$ or $terminated$) by any other events at anytime in between.

Proposition 3 (Direct NESS-cause of a conjunction of literals) *Given an event e happening at time t, $e \in \mathbb{E}_t$, and a conjunction $\psi = l_1 \wedge l_2 \wedge \cdots \wedge l_n$ that is true at time $t', t < t'$.*
e is a direct NESS-cause for ψ if and only if:

1. *$\exists i \in \{1, 2, \cdots n\}$, such that e is a direct NESS-cause for the literal l_i.*

Since ψ is a conjunction, its truthfulness will depend on any constituent literal l_i. As such, an event e that causes the truth of $l_i \in \psi$ would also contribute towards the truthfulness of the conjunction ψ. e then can be considered a *a* cause for ψ, for if e did not happen, ψ would not hold.

Proposition 4 (Direct NESS-cause of a DNF) *Given an event e happening at time t, $e \in \mathbb{E}_t$, and a DNF, minimal and tautology free formula $\psi = \phi_1 \wedge \phi_2 \wedge \cdots \wedge \phi_n$ that is true at time $t', t < t'$.*
e is a direct NESS-cause for ψ if and only if:

1. $\exists i \in \{1, 2, \cdots n\}$, *such that* e *is a direct NESS-cause for the conjunction* ϕ_i.

Proposition 4 completes the intuition behind the NESS account of causation. Indeed, most problematic cases of causation come from disjunctive causes. Once the conditions are organized in clusters, an event can be seamlessly identified as a cause, by determining if it has any influence on the sufficiency of any cluster within the disjunction. An event e is a cause for a disjunction, if it contributes to the truthfulness of any conjunction, which would in turn be sufficient for the truth of the larger disjunction.

Now, we consider the transitivity of NESS causation. An event e_1 can stand in causal relation towards a propositional formula ψ of conditions as well as another event e_2. We define the following two *mutually recursive* functions:

Definition 5 (NESS causal relations) *Let* $\mathbb{E}_t$ *be the set of events occurring at time* t, $\mathcal{P}_t$ *be the set of propositional formulas of precondition that hold true at time* t.

1. $ness\colon \mathbb{E}_t \to \mathcal{P}_{t'}$ *defines a causal relation between an event and a formula.*

2. $actual\colon \mathbb{E}_t \to \mathbb{E}_{t'}$ *defines a causal relation between an event and another event.*

Intuitively, an event $e_1 \in \mathbb{E}_t$ *transitively* causes a formula $\psi \in \mathcal{P}_{t'}$, through an intermediate event $e_2 \in \mathbb{E}_{t_1}, t < t_1 < t'$. An event e_1 causes another event e_2, if e_1 contributed to the truthfulness of a formula ψ that is the precondition for the occurrence of e_2, $pre(e_2) = \psi$. The following two propositions formalize this intuition of transitive causes.

Proposition 5 (NESS-cause) *Given an event* $e \in \mathbb{E}_t$ *and a propositional formula* ψ *true at time* $t', t < t'$.
e *is a NESS-cause for* ψ *if and only if:*

1. *(Base case)* e *is a direct NESS-cause for* ψ

2. *(Recursive case)* $\exists e_1 \in \mathbb{E}_{t_1}, t < t_1 < t'$ *such that*

   - e_1 *is a NESS-cause for* ψ.
   - e *is an actual cause for* e_1.

Proposition 6 (Actual cause) *Given an event* $e_1 \in \mathbb{E}_{t_1}$ *and* $e_2 \in \mathbb{E}_{t_2}$.
e_1 *is an actual cause for* e_2 *if and only if:*

1. e_2 *is a natural event.* $e_2 \in \mathbb{U}$.

2. *There exists a propositional formula* $\psi \in \mathcal{P}_t$ *such that*

   - $t_1 < t \leq t_2$

- $pre(e_2) = \psi$
- e_1 *is a NESS-cause for* ψ

To determine if an event e_1 is an actual cause for another event e_2, we check if e_1 is a NESS-cause to the precondition formula ψ_2 of e_2. To determine if e_1 is a NESS-cause for ψ_2, we check if there is some event e_3 happening between e_1 and e_2, such that e_1 is an actual cause for e_3, and e_3 is a NESS-cause for ψ. The process is repeated until it bottoms out at the base case, where e_1 is a direct NESS-cause for some formula ψ_i (Proposition 4), which in turn is a precondition for some event e_i happening in between e_1 and e_3.

As of now, the relation of actual cause only applies from an arbitrary event to a natural, automatic event. With respect to a volitional action, we would have to employ a weaker causal notion such as 'enables' rather than 'causes' [12]. This will be left for future works.

6.5.2 Axioms of NESS causation

The first group of rules keep track of the actual relation between a compound state (disjunction or conjunction) and its constituents, as well as between fluents and fluents.

$$inertia(as(L,T), as(L,T+1)) :- holds(L,T), holds(L,T+1), literal(L). \tag{25}$$

$$r\_as(as(L,T), as(C,T)) :- conj(C), holds(C,T), in(C,L). \tag{26}$$

$$r\_as(as(C,T), as(D,T)) :- disj(D), holds(D,T), in(D,C), holds(C,T). \tag{27}$$

Rule 25 records the fluent literals that have their truthfulness preserved by the law of inertia. A literal L stands in an actual relation with the conjunction C in which it is in, if both holds true at time T (rule 26). Similarly, a disjunction D whose truth depends on the constituent conjunction C at time T, will also have their relation recorded (rule 27). These information will be used for the direct NESS-cause axioms, as described in propositions 2, 3, and 4.

$$direct\_ness(ao(init(L), -1), ao(L, 0)) :- holds(L, 0), literal(L). \tag{28}$$

$$direct\_ness(ao(E, T), as(F, T+1)) :- initiated(E, F, T). \tag{29}$$

$$direct\_ness(ao(E, T), as(neg(F), T+1)) :- terminated(E, F, T). \tag{30}$$

$$direct\_ness(Event, as(L, T+1)) :- direct\_ness(Event, as(L, T)), \\ inertia(as(L, T), as(L, T+1)). \tag{31}$$

Rule 28 stipulates the NESS-cause for fluent literals that owe their truth merely due to initialization, as opposed to being derived from rules of the causal world. These symbolize past causes that are beyond the temporal bounds. Rules 29, 30, and 31 document the causal relation between an event and a fluent, whose truth is respectively either *initiated*, *terminated*, or preserved through the law of inertia.

Lastly, rule 32 implements direct NESS-causes of an event with respect to a compound state. An event is a direct NESS-cause of some compound state, if it is a direct-NESS cause for one of its constituents, such that there is an *actual* relation between the larger compound and its constituent, as signified in the predicate $r\_as/2$.

$$direct\_ness(Event, as(GD, T)) :- direct\_ness(Event, as(GD\_S, T)), \\ r\_as(as(GD\_S, T), as(GD, T)). \tag{32}$$

The next group of rules is a direct encoding of the propositions 5 and 6:

$$ness(ao(E1, T1), as(GD, T2)) :- direct\_ness(ao(E1, T1), as(GD, T2)). \tag{33}$$

$$ness(ao(E1, T1), as(GD, T3)) :- actual(ao(E1, T1), ao(E2, T2)), \\ ness(ao(E2, T2), as(GD, T3)). \tag{34}$$

$$actual(ao(E1, T1), ao(E2, T2)) :- ness(ao(E1, T1), as(GD, T2)), \\ happens(E2, GD, T2), \\ auto(E2, GD, Eff). \tag{35}$$

Rule 33 implements the base case of NESS-cause. Rules 34 and 35 implement the mutually recursive

structure between NESS-cause and actual-cause.

