Saying “No” Truthfully

What the Murderer at the Door Really Asked

Title: Saying “No” Truthfully: What the Murderer at the Door Really Asked

Student: Halgir Winther Nagata

Course: Master’s Thesis

Date: June 2nd, 2025

Contents

1.0 Introduction

Immanuel Kant's moral philosophy famously demands strict adherence to duty, even when doing so appears to conflict with common moral intuition. One of the most controversial illustrations of this is the so-called axe murderer scenario[1], in which Kant argues that it is always wrong to lie, even to a would-be murderer seeking your friend. This claim has drawn extensive criticism, with many seeing it as a prime example of Kantian ethics' inflexibility and disregard for consequences. The scenario provokes the intuition that preventing an imminent murder should override any abstract principle, and that truth-telling in this context seems not only morally inadequate but morally complicit.

This thesis offers a novel defense of Kant's position. Rather than challenging his prohibition on lying, I propose that the standard interpretation of the scenario mischaracterizes the moral situation. Using tools from the philosophy of language - specifically speech act theory, contextual meaning, and inferentialism - I argue that the murderer's question, "Is your friend inside?", should not be treated as a neutral request for information. Instead, it should be understood as an implicit attempt to coerce complicity in an immoral act. Understood this way, responding "no" is not a lie, but a truthful rejection of a demand for moral participation.

By reframing the question in this way, I show that Kant's categorical imperative can be preserved without sacrificing moral intuition. This reinterpretation offers a way out of the dilemma without appealing to consequentialist reasoning, and opens the door to broader applications where linguistic context alters the moral meaning of speech acts. The thesis begins by establishing the ethical framework of the discussion, situating Kantian deontology alongside consequentialism and virtue ethics. It then reviews Kant's own views on truthfulness, along with key contemporary discussions, before introducing theoretical tools from the philosophy of language. These tools are applied to reinterpret the axe murderer scenario and defend the central claim. Finally, the thesis explores how the insights

[1] Although commonly referred to as the "axe murderer" scenario, Kant himself does not describe the would-be murderer as wielding an axe. The label has become a pedagogical shorthand in contemporary discussions of the example.

developed can inform broader ethical challenges, particularly in the domain of AI alignment.

1.1 Background

Kant's moral system is grounded in the idea that morality must be based on rational principles that apply universally. This is expressed through the categorical imperative, which commands that we act only according to maxims we could will to become universal law (Kant, 1998, p. 31). Within this framework, truthfulness is considered a perfect duty - one that allows no exceptions. Kant insists that lying is always wrong, regardless of the potential consequences, because it undermines the conditions of mutual trust necessary for moral action (Kant, 1889, p. 2).

The infamous axe murderer scenario illustrates the tension this creates. In this thought experiment, a murderer arrives at your door asking whether your friend - whom you are hiding - is inside. Kant claims that it would still be wrong to lie, even to protect your friend's life. This conclusion has struck many as counterintuitive, and it has become a focal point for critiques of Kantian rigidity.

This thesis does not attempt to soften or revise Kant's prohibition on lying. Instead, it questions the assumption that the murderer's question is a simple request for information. Drawing on philosophy of language, I will explore how the meaning of a question can depend heavily on context, intent, and the pragmatic functions of language.

Three theoretical tools will be especially important. Speech act theory, developed by thinkers such as Austin and Searle, explores how saying something is not only a matter of conveying information but also a way of performing actions - promising, commanding, threatening, or, crucially, participating. Wittgenstein's later philosophy stresses the importance of context and practice: meaning is not fixed in the words themselves but arises from how they are used within specific social interactions, or "language games". Finally, Brandom's inferentialism deepens this analysis by emphasizing that meaning is shaped by the network of commitments and entitlements speakers take on when they make claims, positioning themselves within a web of social and normative relations. These

concepts together make possible a richer reading of the murderer's question - one that opens the door to rejecting the demand without violating Kant's duty of truthfulness.

These concepts will allow me to argue that the question "Is your friend inside?" is not ethically neutral - it functions as an implicit attempt to secure your complicity. This reframing will serve as the foundation for the central claim of the thesis.

1.2 Thesis Statement

Kant's axe murderer scenario is often criticized for its moral rigidity. This thesis argues that the murderer's question, "Is your friend inside?", implicitly demands complicity in murder. Using philosophy of language, I show that answering "no" can be a truthful rejection rather than a lie, resolving the scenario's ethical tension without violating Kant's categorical imperative.

2.0 Theoretical framework

This section surveys the theoretical landscape relevant to the reinterpretation of Kant's axe murderer scenario. It is structured in three main parts: first, an overview of key ethical frameworks - including Kantian deontology and its contrast with consequentialism and virtue ethics - that position the moral debate; second, discussion of debates within Kantian ethics concerning the duty of truthfulness; and third, an exploration of theories from the philosophy of language that clarify how speech acts function in context. Together, these perspectives establish the conceptual foundations for this thesis's central question: whether the murderer's question can be understood in a way that permits a morally justified and truthful refusal, without violating Kant's ethical commitments. By positioning the argument within both ethical theory and linguistic analysis, this review prepares the ground for a reinterpretation that remains faithful to Kant while drawing on pragmatic insights into the moral structure of language.

2.1 Ethical frameworks

This section provides a brief overview of the major ethical theories relevant to the argument that follows. While the core of this thesis is situated within Kantian deontology, the broader debate over the duty to truthfulness - especially in high-stakes scenarios like the axe murderer case - frequently draws on competing moral frameworks. Most notably, consequentialism challenges the idea that duties must be followed regardless of outcomes, while virtue ethics shifts attention away from rules or results and toward the moral character of the agent.