6.5.3 Weighted causal responsibility

In this section, I present my own contribution, formalizing notion of weighted causal responsibility and adding its implementation to the framework.

Proposition 7 (Weighted responsibility of a literal condition) *Given a DNF formula $\psi \in \mathcal{P}$ and a fluent literal $l \in \psi$. Let the function $w(l, \psi)$ define the weighted contribution of the fluent literal l with respect to the formula ψ. Let ϕ be any conjunction in the DNF, $\phi \in \psi$.*

$$w(l, \psi) = \frac{|\{\phi \mid \phi \in \psi, l \in \phi\}|}{|\{\phi \mid \phi \in \psi\}|}$$

The weight of a literal is the ratio of the number of conjunctions that the literal is in, to the total number of conjunctions. Each conjunction represents a minimally sufficient set; this is thus equivalent to the idea of weighted responsibility of a condition c mentioned in section 5.

Proposition 8 (Weighted responsibility of an event) *Given a DNF formula $\psi \in \mathcal{P}_t$ and an event $e \in \mathbb{E}_{t'}$. Let the function $w(e, \psi)$ define the weighted contribution of the actual occurrence of e at time t' to the truth of formula ψ at time t. Let ϕ be any conjunction in the DNF, $\phi \in \psi$.*

$$w(e, \psi) = \frac{|\{\phi \mid \phi \in \psi, e \textit{ is a NESS-cause for } \phi\}|}{|\{\phi \mid \phi \in \psi\}|}$$

Similarly, the weighted causal responsibility of an event e with respect to a formula ψ is the number of conjunction in ψ that it is a NESS-cause of, divided by the total number of conjunctions. An event might have causal contribution to many conjunctions. If it is a NESS-cause to all of the conjunctions, it means that for all conjunctions there exists some literal whose truth is determined by the event. In other words, without its occurrence, all conjunctions will not hold due to the lack of some literal condition necessary for their truths. The counterfactual analysis can be easily extended from this fact. This, however, will require a more formal proof.

The following rules implement the aforementioned propositions 7 and 8.

$$weight(as(D,T),L,ratio(K,N)) :- disj(D), holds(D,T), literal(L), time(T),$$
$$K = \#count\{C : conj(C), r\_as(as(L,T),as(C,T)), r\_as(as(C,T),as(D,T))\},$$
$$K > 0,$$
$$N = \#count\{C : conj(C), r\_as(as(C,T),as(D,T))\}. \quad (36)$$

$$weight(as(D,T),ao(E,T1),ratio(K,N)) :- disj(D), holds(D,T), ness(ao(E,T1),as(D,T)), T1 < T,$$
$$K = \#count\{C : conj(C), r\_as(as(C,T),as(D,T)), ness(ao(E,T1),as(C,T))\},$$
$$K > 0,$$
$$N = \#count\{C : conj(C), r\_as(as(C,T),as(D,T))\}. \quad (37)$$

Rule 36 implements the weighted responsibility of a literal L with respect to the truthfulness of the DNF D at time T. Likewise, rule 37 implements the weighted causal responsibility of an event E with respect to the truthfulness of the DNF D at time T. The weight is represented by the term $ratio(K,N)$, which is equivalent to K/N.

6.6 Scope of the Formalism

This section summarizes the scope of the language framework:

1. Discrete timepoints: no continuous change.
2. Events are instantaneous: no durative events that last for more than 1 time point.
3. Boolean fluent: Fluents are not multi-valued.
4. Fluent has determinate values: no uncertainty of fluent values at any time point.
5. No conditional effects: events do not have different effects based on different conditions
6. Concurrency of events, with priority of occurrence
7. Deterministic events: no indeterministic or probabilistic effects of events
8. No state constraints
9. No indirect effects.

These restrictions reduce the complexity of demonstration, allowing us to focus on the causal reasoning aspects of the formalism without losing much essential features of a real-world scenario. In laying out these assumptions, it is acknowledged that there are different logical formalisms that have these features. Thus, it is also the aim of this project to step by step uplift these limitations, incorporating more kinds of events and increasing the expressive power of the formalism. However, this development must also be substantiated by theories of action as much as theories of causation, as noted by Batusov and Soutchanski in their paper [9]:

> It is clear that a broader definition of actual cause requires more expressive action theories that can model not only sequences of actions, but can also include explicit time and concurrent actions. Only after that one can try to analyze some of the popular examples of actual causation formulated in philosophical literature.

Though the problem of concurrency has somewhat been tackled, much are still left to be done, such as the causality of conditional effects and indeterministic events, which are of interest not only to modeling but also theoretical work. It is an interesting question of how much information of the real world we would have lost with these restrictions. For example, restriction 5 requires one precondition state and one effect state for each event, as opposed to conditional effects, which allow an event to have different, disjunctive effect states based on its precondition. In the Event Calculus, this can be modeled as follows [38, 56]:

$$initiates(switch, on, T) \leftarrow holds(off, T), -holds(on, T).$$
$$initiates(switch, off, T) \leftarrow holds(on, T), -holds(off, T).$$

The event $switch$ has different effects based on the precondition state. It turns on if the current state is off and not on; it turns off is the current state is on and not off. The equivalence of these rules in this formalism would require creating a new event for each association of event and precondition state. In other words, we normalize the many-to-one relationship between precondition and event.

$$action(switch\_off, on, off).$$
$$action(switch\_on, off, on).$$

A single event $switch$ has two preconditions on and off, depending on which it will have two different effects. For each event-precondition couple, we created two semantically distinct events $switch\_on$ and $switch\_off$, with each now only having one corresponding effect. It remains open whether we can always

mechanistically normalize the event-precondition relation with appropriate semantics.

6.7 Challenges and Worries

Though I do not intend to engage in the metaphysics of causation, it is worthwhile to reflect on the metaphysical assumptions made in the development of this causal framework. The spelling of assumptions will help us align with the appropriate philosophical theories. In addition, this section will address some worries and challenges in the business of causal modeling.

6.7.1 Causal relata

In a statement of token causation, A causes B, we are describing a causal relation between token A and token B. But what is the ontology of A and B? What are the kinds of things that are related in a causal claim? Some candidates are events, conditions, states, phenomena, processes, facts, etc [22]. The most common answer is events. A causal claim describes a *primitive* relation between two token events. The statement 'My dropping the ball causes the ball to be on the ground' describes a relation between the event of 'dropping the ball' and the event of 'ball being on the ground', as opposed to, for instance, the *state* of 'ball on ground.'

Our causal framework, however, operates on the event-state model, which is rested on the idea of a causal field. A causal field is a totality of conditions necessary and sufficient for the rising of an event. This is dated back to Mill, who takes a total set of conditions, or state, to be the sufficient cause for some effect. Mill discussed an example in which a man dies from eating a particular dish. Eating the particular dish, in general, does not always lead to death. It is only one of the many conditions - the man's health at the time, for instance - that conjointly caused the death. But, for simplicity and convenience of the mind, we tend to single out the act of eating the dish to be the only cause.

Others hold that the event of eating the dish is, in fact, the whole cause, and not just one in the totality of the conditions [22]. The event of eating the dish initiates the outcome. The man's poor health conditions do not directly contribute to the death, but only do so through the initiation of the event. It is really the eating event that makes a difference, accelerating the ongoing health conditions, which have been standing in the background, to the fatal effect that it has. Here, the event does not stand hand in hand with other conditions to make up the whole cause; there seems to be an ontological distinction between event and state. Events serve as active initiators, while conditions/states always serve as enablers, allowing an event to have the effect as it does. The cause, on this ground, is identified with events, while states are always preconditions.

Indeed, the notion of states being be causally efficaceous is problematic. States are deemed passive and inert, whereas events are active and salient. It is difficult to conceive something passive, lacking in power, could by itself give rise something active, outstanding to the mind's attention. The argument for this is articulated as follows [22]. Suppose some state S causes event E to occur at time t. We consider two possibilities:

1. S already holds over an interval $[t', t]$
2. S only starts to hold at time t; in other words, S did not hold for the interval $[t', t)$

In case 1, if S already holds, what makes S cause E later, rather than sooner? If S is sufficient for E at time t, then surely it must already have been sufficient for E at the earlier time t'. If S however only starts to hold (case 2), then we could easily find another event e' that causes the change of state. This event e' then would be the cause.

Replying to these worries, and as a brief justification for the event-states model, I argue that states can *non-causally* give rise to an event. Suppose I set an alarm clock at some time in the future. The state of the alarm is thus turned from off to on, and continues to be on until the set time. Suddenly, when the time comes, it triggers the ringing of the alarm clock. Here the state on remains for a considerable amount of time until eventually giving rise to the event of ringing. From an outsider's view, it appears that some passive state has suddenly amounted to an active event.

Of course, that needs not be the end of an explanation. One could capitalize on the suddenness of the phenomenon, appealing at a micro level to find the true cause. Say, I turned the alarm clock from off to on. What it entailed internally is that the phone was constantly ticking the bits every second. When the tracking bits hit the set time, it activated the speaker system, producing the ringing. This response, however, misses the mark; certainly one could go to an atomic level and claim that particles are constantly moving every nanosecond. But in commonsense reasoning at a macro level, it is sensible enough to say some states could entail the rising of events. States do not *cause* events. Rather, an event could be said to be a salient state of interest. For example, we may have a state that measures the water level in a vase. A tap is pouring water into the vase and the water level rises continuously. When the water level hits a certain mark, it causes a spillover. The state of current water level, combined with the volume of the vase is enough to *determine* the event of a spill-over. The rationale then, is that events can arise out of states, not in virtue of underlying causal power, but of informational convenience. It is important for the informational convenience approach to preserve the intrinsic causality of the world.

As a review, in our framework, event can *cause* states. While states can either *trigger* or *enable* events. An event can be enabled, if it is a volitional action. It can be triggered, or given rise to, if it is a natural events.

States are the means in which an event relates to another. In turn, event allows for the evolution of states. However, it remains a problem whether the choice of events versus states can affect or limit the higher-order causal inferences. Does the causal structure of a scenario change drastically if a token is modeled as an event rather than state, and vice versa? For example, given the statement 'dropping the ball causes it to be on the ground,' can the tokens 'dropping the ball' and 'on the ground' be modeled as events or states interchangeably? Does it have any consequences to causal propositions derivable from the system? Or is there a substantially ontological distinction between states and events that some token can only be modeled as states, while other as events?

6.7.2 Incomplete causal world

Any causal models will inevitably face the epistemic problem of incomplete knowledge of the causal world. There are skeptical worries as to how we can come to know the causal laws of nature in the first place. If laws are just regularities that have come to habituate the mind, then how can we distinguish mere association of events from genuine laws of nature. For example, the cry of a rooster invariably precedes the rise of the sun, but we do not want to say the the rooster's crowing is a cause for the sun to rise.

According to Mill, laws are a special kind of regularities [5]. They are just like regularities, but more stable and general. Our job is to find those regularities that are the most general and organize them into a axiomatic system. An ideal system of causal laws allows us to discover new facts, entail more truths, and explain phenomena in the causal world using the foundational axioms within the system. If the succession of rooster's crowing and the sun's rising is stable enough a pattern, then they would deserve a place in the system of laws. What this means for our task of causal modeling is that, it alleviates the epistemic anxieties of finding the utmost objective laws, and of justifying whether a regularity is a genuine law or not. The goal for the formalism is to enforce a coherence in our body of causal knowledge, even if it cannot directly track the truths in the causal world. But at the same time, it should be flexible enough for the discovery (unexpected) inconsistencies, thus allowing for the revision of the causal laws.

One could imagine a system containing incomplete knowledge of the causal world. Upon being fed new information, it can automatically identify inconsistencies and narrow down the rules to be updated. For instance, generally when I light a match, there will be fire, $light\_match \rightarrow fire$. But suppose that I come across a new environment and discover that without oxygen, I cannot light up the match to fire. Thus, I would have to revise the rule: $light\_match \wedge oxygen \rightarrow fire$. This is a non-trivial task, also known as the Epistemological Frame Problem of AI [57]. In a system with large database knowledge of causal rules, how do we determine the scope of relevant rules to be updated? What rules will be likely updated, given that

now we know oxygen is an important factor in making fire? And how do we update these rules iteratively and consistently?

7 Result

In this section, I present the implementation results of the causal language in ASP and their applications on certain problems of causation. All ASP programs in this project are implemented on the Clingo software (v5.6), which consists of a grounder and a solver. Additionally, I develop a Python framework for debugging and testing ASP programs as well as running encoded ethical dilemmas, using Clingo's Python API. [6]

7.1 Example 1: Electrocution

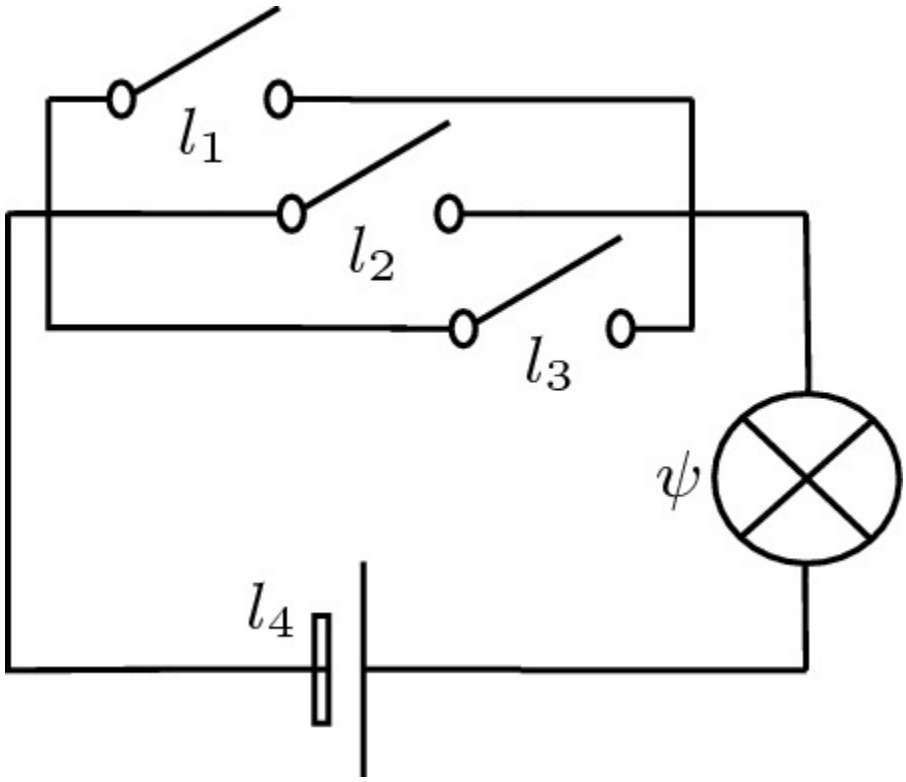


Figure 2: Electrical circuit of electrocution

This example is inspired from Sarmiento et al.'s paper [47]. Imagine an electrical circuit containing four switches l_1, l_2, l_3, l_4. A prisoner is connected to the electrical circuit and would be electrocuted when it is

[6] https://potassco.org/clingo/python-api/5.6/clingo/

closed (Figure 2). l represents a closed switch, while $\overline{l}$ represents an open switch. Four events $e_{\neg 1}, e_2, e_3, e_4$ represent turning on/off the corresponding switches.