By situating the discussion within this broader landscape, the thesis aims to clarify its commitments and the nature of the problem it addresses. The goal is not to resolve long-standing disagreements between these traditions, but to show how a specifically Kantian conception of duty and truthfulness can remain philosophically defensible when examined through the lens of linguistic context and speech act theory.

2.1.1 Kantian deontology

Kantian deontology is a moral framework rooted in the principle that right action is determined not by consequences but by reason. According to Kant, morality arises from the rational will, and the moral worth of an action depends on the maxim from which it is performed - the principle the agent acts upon - not on the action's outcomes. What makes an action morally right is that its maxim can be willed as a universal law: that is, it could be consistently adopted by all rational agents without contradiction. This idea is encapsulated in the categorical imperative, which Kant formulates in several complementary ways.

The core expression of Kant's moral theory is the Formula of Universal Law. It asks whether the maxim guiding one's action could be willed as a universal law that all rational agents would follow. This is not a test of whether a maxim is popular or desirable, but whether it is logically coherent when universalized. If the maxim would generate a contradiction when applied universally, then acting on it is impermissible. The Formula of Universal Law focuses on the consistency and rational structure of moral principles.

> "*There is, therefore, only a single categorical imperative and it is this:* act only in accordance with that maxim through which you can at the same time will that it become a universal law." (Kant, 1998, p. 31)

The second is the Formula of Humanity, which commands that one must always treat humanity, whether in oneself or in others, as an end in itself, and never merely as a means. This formulation highlights the intrinsic moral worth of rational beings. It prohibits any action that uses others for one's own purposes without regard for their autonomy or capacity for rational choice. Lying, coercion, and manipulation are violations of this principle, since they involve undermining the other's status as a rational agent.

> "*The practical imperative will therefore be the following:* So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means." (Kant, 1998, p. 38)

The third is the Formula of the Kingdom of Ends, which asks us to regard ourselves and others as members of a community of rational agents, each of whom is both subject to and author of the moral law. This formulation brings together the universal and personal dimensions of morality. It envisions moral agents not as isolated individuals, but as co-legislators in a shared moral order governed by reason.

> "*The concept of every rational being as one who must regard himself as giving universal law through all the maxims of his will, so as to appraise himself and his actions from this point of view, leads to a very fruitful concept dependent upon it, namely that* of a kingdom of ends." (Kant, 1998, p. 41)

Kant maintains that moral duties are grounded in reason and therefore must be internally consistent. A central feature of his moral theory is that genuine duties cannot conflict. If two duties appear to be in contradiction in a particular case, this signals a misunderstanding of either the duties themselves or the situation to which they are being applied (Kant, 1998, p. 33). For Kant, the moral law must be

capable of guiding action without contradiction; otherwise, it would cease to be law in any meaningful sense. This commitment to consistency reflects his broader view that morality is not situational or contingent but universal and rationally coherent.

Crucial to Kant's system is the duty to truthfulness (Kant, 1991, p. 225). For Kant, truthfulness is not simply a social good or a pragmatic tool - it is a requirement of respecting others as rational beings. Lying is wrong not because of what it might lead to, but because it deprives others of the information they need to exercise their rational agency. When one lies, one treats another person as a mere means to an end, violating the Formula of Humanity. This prohibition applies universally, even in cases where lying might seem to serve benevolent purposes.

Kant's commitment to absolute truthfulness has provoked enduring controversy, especially when applied to extreme moral situations. The best-known example is the so-called axe murderer at the door scenario, where telling the truth appears to enable great harm. Kant, however, insists that even in such cases, one may not lie (Kant, 1889, p. 1). His position, while counterintuitive to many, reflects his conviction that the moral law must hold universally if it is to hold at all. To lie, even with the intention of preventing harm, is to violate a principle that cannot be coherently willed as a universal law. This scenario and surrounding literature will be examined in detail in the next section.

The present thesis works squarely within the Kantian deontological framework. It accepts the structure of duties, the authority of the categorical imperative, and the centrality of maxims in moral evaluation. At the same time, it seeks to reexamine how the duty to truthfulness operates when the moral meaning of a speech act is not transparent at the level of grammar or surface form. The approach taken here does not involve modifying or weakening Kant's system, but rather understanding how it should be applied in contexts where language itself has complex moral function. To do that, one must take seriously not only what is said, but what is being done in the act of speaking.

2.1.2 Consequentialism

Consequentialism is a broad family of moral theories that assess the rightness or wrongness of actions based on their outcomes. At its core, consequentialism holds that what makes an action morally right is its ability to bring about the best overall consequences (Bentham, 1789, p. 61). In contrast to deontological ethics, which evaluates actions by reference to their underlying principles or maxims, consequentialism is forward-looking: it judges actions by the effects they are likely to produce. If lying, stealing, or even harming someone results in a better overall outcome, then those actions may be not only permissible but morally required.

The most influential form of consequentialism is utilitarianism, developed by thinkers such as Jeremy Bentham and John Stuart Mill (Sinnott-Armstrong, Walter, 2023, §1). Utilitarianism evaluates actions by the principle of utility - commonly expressed as the imperative to maximize happiness or minimize suffering. According to this view, moral agents should act in ways that produce the highest total amount of well-being, often simplified to promoting the greatest good for the greatest number. Bentham formulated this in quantitative terms, proposing that pleasure and pain could be measured and compared.

> "*Pleasures then, and the avoidance of pains, are the* ends *that the legislator has in view; it behoves him therefore to understand their* value. *Pleasures and pains are the instruments he has to work with: it behoves him therefore to understand their force, which is again, in other words, their value*." (Bentham, 1789, p. 31)

Mill, while retaining the basic structure, introduced a qualitative dimension, suggesting that some pleasures are more valuable than others (Mill, 2009, p. 16). Despite their differences, both thinkers maintained that moral decision-making requires weighing alternatives and choosing the one with the best overall results.