Initially, $S_0 = \{l_1, \overline{l_2}, \overline{l_3}, \overline{l_4}\}$. The events happening at time $t = 0$ are$E_0 = \{e_{\neg 1}, e_2\}$, and at time $t = 1$ are $E_1 = \{e_3, e_4\}$. From the circuit, the electrocution condition can be written in the following DNF formula $\psi = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$. As a result, it is expected that at the end of these events sequence the prisoner will be electrocuted. The problem of interest is to determine the causes for the electrocution as well as each's weighted causal contribution.

Running this encoded scenario (Appendix B.1) along with the causal axioms and event axioms yields the following result (Figures 3 and 4).

```
Event traces:
holds(l1,0)
holds(neg(l2),0)
holds(neg(l3),0)
holds(neg(l4),0)
holds(true,0)
happens(e2,true,0)
happens(e1,true,0)
holds(true,1)
holds(neg(l3),1)
holds(neg(l4),1)
holds(neg(l1),1)
holds(l2,1)
happens(e4,true,1)
happens(e3,true,1)
holds(true,2)
holds(l3,2)
holds(l4,2)
holds(l2,2)
holds(neg(l1),2)
holds(electrocuteCond3,2)
holds(electrocuteCond,2)
holds(electrocuteCond2,2)
happens(electrocute,electrocuteCond,2)
----
```

Figure 3: Event traces

Figure 3 shows the traces of event happening along with evolution of the world, sorted by time. In debugging mode, it will show only the key predicates indicating the occurrence of event $happens/3$ and states of the world $holds/2$. In this example, the expected fact at the end of the events sequence is $happens(electrocute, electrocuteCond, 2)$, indicating that the prisoner was electrocuted at time 2.

Figure 4 shows the causal inferences based on the emitted traces of events. e_2, e_3, e_4 are both considered as causes for the electrocution under the NESS account of causation. Furthermore, the causal weight of e_2 and e_3 are each $1/2$, whereas the weight of e_4 is $2/2 = 1$. Thus, it can be derived from the output that e_4 is counterfactually necessary for the outcome of electrocution.

```
----
Causal traces:
ness(ao(e1,0),as(neg(l1),1))
ness(ao(e1,0),as(neg(l1),2))
ness(ao(e2,0),as(l2,1))
ness(ao(e2,0),as(l2,2))
ness(ao(e2,0),as(electrocuteCond2,2))
ness(ao(e2,0),as(electrocuteCond,2))
ness(ao(e3,1),as(l3,2))
ness(ao(e3,1),as(electrocuteCond3,2))
ness(ao(e3,1),as(electrocuteCond,2))
ness(ao(e4,1),as(l4,2))
ness(ao(e4,1),as(electrocuteCond2,2))
ness(ao(e4,1),as(electrocuteCond3,2))
ness(ao(e4,1),as(electrocuteCond,2))
weight(as(electrocuteCond,2),ao(e4,1),ratio(2,2))
weight(as(electrocuteCond,2),ao(e3,1),ratio(1,2))
weight(as(electrocuteCond,2),ao(e2,0),ratio(1,2))
weight(as(electrocuteCond,2),l2,ratio(1,2))
weight(as(electrocuteCond,2),l3,ratio(1,2))
weight(as(electrocuteCond,2),l4,ratio(2,2))
----
Errors:
```

Figure 4: Causal traces

7.2 Example 2: a neuron diagram

This section demonstrates a way to translate neuron diagrams within our causal frameworks.

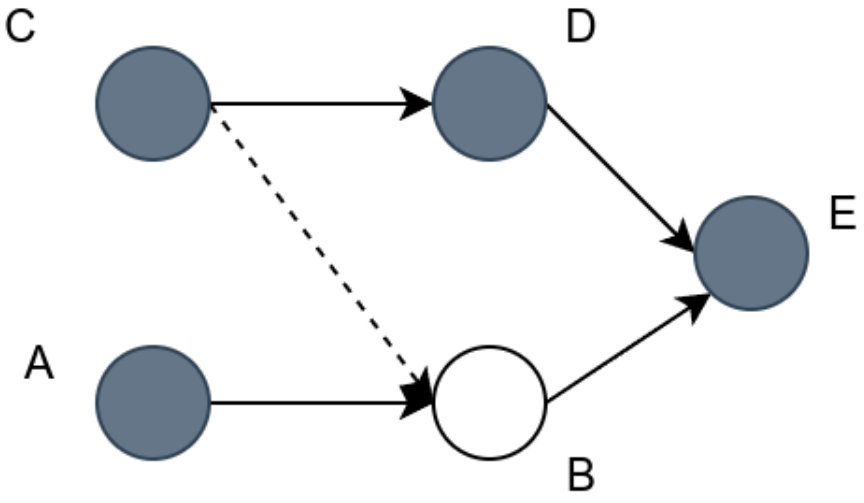


Figure 5: Neuron diagram: preemption [40]

Circles represent neurons. Arrows represent token causal influence from one neuron to another; straight arrows are excitatory signal, whereas dotted arrows are inhibitory signal. Shaded neuron indicates its firing; unshaded indicates it is not fired. In figure 5, neurons C and A fire at the same time. C sends an excitatory signal to D, causing it to fire. A excites B, but since C also inhibits B, B does not fire. Lastly, D excites E to be fired.

The neuron diagram in Figure 5 is a problematic case of preemption. We want to say that C is the cause for the firing of E, while A is not. Here, C is the *preempting* cause. Let us first start with simplistic versions of counterfactual and regularity account. Counterfactually, C is not a cause for E, because had C not fired, E would still have fired due to A. Similarly for A, had A not fired, C would have been enough for E. Thus neither of A and C are causes for E. Meanwhile, on the regularity account, both of them are counted *wrongly* as causes. For every time C fires, it suffices the firing of E. Likewise, every time A fires, it suffices the firing of E.

The flaw in these reasoning is that it has not spelled out all the conditions for the firing of each event. A is only a cause for E in the absence of other factors, namely C. The sufficient conditions for B is as follows:

$$A \wedge \neg C \rightarrow B \tag{38}$$

and for E:

$$D \vee B \rightarrow E \tag{39}$$

Although A is instantiated, it is by itself not sufficient for the instantiation of B. Because B does not follow through all the way, E is not caused by B. Meanwhile, the causal path $C-D-E$ are completely instantiated. Thus, E is caused by D and C rather than B and A.