This approach stands in direct contrast to Kantian deontology. For a consequentialist, moral rules are only useful to the extent that they generally promote good outcomes. They may be broken, without moral fault, when doing

so would prevent harm or produce a greater benefit. In cases where following a rule would lead to suffering or injustice, a consequentialist will typically favor breaking the rule. Lying is not inherently wrong according to this view; it is wrong when it leads to more harm than good, and right when it prevents harm or promotes wellbeing.

In the axe murderer scenario, a consequentialist would generally argue that lying to the would-be murderer is morally obligatory, since doing so would likely prevent a killing. Kant, by contrast, holds that the duty to truthfulness is exceptionless. To lie, even to save a life, is to violate a moral law that must hold universally. The conflict between these two positions has become a focal point in debates about the limits of rule-based ethics and the role of context in moral reasoning.

This thesis is situated firmly within the deontological tradition. It does not attempt to resolve the axe murderer problem by appealing to consequences or justifying lies on the basis of predicted outcomes. However, understanding the consequentialist challenge is essential, because it helps clarify what is at stake in defending Kant's position. The central contribution of this thesis is to show that one need not abandon Kant's core commitments to truthfulness and universal law in order to respond intelligibly to moral pressure. Instead, the solution lies in reinterpreting what is being asked of the moral agent - not in sacrificing principle for outcome.

2.1.3 Virtue ethics

Virtue ethics is a tradition that grounds morality not in rules or outcomes, but in the character of the moral agent. Rather than asking what one ought to do in a given situation, virtue ethics asks what kind of person one ought to be. The focus is on cultivating stable moral dispositions - virtues - such as honesty, courage, temperance, and justice (Aristotle, 2004, p. 23). When properly developed, these traits guide the agent toward appropriate action in a wide range of situations, without the need for fixed rules or calculative reasoning about consequences.

The roots of virtue ethics lie in ancient philosophy, especially in the work of Aristotle, who described virtue as a mean between extremes and emphasized the development of moral character through habituation (Aristotle, 2004, p. 29). For Aristotle, the goal of ethics is to achieve eudaimonia - often translated as “flourishing” or “living well” - which requires not only external goods, but also the cultivation of intellectual and moral virtues (Aristotle, 2004, p. 206).

While virtue ethics remains an influential and valuable tradition, it plays a limited role in the present thesis. The debate surrounding the axe murderer scenario is typically framed in terms of a conflict between rule-based ethics (as exemplified by Kantian deontology) and consequence-based reasoning (as exemplified by utilitarianism). The central issue concerns whether there are moral absolutes, such as the duty not to lie, or whether such duties can be overridden by concern for outcomes.

Virtue ethics approaches this problem from a different angle: it might ask what a virtuous person would do in such a situation, or how honesty and compassion should be balanced. Such an approach might suggest that protecting a friend from harm, even at the cost of bending the truth, expresses virtues like compassion, loyalty, or courage.

However, this approach does not directly engage the specific structure of the tension between moral law and consequence that this thesis aims to address. Accordingly, virtue ethics is mentioned here for completeness and conceptual orientation, but it will not be developed further. The remainder of the thesis focuses on the tension between Kantian deontology and consequentialist reasoning, particularly as it emerges in interpretations of the duty to truthfulness.

2.1.4 Summary: ethical frameworks

This overview of ethical frameworks has outlined the major philosophical approaches relevant to the analysis of the axe murderer scenario. Kantian deontology emphasizes universal moral law and the inviolability of perfect duties, while consequentialism challenges this rigidity by insisting that moral rightness depends on outcomes. Virtue ethics, though less central to this thesis, reframes

the moral problem in terms of the character of the agent rather than strict duties or consequences.

While all three frameworks offer valuable perspectives, this thesis situates itself firmly within the Kantian tradition, accepting the binding force of moral law and the centrality of the categorical imperative. The challenge it addresses is not whether Kant's system should be abandoned in favor of outcome- or character-based approaches, but whether the duty to truthfulness, properly understood, leaves space for a morally justified refusal in cases where surface-level interpretations would demand complicity in wrongdoing. The following sections turn to Kant's own writings and surrounding literature to examine this tension in detail.

2.2 The axe murderer scenario

The moral dilemma posed by the axe murderer scenario has become one of the most widely cited and heavily debated aspects of Kantian ethics. Kant's position - that one must not lie even to a would-be murderer asking for the location of an innocent victim - has drawn criticism for what many see as a morally implausible level of rigidity (Wood, 2008, p. 240). Yet much of the philosophical literature responding to this scenario does not reject Kant's system wholesale. Instead, it seeks to interpret, refine, or internally critique Kant's framework in order to resolve the tension between his commitment to moral law and our common-sense intuitions about moral responsibility.

This section considers three representative approaches to the issue. First, it presents Kant's own arguments on truthfulness, focusing on his understanding of perfect duties and the moral logic that underlies his categorical prohibition on lying. Next, it turns to Michael Cholbi's internal critique of Kant's stance, in which he argues - through what he calls the Deontic Symmetry Thesis - that Kant's moral reasoning should lead to the opposite conclusion: that one must lie to preserve the rational life of another. Finally, the section considers James Mahon's interpretive defense, which maintains that Kant's critics misread the scope of his prohibition by failing to distinguish lying from other forms of potentially deceptive but morally permissible speech, such as evasion or reticence.

Although these approaches differ in method and conclusion, they share a common assumption: that the moral content of the scenario lies in the truth or falsehood of the speaker's response. Whether they argue that the speaker must lie, may remain silent, or must tell the truth, they all treat the question "Is your friend inside?" as if it is a morally neutral inquiry whose ethical weight depends entirely on how the addressee responds. This thesis challenges that assumption. Rather than focusing on whether the answer is true or false, it proposes that the question itself carries implicit normative commitments that have been overlooked in the standard interpretations. By applying tools from the philosophy of language, the thesis will argue that the murderer's question is not a straightforward request for information but a morally loaded demand for complicity - one that may be truthfully refused without violating Kant's ethical principles.

2.2.1 Kant on telling the truth

Kant's ethics are rooted in a firm commitment to the universality and unconditional nature of moral principles. For Kant, moral worth does not depend on consequences but on acting out of duty, in accordance with principles that can be universalized (Kant, 1998, p. 31). Among these, the duty of truthfulness holds a particularly strict place.

> "*The greatest violation of man's duty to himself regarded merely as a moral being (the humanity in his own person) is the contrary of truthfulness, lying.*" (Kant, 1991, p. 225)

Kant's most concentrated defense of this position appears in his short essay *On a Supposed Right to Lie from Benevolent Motives*, a response to Benjamin Constant's argument that one may sometimes be morally obligated to lie - particularly when doing so would prevent unjust harm, such as murder. Constant argues that rights arise from social contracts, and that truth is only obligatory between those who have a right to it (Kant, 1889, p. 1).[2] Kant rejects this premise

[2] I was not able to find an English translation of the particular work where Constant argues this, so I accept Kant's presentation of his argument for the purpose of this thesis.

outright, asserting instead that truthfulness is not grounded in social relations or consequences but in the moral law itself (Kant, 1889, p. 2). To lie, even with benevolent intent, is to act on a maxim that could not be willed as a universal law without contradiction.

More fundamentally, Kant holds that lying undermines the very conditions that make moral law possible. A lie treats the listener as a mere means, manipulating their capacity for rational judgment. Worse still, lying severs the trust that binds human beings in a moral community. If lies are allowed even occasionally, the integrity of communication - and therefore of law and obligation - breaks down. Kant writes that a lie "always injures another; if not another individual, yet mankind generally, since it vitiates the source of justice" (Kant, 1889, p. 1).

Kant's rejection of lying is not naivety. He does not claim that people always tell the truth or that lies never go undetected. Rather, he insists that "every rational being must act as if he were by his maxims at all times a lawgiving member of the universal kingdom of ends" (Kant, 1998, p. 45), in which moral law is universally respected. This idealism is central to his moral philosophy: right action is not defined by outcome but by principle. Thus, even if telling the truth in a particular instance results in tragedy - such as the murder of one's friend - it does not follow that the truth-teller has done wrong. The blame lies only with the murderer, whose will is the source of harm.

This is the foundation upon which Kant's critics build their attack. The thought experiment involving the would-be murderer has become a staple in both introductory ethics and academic criticism of deontology. Yet Kant's position remains consistent: if moral law is to bind rational agents categorically, then truthfulness must remain inviolable, even when it may lead to harmful consequences. It is this very absolutism that invites reinterpretation - not in order to reject it, but to consider whether the speech act in question is being properly understood in the first place.

2.2.2 Cholbi on extending the duty of self-preservation

In *The Murderer at the Door: What Kant Should Have Said*, Michael Cholbi provides a rigorous critique of Kant's categorical prohibition on lying, aiming to reconcile Kantian deontology with common moral intuitions regarding the permissibility of lying to protect another person's life. Cholbi begins by clearly identifying what he perceives as Kant's central error: the claim that truthfulness, being a perfect duty, can never be overridden, even in situations where adhering to it would directly contribute to an innocent person's death.

> "*[...] the perfect duty of truthfulness could at least in principle be overridden by another perfect duty, that of self-preservation.*" (Cholbi, 2009, p. 31)

To expose this supposed error, Cholbi introduces the Deontic Symmetry Thesis (DST), which states that if an agent holds a perfect duty toward herself, then logically, she must hold an equivalent perfect duty toward others.

> "*Deontic Symmetry Thesis (DST): All other things equal, any act with deontic valence V performed by agent A in which A is also the act's patient will have the same valence V if another agent B is the patient of A's act instead, and vice versa.*" (Cholbi, 2009, p. 34)

Cholbi defends DST by appealing directly to Kant's own principle of treating rational nature as an end in itself. Rational beings, according to Kant, possess an absolute moral worth that must be universally respected. If one has a perfect moral obligation not to harm oneself, grounded in the inherent worth of rational nature, then this obligation, Cholbi argues, cannot be coherently restricted to one's own rational life; it necessarily and symmetrically extends to all rational beings (Cholbi, 2009, p. 20).

From this thesis, Cholbi derives a solution for Kant's murderer-at-the-door scenario. Kant argues that one may not lie even to a would-be murderer seeking the whereabouts of an innocent victim, asserting that to do so violates the perfect duty of truthfulness. However, Cholbi contends that Kant fails to recognize

another, equally binding perfect duty - the duty to preserve rational life. Since DST mandates symmetrical obligations toward oneself and others, preserving the rational life of another is no less a perfect duty than preserving one's own life. Cholbi emphasizes that in Kant's ethical system, perfect duties cannot coherently conflict, as that would indicate a fundamental incoherence within moral reasoning itself (Cholbi, 2009, p. 27). Thus, if truthfulness leads directly and predictably to the destruction of rational life, truthfulness must, on Kant's own terms, lose its binding moral status in that specific scenario.

Cholbi's conclusion - that Kantian ethics, correctly interpreted, would mandate lying in the murderer-at-the-door scenario - is unique precisely because it aims to remain firmly within the Kantian framework. He explicitly avoids consequentialist reasoning, arguing instead from Kant's fundamental principles of rational respect and universalizability. The duty of truthfulness retains its general status as perfect duty, but it is fundamentally constrained by the even more basic Kantian commitment to the preservation of rational nature itself. Cholbi thus positions his analysis as a correction and clarification of Kantian ethics, rather than as a rejection or critique from an external consequentialist standpoint.