Another issue is that actual causation analysis requires tracing the entire causal process, to the extent that all the relevant causal laws are given. In the neuron diagram, the neuronal path A to E, and likewise C to E is not a primitive causal relation. They cannot be analyzed by themselves by ignoring all the other events happening in between. The importance of a complete analysis of the causal process is stressed by Wright [64]:

> one must include in the causal analysis the entire causal process up to the time of the occurrence of the consequence. The definition of a causal law in the immediately preceding paragraph assures this by requiring that the instantiation of the consequent of the causal law occur immediately when all of the conditions in the antecedent of the causal law have been instantiated. A

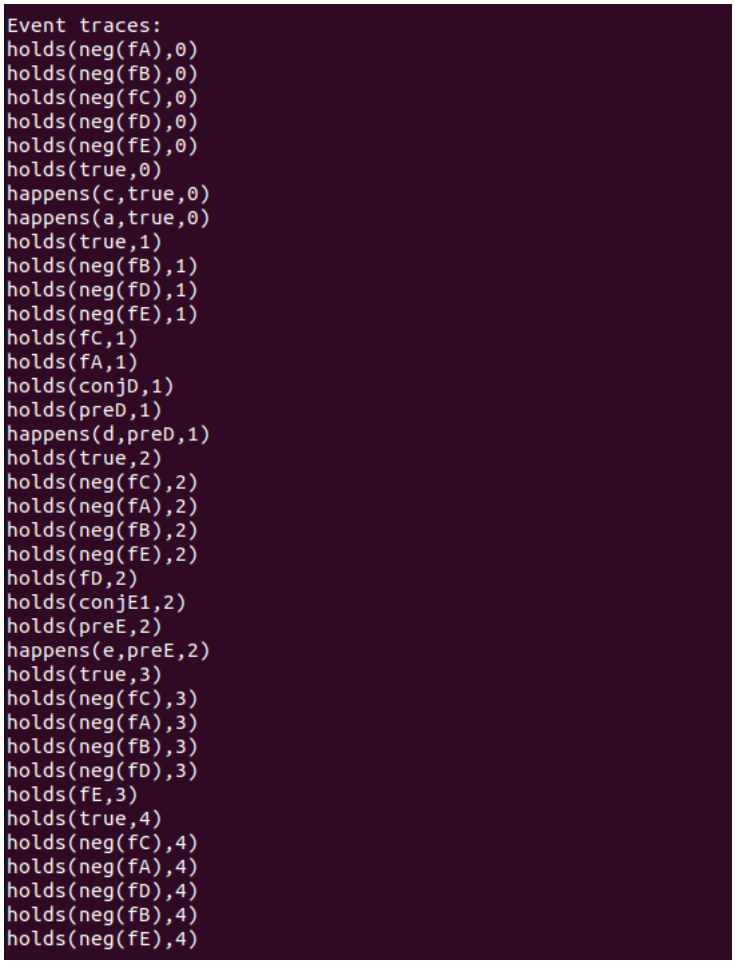


Figure 6: Event traces of neuron diagram

> causal process consists of the instantiation of one or more simultaneously or successively operative causal laws. Another critical feature of causal laws – and the related concept of causal sufficiency as distinct from mere lawful sufficiency – is their successional or directional nature, according to which the instantiation of the conditions in the antecedent of the causal law causes the instantiation of the consequent, but not vice versa.

Collapsing the causal process between A and E implies that there is a natural causal law connect A directly to E. In other words, E would be an intrinsic effect of A. But it is not the case in this neuron diagram, A only has an influence on E through the intermediary of B. Taking A directly to E would result in a different neuron diagram, with totally different causal traces.

The neurons' causal connection can be implemented in our causal framework, as detailed in the appendix B.2. Running the problem encoding, with causal and event axioms yields the following causal traces (Figure 7). C is an actual cause for D and E, as seen in $actual(ao(c,0), ao(d,1))$, and $actual(ao(c,0), ao(e,2))$.

```
Causal traces:
ness(ao(a,0),as(fA,1))
ness(ao(c,0),as(fC,1))
ness(ao(c,0),as(conjD,1))
ness(ao(c,0),as(preD,1))
ness(ao(c,0),as(fD,2))
ness(ao(c,0),as(conjE1,2))
ness(ao(c,0),as(preE,2))
ness(ao(c,0),as(fE,3))
actual(ao(c,0),ao(d,1))
actual(ao(c,0),ao(e,2))
ness(ao(a,1),as(neg(fA),2))
ness(ao(a,1),as(neg(fA),3))
ness(ao(a,1),as(neg(fA),4))
ness(ao(c,1),as(neg(fC),2))
ness(ao(c,1),as(neg(fC),3))
ness(ao(c,1),as(neg(fC),4))
ness(ao(d,1),as(fD,2))
ness(ao(d,1),as(conjE1,2))
ness(ao(d,1),as(preE,2))
ness(ao(d,1),as(fE,3))
actual(ao(d,1),ao(e,2))
weight(as(preD,1),ao(c,0),ratio(1,1))
weight(as(preD,1),fC,ratio(1,1))
ness(ao(d,2),as(neg(fD),3))
ness(ao(d,2),as(neg(fD),4))
ness(ao(e,2),as(fE,3))
weight(as(preE,2),ao(d,1),ratio(1,1))
weight(as(preE,2),ao(c,0),ratio(1,1))
weight(as(preE,2),fD,ratio(1,1))
ness(ao(e,3),as(neg(fE),4))
----
Errors:
```

Figure 7: Causal traces of neuron diagram example

Meanwhile, A is not an actual cause for anything.

The formalization of weighted causal responsibility, however, does not work for this example. As seen in the output, the weighted contribution of c with respect to e is exactly 1. However, e in this example does not counterfactually depend on c. For a could still have caused e to fire, had it not been for c. I suspect this has to do with the fact that the causal weight formula does not fully capture transitivity. This needs further examination in future works.

Besides, there were no errors, as shown.

8 Conclusion

In this paper, I motivated the need for a rigorous formalism of ethics and explored some of the ideal standards in the backdrop of the outstanding challenges in computational ethics. One aspect of such challenges is how to encode an ethical scenario exhaustively and faithfully, without any unintentional biases. I ar-

gued that an exhaustive and general description of a scenario requires an extensive understanding of the causal world. Thus, causal modeling is an integral part of ethical modeling. Starting from contemporary theories of actual causation, I introduced the NESS test, commonly used in the law to determine liability. I then presented a framework for reasoning about ethics and actions, extended from previous influential works, especially in [47, 11]. I demonstrated the application of the framework on some problematic cases of overdetermined causation. Finally, to highlight my contributions:

- I noted the difficulty in general computational ethics and motivated the need for causal modeling within ethical formalism.
- I advanced the causal/ethical framework, adding rules formalizing weighted causal responsibility and delayed effects.
- I developed a Python framework, which makes it easier to debug and test the formalism, as well as running ethical scenarios.
- I added a number of problematic cases of causation and ethics.

9 Future works

This project is only a first step towards a more rigorous and complete formalism of ethics and causation. As discussed over the course of this paper, there are still outstanding problems in the overall approach, as well as in the specific implementation of the formalism. Many restrictions were set within this project for pragmatic purposes. In the future, more restrictions will be lifted, and more features will be iteratively added. This heads towards a complete account of the causal world (within the bound of our knowledge). This must, however, be accompanied with adequate theoretical justifications drawn from the philosophical literature.

There are specific milestones I wish to achieve in the long run:

- Incorporating Inductive Learning for Answer Set Programming [33]. This is headed towards a more hybrid approach, allowing the AI agent to learn new ethical rules, in addition to the encoded principles [18].
- Natural language to ASP, and vice versa [53, 15]. Natural languages are an abundant resource for ethical scenarios. One can imagine an agent that can automatically parse from text and construct a formal description of the scenario, in accordance with the pre-encoded axioms about the causal

world. In return, its ethical evaluation can be translated into natural language, thus providing a human-readable explanation for its decision.