2.2.3 Mahon clarifies Kant on lying

James Mahon offers an interpretive defense of Kant's prohibition on lying, not by revising the doctrine, but by clarifying what Kant means by a "lie". In both *Kant on Lies, Candour and Reticence* and *Kant and the Perfect Duty to Others Not to Lie*, Mahon argues that critics frequently misrepresent Kant's position by treating it as a blanket condemnation of all deceptive speech. Rather, Mahon argues that Kant's idea of a lie is narrow and specific.

> "*Those commentators who seek to criticize Kant's prohibition against lying would do well to pay attention to the precise but comparatively narrow scope of the prohibition.*" (Mahon, 2003, p. 123)

A lie, in Kant's moral vocabulary, is not merely any communicative act that misleads, but a very particular kind of speech act.

> "*According to Kant, a lie is an untruthful declaration. […] A person makes an* untruthful declaration *when she makes a statement that does not correspond to what she believes to be true, and invites someone to believe that statement to be true.*" (Mahon, 2003, p. 102)

This precision is crucial to Mahon's defense. By distinguishing lying from other forms of misleading communication - such as reticence, evasion, or omission - Mahon opens interpretive wriggle room for morally permissible action under Kantian ethics that does not violate the duty of truthfulness. In the case of the murderer at the door, for instance, Mahon argues that one may morally refuse to answer, or evade the question in a way that avoids direct falsehood (Mahon, 2003, p. 121). These alternatives, while potentially deceptive in outcome, do not involve the assertion of a falsehood and therefore do not qualify as lies under Kant's strict definition.

Mahon further argues that many criticisms of Kant's view are predicated on the false assumption that Kant requires full transparency in all communicative situations. This, Mahon suggests, is a caricature. On his reading, Kantian ethics allows for considerable moral subtlety so long as the agent avoids knowingly asserting what they believe to be false. Thus, Kant's absolutism about truthfulness does not need to be interpreted as a demand for complete candour (Mahon, 2003, p. 114).

Mahon's analysis centers on the structure of the speech act itself, limiting moral prohibition to cases of explicit false assertion. This reading offers an interpretation of Kant's doctrine in which moral agents retain the ability to resist immoral demands without violating the duty of truthfulness, so long as they avoid asserting what they believe to be false.

2.2.4 Summary: the axe murderer scenario

The debate within Kantian ethics has centered on how to interpret the duty of truthfulness in cases of moral conflict. Some, like Cholbi, argue that Kant's own principles lead to a duty to lie, while others, like Mahon, defend the prohibition by narrowing the definition of lying. These approaches differ in their evaluation of

Kant's consistency but share a common focus on the speaker's response. The next section shifts attention to the structure of the question itself, drawing on the philosophy of language to reconsider what kind of moral act the speech situation initiates.

2.3 Philosophy of language

Contemporary philosophy of language provides several conceptual tools for understanding speech not merely as the transmission of information, but as a form of action embedded in normative contexts. These tools help reveal how certain utterances, such as the murderer's question in Kant's scenario, may carry moral implications that are not captured by their surface meaning alone.

Speech act theory, as developed by J. L. Austin and further refined by John Searle, begins with the insight that to speak is to do something: to promise, warn, request, or command. Austin distinguishes between the locutionary act (what is said), the illocutionary act (what is meant or intended), and the perlocutionary act (the effect on the hearer). In the case of a question like "Is your friend inside?", this structure reveals that the utterance may not be a simple request for information, but an attempt to secure cooperation or coerce moral participation.

Ludwig Wittgenstein's later work, particularly his concept of language games, further supports the idea that the meaning of an utterance is determined by its role in a broader social activity. What a statement does - how it functions within a given context - matters as much as what it says. A question posed in the context of threat or power imbalance does not belong to the same language game as a question posed in good faith. Thus, to interpret the murderer's question as a morally neutral inquiry is to abstract it from the social and pragmatic context that gives it force.

Extending this, Robert Brandom's inferentialism frames meaning in terms of the normative roles that speech acts play within a discursive practice. For Brandom, to make a statement is to undertake a set of inferential commitments: one becomes responsible for the implications of what one says, for what it licenses others to conclude, and for how it fits into the wider web of reasons and obligations. On this

view, a question like “Is your friend inside?” is not merely a request for data - it initiates a normative exchange that presupposes certain roles and entitlements, including potential complicity. Brandom’s framework allows one to analyze how responding to the murderer involves more than reporting facts; it involves participating in a morally loaded structure of inference and accountability.

Together, these theories converge on the insight that language use is not morally neutral. The ethical weight of a speech act may lie not in the truth or falsity of its surface content, but in the social function it performs and the normative pressures it exerts. These insights will form the basis for reinterpreting the murderer’s question in a way that both explains the intuitive discomfort with Kant’s original response and preserves his core ethical commitments.

2.3.1 The linguistic turn

The linguistic turn refers to a major development in twentieth-century philosophy, marked by a shift in focus toward language as the primary medium through which philosophical problems are framed, explored, and understood (Wolf, 2025, §1.a). Rather than treating language as just a tool for conveying thought or representing reality, philosophers increasingly came to see language itself as shaping the possibilities of meaning, knowledge, and experience. The linguistic turn does not refer to a single theory or school but to a general tendency: the recognition that to make progress on philosophical questions, one must first investigate the structures, uses, and functions of language.