- Causal explanation: examples used in this project are only for demonstrations and thus simple in nature. When the causal scenario grows complex enough, we cannot examine the entire traces of actual causation. Thus, we must narrow down relevant causal relations and construct a reasonable narrative. For example, the fact that I can create a fire from my lighter depends on several factors, such as oxygen, remaining gas, malfunctioned lighter, etc. What conditions, out of the total causal field, should be output as the cause *of relevance*?

In the short run, the implementations will need a clean-up.

Appendices

A Full Implementation of the Language

All codes with annotation are available at https://github.com/vulecoff/ilasp-ethics-thesis

A.1 Event Axioms

Events

$$event(A, GD, Eff) :- action(A, GD, Eff).$$
$$event(U, GD, Eff) :- auto(U, GD, Eff).$$
$$event(E) :- event(E, GD, Eff).$$
$$time(0..T) :- maxTime(T).$$

Event precondition axioms

$$triggered(A, GD, T) :- action(A, GD, Eff), performs(A, T), holds(GD, T).$$
$$triggered(U, GD, T) :- auto(U, GD, Eff), holds(GD, T).$$

$$overtaken(E1, T) :- triggered(E1, \_, T), happens(E2, \_, T), priority(E2, E1), E1! = E1.$$
$$happens(E, GD, T) :- triggered(E, GD, T), not\ overtaken(E, T).$$

Effect of event axioms

$$\begin{aligned} apply(E, Eff, T) :- & \ happens(E, GD, T), event(E, GD, Eff), \\ & not\ delayedEff(E, Eff, TD) : time(TD). \\ apply(E, Eff, T + TD - 1) :- & \ happens(E, GD, T), event(E, GD, Eff), delayedEff(E, Eff, TD), \\ & time(TD), maxTime(M), T + TD \leq M, TD > 0. \end{aligned}$$

$$initiated(E, F, T) :- apply(E, Eff, T), in(Eff, F), holds(neg(F), T), fluent(F).$$
$$terminated(E, F, T) :- apply(E, Eff, T), in(Eff, neg(F)), holds(F, T), fluent(F).$$

Fluent axioms, inertia of fluents

$$holds(F, 0) :- initiallyP(F), fluent(F).$$
$$holds(neg(F), 0) :- initiallyN(F), fluent(F).$$
$$holds(F, T + 1) :- initiated(E, F, T), fluent(F).$$
$$holds(neg(F), T + 1) :- terminated(E, F, T), fluent(F).$$

$$\begin{aligned} holds(F, T + 1) :- & \ holds(F, T), fluent(F), time(T), maxTime(T2), T < T2, \\ & not\ terminated(E, F, T) : event(E). \\ holds(neg(F), T + 1) :- & \ holds(neg(F), T), fluent(F), time(T), maxTime(T2), T < T2, \\ & not\ initiated(E, F, T) : event(E). \\ :- & \ holds(F, T), holds(neg(F), T), time(T). \end{aligned}$$

Conjunctions and disjunctions of fluents

$$literal(neg(F)) :- fluent(F).$$
$$literal(F) :- fluent(F).$$
$$in(L, L) :- literal(L).$$
$$holds(true, T) :- time(T).$$

$$holds(C, T) :- conj(C), time(T), holds(F, T) : in(C, F); not\ event(\_, \_, C).$$
$$holds(D, T) :- disj(D), holds(F, T), time(T), in(D, F), not\ event(\_, \_, D).$$

A.2 Causal Axioms

Actual relation between formulas and its constituent literals

$$inertia(as(L, T), as(L, T + 1)) :- holds(L, T), holds(L, T + 1), literal(L).$$
$$r\_as(as(L, T), as(C, T)) :- conj(C), holds(C, T), in(C, L).$$
$$r\_as(as(C, T), as(D, T)) :- disj(D), holds(D, T), in(D, C), holds(C, T).$$

Direct NESS

$$direct\_ness(ao(init(L), -1), ao(L, 0)) :- holds(L, 0), literal(L).$$
$$direct\_ness(ao(E, T), as(F, T + 1)) :- initiated(E, F, T).$$
$$direct\_ness(ao(E, T), as(neg(F), T + 1)) :- terminated(E, F, T).$$
$$direct\_ness(Event, as(L, T + 1)) :- direct\_ness(Event, as(L, T)),\ inertia(as(L, T), as(L, T + 1)).$$
$$direct\_ness(Event, as(GD, T)) :- direct\_ness(Event, as(GD\_S, T)),\ r\_as(as(GD\_S, T), as(GD, T)).$$

NESS cause and actual cause

$$
\begin{aligned}
ness(ao(E1,T1),as(GD,T2)) :- & \ direct\_ness(ao(E1,T1),as(GD,T2)).\\
ness(ao(E1,T1),as(GD,T3)) :- & \ actual(ao(E1,T1),ao(E2,T2)),\\
& \ ness(ao(E2,T2),as(GD,T3)).\\
actual(ao(E1,T1),ao(E2,T2)) :- & \ ness(ao(E1,T1),as(GD,T2)),\\
& \ happens(E2,GD,T2),auto(E2,GD,Eff).
\end{aligned}
$$

Weight of responsibility

$$
\begin{aligned}
weight(as(D,T),L,ratio(K,N)) :- & \ disj(D),holds(D,T),literal(L),time(T),\\
& K = \#count\{C : conj(C),r\_as(as(L,T),as(C,T)),r\_as(as(C,T),as(D,T))\},\\
& K > 0,\\
& N = \#count\{C : conj(C),r\_as(as(C,T),as(D,T))\}.\\
weight(as(D,T),ao(E,T1),ratio(K,N)) :- & \ disj(D),holds(D,T),ness(ao(E,T1),as(D,T)),T1 < T,\\
& K = \#count\{C : conj(C),r\_as(as(C,T),as(D,T)),ness(ao(E,T1),as(C,T))\},\\
& K > 0,\\
& N = \#count\{C : conj(C),r\_as(as(C,T),as(D,T))\}.
\end{aligned}
$$

B Examples

B.1 Electrocution

$$maxTime(2).$$

$$fluent(l1; l2; l3; l4).$$

$$initiallyN(l2; l3; l4).\ initiallyP(l1).$$

$$disj(electrocuteCond).$$

$$in(electrocuteCond, (electrocuteCond1; electrocuteCond2; electrocuteCond3)).$$

$$conj(electrocuteCond1).$$

$$in(electrocuteCond1, l1).in(electrocuteCond1, l4).$$

$$conj(electrocuteCond2).$$

$$in(electrocuteCond2, l2).in(electrocuteCond2, l4).$$

$$conj(electrocuteCond3).$$

$$in(electrocuteCond3, l3).in(electrocuteCond3, l4).$$

$$action(e1, true, neg(l1)).$$

$$action(e2, true, l2).$$

$$action(e3, true, l3).$$

$$action(e4, true, l4).$$

$$auto(electrocute, electrocuteCond, true).$$

$$performs(e1, 0).\ performs(e2, 0).$$

$$performs(e3, 1).\ performs(e4, 1).$$

B.2 Neuron preemption

$$maxTime(4).$$

$$fluent(fA; fB; fC; fD; fE).$$

$$initiallyN(fA; fB; fC; fD; fE).$$

$$action(a, true, fA).\ action(a, true, neg(fA)).\ delayedEff(a, neg(fA), 2).$$

$$action(c, true, fC).\ action(c, true, neg(fC)).\ delayedEff(c, neg(fC), 2).$$

$$auto(d, preD, fD).\ auto(d, preD, neg(fD)).\ delayedEff(d, neg(fD), 2).$$

$$disj(preD).\ in(preD, conjD).$$

$$conj(conjD).\ in(conjD, fC).$$

$$auto(b, preB, fB).\ auto(b, preB, neg(fB)).\ delayedEff(b, neg(fB), 2).$$

$$disj(preB).\ in(preB, conjB).$$

$$conj(conjB).\ in(conjB, fA).\ in(conjB, neg(fC)).$$

$$auto(e, preE, fE).\ auto(e, preE, neg(fE)).\ delayedEff(e, neg(fE), 2).$$

$$disj(preE).\ in(preE, (conjE1; conjE2)).$$

$$conj(conjE1; conjE2).\ in(conjE1, fD).\ in(conjE2, fB).$$

$$performs(a, 0).\ performs(c, 0).$$

References

[1] David Abel, James MacGlashan, and Michael L. Littman. "Reinforcement Learning as a Framework for Ethical Decision Making". In: *AAAI Workshop: AI, Ethics, and Society*. 2016. URL: https://api.semanticscholar.org/CorpusID:14717578.

[2] Michael Anderson, Susan Anderson, and Chris Armen. "Towards Machine Ethics". In: July 2004.

[3] Michael Anderson and Susan Leigh Anderson. In: *Paladyn, Journal of Behavioral Robotics* 9.1 (2018), pp. 337–357. DOI: doi:10.1515/pjbr-2018-0024. URL: https://doi.org/10.1515/pjbr-2018-0024.

[4] Michael Anderson, Susan Leigh Anderson, and Chris Armen. "MedEthEx: a prototype medical ethics advisor". In: *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*. IAAI'06. Boston, Massachusetts: AAAI Press, 2006, pp. 1759–1765. ISBN: 9781577352815.

[5] Holger Andreas and Mario Guenther. "Regularity and Inferential Theories of Causation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University, 2021.

[6] Stuart Armstrong. "Motivated Value Selection for Artificial Agents". In: *AI and Ethics*. 2015. URL: https://api.semanticscholar.org/CorpusID:8607080.

[7] Edmond Awad et al. "Computational ethics". In: *Trends in Cognitive Sciences* 26.5 (2022), pp. 388–405. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2022.02.009. URL: https://www.sciencedirect.com/science/article/pii/S1364661322000456.

[8] Edmond Awad et al. "The Moral Machine experiment". In: *Nature* 563.7729 (Nov. 2018), pp. 59–64. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0637-6. URL: https://doi.org/10.1038/s41586-018-0637-6.

[9] Vitaliy Batusov and Mikhail Soutchanski. "Situation Calculus Semantics for Actual Causality". In: *International Symposium on Commonsense Reasoning*. 2018. URL: https://api.semanticscholar.org/CorpusID:19210399.

[10] Martin Mose Bentzen and Felix Lindner. "A Formalization of Kant's Second Formulation of the Categorical Imperative". In: *CoRR* abs/1801.03160 (2018). arXiv: 1801.03160. URL: http://arxiv.org/abs/1801.03160.

[11] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. "A Declarative Modular Framework for Representing and Applying Ethical Principles". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '17. São Paulo, Brazil: International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 96–104.

[12] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. "Event-Based and Scenario-Based Causality for Computational Ethics". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '18. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 147–155.

[13] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. "Modelling Moral Reasoning and Ethical Responsibility with Logic Programming". In: *Logic for Programming, Artificial Intelligence, and Reasoning*. Ed. by Martin Davis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 532–548. ISBN: 978-3-662-48899-7.

[14] Vincent Bonnemains, Claire Saurel, and Catherine Tessier. "Embedded ethics: some technical and ethical challenges". In: *Ethics and Information Technology* 20.1 (Mar. 2018), pp. 41–58. ISSN: 1572-8439. DOI: 10.1007/s10676-018-9444-x. URL: https://doi.org/10.1007/s10676-018-9444-x.

[15] SIMONE CARUSO et al. "CNL2ASP: Converting Controlled Natural Language Sentences into ASP". In: *Theory and Practice of Logic Programming* 24.2 (2024), pp. 196–226. DOI: 10.1017/S1471068423000388.

[16] James P. Delgrande et al. *Current and Future Challenges in Knowledge Representation and Reasoning*. 2023. arXiv: 2308.04161 [cs.AI].

[17] Abeer Dyoub, Stefania Costantini, and Francesca A. Lisi. "Logic Programming and Machine Ethics". In: *Electronic Proceedings in Theoretical Computer Science* 325 (Sept. 2020), pp. 6–17. ISSN: 2075-2180. DOI: 10.4204/eptcs.325.6. URL: http://dx.doi.org/10.4204/EPTCS.325.6.

[18] Abeer Dyoub, Stefania Costantini, and Francesca A. Lisi. "Towards Ethical Machines Via Logic Programming". In: *Electronic Proceedings in Theoretical Computer Science* 306 (Sept. 2019), pp. 333–339. ISSN: 2075-2180. DOI: 10.4204/eptcs.306.39. URL: http://dx.doi.org/10.4204/EPTCS.306.39.

[19] Thomas Eiter, Giovambattista Ianni, and Thomas Krennwallner. "Answer Set Programming: A Primer". In: vol. 5689. Jan. 2009, pp. 40–110. ISBN: 978-3-642-03753-5. DOI: 10.1007/978-3-642-03754-2_2.

[20] Philippa Foot. "The Problem of Abortion and the Doctrine of the Double Effect". In: *Oxford Review* 5 (1967), pp. 5–15.

[21] J. Dmitri Gallow. "The Metaphysics of Causation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022.

[22] Antony Galton. "States, processes and events, and the ontology of causal relations". In: *Frontiers in Artificial Intelligence and Applications* 239 (Jan. 2012), pp. 279–292. DOI: 10.3233/978-1-61499-084-0-279.

[23] J. Ganascia. "Ethical System Formalization using Non-Monotonic Logics". In: 2007. URL: https://api.semanticscholar.org/CorpusID:6104466.

[24] Michael Gelfond and Vladimir Lifschitz. "Action Languages". In: *ETAI* 3 (Apr. 1999).

[25] Naveen Sundar Govindarajulu and Selmer Bringsjord. "On Automating the Doctrine of Double Effect". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 4722–4730. ISBN: 9780999241103.

[26] Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2016. ISBN: 0262035022.

[27] Patrik Haslum et al. "An Introduction to the Planning Domain Definition Language". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13 (Apr. 2019), pp. 1–187. DOI: 10.2200/S00900ED2V01Y201902AIM042.

[28] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. *An Overview of Catastrophic AI Risks*. 2023. arXiv: 2306.12001 [cs.CY].

[29] Liwei Jiang et al. *Can Machines Learn Morality? The Delphi Experiment*. 2021. eprint: arXiv:2110.07574.

[30] Deborah G. Johnson and Mario Verdicchio. "Ethical AI is Not about AI". In: *Commun. ACM* 66.2 (Jan. 2023), pp. 32–34. ISSN: 0001-0782. DOI: 10.1145/3576932. URL: https://doi.org/10.1145/3576932.

[31] Gabbrielle Johnson. *Algorithmic Bias: On the Implicit Biases of Social Technology*. May 2020. URL: http://philsci-archive.pitt.edu/17169/.

[32] Frances Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York ; Oxford University Press, 2007.