> “*Since traditional philosophy has been (so the argument goes) largely an attempt to burrow beneath language to that which language expresses, the adoption of the linguistic turn presupposes the substantive thesis that there is nothing to be found by such burrowing*.” (Rorty, 1992, p. 10)

While earlier analytic philosophers laid groundwork by examining the logical structure of language, the shift became fully consolidated and named with the publication of Richard Rorty’s edited volume *The Linguistic Turn*, which brought together key essays reflecting this new focus and cemented “the idea of a linguistic turn as a sea change in the history of philosophy” (Ramberg & Dieleman,

2024, §4.1). Rorty's collection helped crystallize the idea that across diverse philosophical fields - epistemology, metaphysics, ethics - the investigation of language was no longer peripheral; it had become central. Philosophers increasingly recognized that many longstanding puzzles were entangled with the way we formulate, express, and negotiate meaning in linguistic practices.

> "*I shall mean by 'linguistic philosophy' the view that philosophical problems are problems which may be solved (or dissolved) either by reforming language, or by understanding more about the language we presently use.*" (Rorty, 1992, p. 3)

This shift reshaped philosophical inquiry across domains. In epistemology, the focus turned to how knowledge claims are framed in language; in metaphysics, to how linguistic frameworks determine what is considered real or possible; and in ethics, to how moral concepts are expressed, enacted, and understood within social practices (Wolf, 2025, §5). The linguistic turn introduced a broader emphasis on pragmatics, use, and context. Meaning was no longer seen as something that simply attached to words or sentences in isolation, but as something that arose from their function within shared forms of life.

Several of the thinkers central to this thesis's argument stand within this broader context. Ludwig Wittgenstein, in his later work, emphasized that meaning is inseparable from use and that words gain their significance within "language games" governed by social rules and practices. J.L. Austin and John Searle, through speech act theory, explored how language functions as a form of action: when we speak, we do things like promise, warn, or command - not merely state facts. Robert Brandom extended these insights into a sophisticated account of inferential roles, arguing that to make a claim is to participate in a network of normative commitments, shaping what others are entitled to infer or expect.

For the purposes of this thesis, the relevance of the linguistic turn lies in its core insight that the moral significance of speech cannot be reduced to its literal or surface content. When evaluating duties like Kant's prohibition on lying, we must also attend to the pragmatic and normative dimensions of the communicative act. This approach helps uncover what is morally at stake when a speaker answers

a question - not just whether the answer matches an external fact, but what role the speaker plays in the interaction, what commitments they take on, and what they help bring about through their participation in a shared practice of meaning. The reinterpretation developed in the following sections builds directly on this orientation, bringing the tools of the linguistic turn to bear on one of the most famous moral puzzles in Kantian ethics.

2.3.2 Speech acts with Austin and Searle

Speech act theory, as developed by J. L. Austin and later expanded by John Searle, offers a foundational framework for understanding language not merely as a tool for communicating information but as a form of action in itself. This perspective enables a shift in moral analysis from focusing solely on the semantic content of utterances to considering the kinds of actions speakers perform when they speak. This is particularly relevant to the axe murderer scenario, where the moral stakes are not confined to what is said, but extend to the function of the question itself within the social and ethical situation in which it is uttered.

Austin's model distinguishes between three dimensions of a speech act. The locutionary act refers to the act of saying something and its literal meaning. The illocutionary act captures what the speaker is doing in saying it - asserting, asking, commanding, promising, and so forth. The perlocutionary act describes the effect the utterance has on the hearer, such as persuading, frightening, or convincing (Austin, 1962, p. 101). In the axe murderer scenario, the question "Is your friend inside?" appears, on its face, to be a locutionary act - a simple request for information. However, speech act theory draws attention to the illocutionary force embedded in such utterances, which may involve implicit directives, threats, or attempts to elicit cooperation. By analyzing the utterance at the level of illocution, it becomes possible to consider whether the question is in fact a coercive act rather than a morally neutral inquiry.

A key distinction in speech act theory is between propositional content and illocutionary force. The propositional content refers to the individually literal meaning of the words being said - the informational or factual component that can be evaluated for truth or falsity. By contrast, the illocutionary force concerns

what the speaker is doing by saying it: asking a question, making a promise, issuing a command, offering a warning, or performing some other communicative act (Searle, 1969, p. 30). This difference matters because the same propositional content can take on very different moral and practical significance depending on the illocutionary force behind it. In the context of the axe murderer scenario, this distinction invites us to look beyond the surface-level meaning of the question "Is your friend inside?" and ask what kind of act the speaker is performing, and what kind of participation they are implicitly demanding from the hearer.

Searle builds on Austin's framework by providing a taxonomy of illocutionary act types, including assertives (statements of fact), directives (attempts to get the hearer to do something), commissives (commitments to future action), expressives (expressions of attitudes), and declarations (speech acts that change the social world) (Searle, 1979, p. 12). A question such as "Is your friend inside?" may appear to be a directive masked as an interrogative - especially when issued in a context where one party holds power or intends harm.

> "*Thus, for example, the sentence, 'Could you do this for me?' in spite of the meaning of the lexical items and the interrogative illocutionary force-indicating devices is not characteristically uttered as a subjunctive question concerning your abilities; it is characteristically uttered as a request.*" (Searle, 1969, p. 68)

Searle emphasizes that illocutionary force is not just a matter of speaker intent but is also governed by shared social conventions and background assumptions (Searle, 1969, p. 35). In morally charged contexts, such as threats or interrogations, speech acts often function under implied pressure, and their surface form may be misleading with respect to their actual pragmatic role.

These insights are significant for evaluating the Kantian scenario. Kant's framework, as typically applied to this case, focuses on whether the speaker tells the truth or lies in response to the murderer's question. But if the question itself is not a neutral request, and instead operates as a coercive directive or a morally loaded invitation to complicity, then the speech act is not properly characterized

merely by its semantic content. From this perspective, evaluating the morality of the agent's response requires an analysis of the speech situation as a whole, including the act that initiated it.

The implications for Kantian ethics are considerable. If language functions as action, then the moral assessment of truth-telling cannot be limited to the literal truth-value of propositions. The agent must also consider the illocutionary structure of the exchange, and whether participating in the speech act amounts to endorsing or enabling moral wrongdoing. Austin and Searle provide the conceptual tools to analyze speech at this level, setting the stage for a reinterpretation of the axe murderer scenario in which Kant's prohibition on lying may be preserved - not by redefining what a lie is, but by challenging what sort of act the question really is.

2.3.3 Context and language games with Wittgenstein

Ludwig Wittgenstein's later philosophy, particularly as developed in *Philosophical Investigations*, offers a framework for thinking about meaning that resists abstraction and emphasizes the concrete practices in which language is embedded. Wittgenstein rejects the idea that words have meanings independent of how they are used. Instead, he famously claims that "the meaning of a word is its use in the language" (Wittgenstein, 1968, §43). This shift - from seeing meaning as a static relation between words and things to understanding it as a function of social activity - has major implications for evaluating moral speech, especially in contexts such as coercion, threat, or deception.

Wittgenstein introduces the concept of language games to explain how language functions as a rule-governed activity. A language game is a structured form of linguistic interaction governed by implicit or explicit norms, always situated within a particular form of life (Wittgenstein, 1968, §7). These forms of life are the background practices, institutions, and social conditions in which language is intelligible. Crucially, the same sentence can participate in different language games depending on the context: "Is your friend inside?" can be an innocent question, a test of loyalty, a veiled threat, or an implicit accusation, depending on who asks, to whom, and under what circumstances.

This framework is directly relevant to the axe murderer scenario. Standard treatments of the case treat the question as a factual inquiry, assuming that its meaning is transparent and stable across contexts. But from a Wittgensteinian perspective, the meaning of the murderer's question is not reducible to its words' propositional content. It functions differently in the context of threat, asymmetry, and imminent harm. The utterance does not merely seek information; it places moral and social pressure on the hearer, framing their possible responses within a narrow range of complicity. In this sense, the question belongs to a different language game than ordinary requests for information. Its function is shaped by the norms of coercion and by the hearer's recognition of the speaker's intentions and authority.

What follows from this is not simply that context influences meaning, but that the moral evaluation of speech must consider the language game in which the utterance occurs. If the question is not functioning as a genuine inquiry, then the hearer's moral obligations may not be those that apply in ordinary contexts of truth-telling. The utterance may be structurally defective or manipulative, such that the ethical demands it appears to place on the hearer are themselves immoral. This perspective supports the broader claim of this thesis: that the moral weight of an utterance lies not only in its propositional content but in its pragmatic function. Wittgenstein's emphasis on use, rule-following, and shared practice helps make room for a reinterpretation of the axe murderer scenario that avoids consequentialism without abandoning Kant's ethical principles.

2.3.4 Inferentialism with Brandom

Robert Brandom's inferentialist philosophy of language, developed in *Making It Explicit*, further shifts attention away from the literal content of utterances toward the normative structures they create and participate in. His work draws significantly on earlier insights developed by Wilfrid Sellars, who rejected the idea that meaning stems directly from a representational link between language and the world, arguing instead that meaning arises within norm-governed inferential practices (Brandom, 2001, p. 90). For Brandom, the meaning of a statement is not determined by reference or description alone, but by the inferential role it plays

within a broader network of claims, reasons, and commitments. To make an utterance is to position oneself within a space of reasons, undertaking specific obligations to justify the claim, to infer further consequences from it, and to recognize the entitlements and commitments it creates for others. This is to participate in the "*game of giving and asking for reasons*" (Brandom, 2001, p. 106).

Brandom describes language as a normative practice in which speakers and listeners are constantly managing claims and responsibilities in a practice Brandom calls *deontic scorekeeping* (Brandom, 2001, p. 142). An assertion is not simply the presentation of a fact; it is the act of staking a claim within a community governed by rules of reasoning and justification. Similarly, asking a question is not merely a request for information; it is an act that initiates a particular normative situation, creating expectations and potential obligations for the hearer. The content of what is said is inseparable from the social role it plays in the discursive practice.

Applying Brandom's framework to the axe murderer scenario highlights the normative pressures embedded in the murderer's question. "Is your friend inside?" does not function merely as a neutral request for information. It creates a structure of inferential expectations: that the hearer will respond cooperatively, that the information provided will be used for particular ends, and that the speaker and hearer are engaged in a recognizable social practice of question and answer. In the coercive context of the axe murderer scenario, responding even with a literal truth may involve accepting inferential commitments that contribute to immoral ends, namely the violation of rational nature through the murder of the victim.

This perspective reinforces the claim that the moral evaluation of speech acts cannot be confined to the truth or falsity of their surface content. From an inferentialist standpoint, participation in certain speech practices can itself be morally problematic if it involves adopting roles or commitments incompatible with respecting rational agents as ends. Brandom's work thus supports the reinterpretation of the axe murderer scenario not by altering Kant's ethical framework, but by revealing that the speech situation itself may be compromised at the level of inferential structure. What this means is that the problem lies not

merely in the literal content of the murderer's question but in the network of normative relations it invokes: by answering, the agent is drawn into a chain of reasoning aimed at an immoral outcome. The inferential commitments at play are already oriented toward wrongdoing, so that even a truthful answer risks becoming complicit in a process that violates Kantian moral demands. Recognizing the normative commitments one would undertake by responding to the murderer allows for the possibility of a truthful refusal - one that preserves both the duty to truthfulness and the duty to protect rational life.