[33] Mark Law, Alessandra Russo, and Krysia Broda. "Inductive Learning of Answer Set Programs". In: *Logics in Artificial Intelligence*. Ed. by Eduardo Fermé and João Leite. Cham: Springer International Publishing, 2014, pp. 311–325. ISBN: 978-3-319-11558-0.

[34] Emily LeBlanc, Marcello Balduccini, and Joost Vennekens. "Explaining Actual Causation via Reasoning About Actions and Change". In: *Logics in Artificial Intelligence*. Ed. by Francesco Calimeri, Nicola Leone, and Marco Manna. Cham: Springer International Publishing, 2019, pp. 231–246. ISBN: 978-3-030-19570-0.

[35] David Lewis. "Causation". In: *Journal of Philosophy* 70.17 (1973), pp. 556–567. DOI: 10.2307/2025310.

[36] Alison McIntyre. "Doctrine of Double Effect". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University, 2023.

[37] Michael S. Moore. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press, 2009.

[38] Erik T. Mueller. *Commonsense Reasoning: An Event Calculus Based Approach*. 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2014. ISBN: 9780128016473.

[39] Charles L. Ortiz Jr. "Explanatory update theory: Applications of counterfactual reasoning to causation". In: *Artificial Intelligence* 108.1 (1999), pp. 125–178. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00004-1. URL: https://www.sciencedirect.com/science/article/pii/S0004370299000041.

[40] L. A. Paul and Ned Hall. *Causation: A User's Guide*. Oxford University Press, Apr. 2013. ISBN: 9780199673445. DOI: 10.1093/acprof:oso/9780199673445.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199673445.001.0001.

[41] Luís Moniz Pereira and Ari Saptawijaya. "Modelling Morality with Prospective Logic". In: *Progress in Artificial Intelligence*. Ed. by José Neves, Manuel Filipe Santos, and José Manuel Machado. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 99–111. ISBN: 978-3-540-77002-2.

[42] Lewis Petrinovich and Patricia O'Neill. "Influence of wording and framing effects on moral intuitions". In: *Ethology and Sociobiology* 17.3 (1996), pp. 145–171. ISSN: 0162-3095. DOI: https://doi.org/10.1016/0162-3095(96)00041-6. URL: https://www.sciencedirect.com/science/article/pii/0162309596000416.

[43] Franziska Poszler, Edy Portmann, and Christoph Lütge. "Formalizing ethical principles within AI systems: experts' opinions on why (not) and how to do it". In: *AI and Ethics* (Feb. 2024). ISSN: 2730-5961. DOI: 10.1007/s43681-024-00425-6. URL: https://doi.org/10.1007/s43681-024-00425-6.

[44] T.M. Powers. "Prospects for a Kantian Machine". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 46–51. DOI: 10.1109/MIS.2006.77.

[45] R. Reiter. "A logic for default reasoning". In: *Artificial Intelligence* 13.1 (1980). Special Issue on Non-Monotonic Logic, pp. 81–132. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(80)90014-4. URL: https://www.sciencedirect.com/science/article/pii/0004370280900144.

[46] Nicholas Rescher. "How Wide Is the Gap Between Facts and Values?" In: *Philosophy and Phenomenological Research* 50 (1990), pp. 297–319. ISSN: 00318205. URL: http://www.jstor.org/stable/2108045 (visited on 03/14/2024).

[47] Camilo Sarmiento et al. *Action Languages Based Actual Causality for Computational Ethics: a Sound and Complete Implementation in ASP*. 2022. eprint: arXiv:2205.02919.

[48] Carolina Sartorio. "575 Causation and Ethics". In: *The Oxford Handbook of Causation*. Oxford University Press, Nov. 2009. ISBN: 9780199279739. DOI: 10.1093/oxfordhb/9780199279739.003.0027. eprint: https://academic.oup.com/book/0/chapter/357702810/chapter-ag-pdf/45510563/book_42621_section_357702810.ag.pdf. URL: https://doi.org/10.1093/oxfordhb/9780199279739.003.0027.

[49] Carolina Sartorio. "Causation and Responsibility". In: *Philosophy Compass* 2.5 (2007), pp. 749–765. DOI: 10.1111/j.1747-9991.2007.00097.x.

[50] Carolina Sartorio. "Disjunctive Causes". In: *The Journal of Philosophy* 103.10 (2006), pp. 521–538. ISSN: 0022362X. URL: http://www.jstor.org/stable/20619970 (visited on 03/08/2024).

[51] Ken Satoh and Satoshi Tojo. "Disjunction of Causes and Disjunctive Cause: a Solution to the Paradox of Conditio Sine Qua Non using Minimal Abduction". In: *International Conference on Legal Knowledge and Information Systems*. 2006. URL: https://api.semanticscholar.org/CorpusID:17546287.

[52] Jonathan Schaffer. "Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation". In: *Contemporary Debates in Philosophy of Science*. Ed. by Christopher Read Hitchcock. Blackwell, 2004, pp. 197–216.

[53] Rolf Schwitter. *Controlled Natural Language Processing as Answer Set Programming: an Experiment*. 2014. arXiv: 1408.2466 [cs.CL].

[54] Samuel T. Segun. "From machine ethics to computational ethics". In: *AI & SOCIETY* 36.1 (Mar. 2021), pp. 263–276. ISSN: 1435-5655. DOI: 10.1007/s00146-020-01010-1. URL: https://doi.org/10.1007/s00146-020-01010-1.

[55] L. A. Selby-Bigge, ed. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Oxford: Oxford University Press, 1975.

[56] Murray Shanahan. "The Event Calculus Explained". In: *Artificial Intelligence Today: Recent Trends and Developments*. Ed. by Michael J. Wooldridge and Manuela Veloso. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 409–430. ISBN: 978-3-540-48317-5. DOI: 10.1007/3-540-48317-9_17. URL: https://doi.org/10.1007/3-540-48317-9_17.

[57] Murray Shanahan. "The Frame Problem". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University, 2016.

[58] Peter Singer. "Sidgwick and Reflective Equilibrium". In: *The Monist* 58.3 (1974), pp. 490–517. DOI: 10.5840/monist197458330.

[59] Zeerak Talat et al. "A Word on Machine Ethics: A Response to Jiang et al. (2021)". In: *CoRR* abs/2111.04158 (2021). arXiv: 2111.04158. URL: https://arxiv.org/abs/2111.04158.

[60] Suzanne Tolmeijer et al. "Implementations in Machine Ethics: A Survey". In: *ACM Comput. Surv.* 53.6 (Dec. 2021). ISSN: 0360-0300. DOI: 10.1145/3419633. URL: https://doi.org/10.1145/3419633.

[61] Fiona Woollard and Frances Howard-Snyder. "Doing vs. Allowing Harm". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University, 2022.

[62] John P. Wright. *Hume's "A Treatise of Human Nature": An Introduction*. Cambridge Introductions to Key Philosophical Texts. Cambridge University Press, 2009.

[63] Richard W. Wright. "Causation in Tort Law". In: *California Law Review* 73.6 (1985), pp. 1735–1828. ISSN: 00081221. URL: http://www.jstor.org/stable/3480373 (visited on 03/13/2024).

[64] Richard W. Wright. "The Ness Account of Natural Causation: A Response to Criticisms". In: *Critical Essays on "Causation and Responsibility"*. Ed. by Markus Stepanians and Benedikt Kahmen. De Gruyter, 2013, pp. 13–66.

[65] Han Yu et al. "Building Ethics into Artificial Intelligence". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, 2018, pp. 5527–5533. ISBN: 9780999241127.