2.3.5 Summary: philosophy of language

The philosophy of language provides essential tools for deepening our understanding of the moral dimensions of speech acts. As this section has shown, thinkers like Austin, Searle, Wittgenstein, and Brandom reveal that language is not merely a medium for conveying factual information but a site of action, commitment, and normative consequence. Speech acts perform roles, create expectations, and position speakers and hearers within complex networks of social meaning. For moral philosophy, this means that evaluating the ethical weight of what is said cannot stop at propositional content; it must also account for what the speaker is doing and what moral roles they are inhabiting by participating in a communicative exchange.

This insight is crucial for the present project, which aims to reinterpret Kant's duty to truthfulness not by weakening or bypassing his ethical system, but by applying it more rigorously to the pragmatic dimensions of language. The following sections will build on this foundation, arguing that the moral status of the agent's response to the axe murderer cannot be understood by focusing solely on literal truth or falsehood. Instead, it must be assessed in light of the broader structure of the speech act and the normative commitments it entails.

2.4 Summary: theoretical frameworks

The literature surveyed here spans ethical theory, focused debates within Kantian moral philosophy, and key insights from the philosophy of language. Together, these bodies of work outline the conceptual landscape in which the axe murderer problem has been debated: whether Kant's duty to truthfulness holds firm even in

extreme cases, and how the moral meaning of speech acts can be analyzed beyond their surface content.

While there has been substantial engagement with both the ethical and linguistic dimensions individually, little work has been done to bring these strands together in a systematic way. My thesis aims to fill that gap by using linguistic analysis to reinterpret the speech act at the center of the axe murderer scenario, offering a resolution that remains faithful to Kant's moral framework while addressing the intuitive tensions it has long provoked.

3.0 Critique & discussion

Before presenting a reinterpretation of the axe murderer scenario, it is necessary to engage with the most influential attempts to defend, modify, or reject Kant's position. This section focuses on two such efforts: Michael Cholbi's proposal to resolve the dilemma by appealing to a symmetrical duty to preserve rational life, and James Mahon's attempt to preserve Kantian truthfulness by appealing to non-deceptive forms of misleading communication. Both approaches aim to maintain fidelity to Kant's ethics while softening its perceived rigidity. However, as the following critiques will show, each position either introduces theoretical tensions or fails to fully reckon with the moral function of language. These shortcomings motivate the need for a deeper reconsideration of what is being asked in the scenario, and how Kant's ethics might respond without compromise.

3.1 Cholbi

Michael Cholbi attempts to resolve the tension in Kant's axe murderer scenario by proposing the Deontic Symmetry Thesis (DST). According to Cholbi, if one holds a perfect duty of self-preservation - as Kant affirms - then consistency requires recognizing an equivalent perfect duty to preserve the lives of others. Since rational nature is the ground of moral worth for all persons, not just oneself, Kant's framework must, Cholbi argues, extend the duty of preservation symmetrically. In the case of the murderer at the door, lying would therefore be not merely permissible but morally required, insofar as it is the only available means to preserve the victim's rational life (Cholbi, 2009, p. 46).

Cholbi's approach remains committed to many of Kant's core principles. He does not appeal to maximizing happiness or weighing outcomes in the utilitarian sense. Rather, he frames the duty to lie as the necessary implication of Kant's own respect for rational nature. In doing so, Cholbi attempts to correct what he sees as an inconsistency internal to Kant's system, rather than abandoning its deontological structure.

However, I argue that Cholbi's solution introduces consequentialist reasoning into Kant's framework. His argument depends on the foreseeable consequences of telling the truth or lying: the death or survival of the victim. While he frames the duty to preserve life as a perfect duty, the moral reasoning ultimately hinges on the effect of the agent's actions. If telling the truth predictably leads to the destruction of rational life, Cholbi concludes that the duty of truthfulness must yield. But this reasoning structurally resembles consequentialism: it evaluates the moral rightness of an action based on its expected outcomes, rather than the intrinsic form or universalizability of the maxim governing the action.

This shift undermines a central pillar of Kant's ethics. Kant insists that moral duties must be determined independently of empirical consequences; principles are to be universalized without regard for outcomes (Kant, 1998, p. 31). If duties are calculated based on which action better preserves rational nature in a particular case, then the deontological framework gives way to a morality based on outcomes. The agent becomes responsible for calculating effects rather than adhering unconditionally to rational principles.

Cholbi's argument also depends on an extension of Kant's duty of self-preservation that is not clearly supported by Kant's own texts. While Kant prohibits suicide on the grounds that it treats one's rational nature merely as a means, this is a prohibition directed at self-harming action (Kant, 1998, p. 32). Similarly, Kant condemns homicide as the direct destruction of another's rational nature. In both cases, the agent bears responsibility as the primary cause of the destruction. What Kant does not assert is that an agent has a perfect duty to preserve another person's life from destruction at the hands of a third party - particularly not by violating another perfect duty such as truthfulness. To infer

such a duty introduces an element of empirical foresight that Kant explicitly rejects as a basis for moral decision-making. In Kant's view, the moral worth of an action lies in its maxim, not in its consequences (Kant, 1998, p. 31). Cholbi's reinterpretation therefore depends not only on a consequentialist logic, but also on a normative symmetry between suicide and third-party murder that Kant himself does not endorse.

Cholbi's solution compromises Kant's insistence that morality must be grounded in pure reason, not empirical foresight. My solution will seek to remain within Kant's formalism by questioning whether the speech act itself demands a straightforward choice between truth and falsehood, rather than shifting the framework toward a consequentialist evaluation.

3.2 Mahon

James Mahon defends Kant's prohibition on lying by distinguishing it from other forms of misleading communication. In Mahon's view, while Kant strictly forbids knowingly asserting a falsehood with the intent to deceive, he does not prohibit other strategies such as reticence, evasive ambiguity, or even sarcasm (Mahon, 2003, p. 119). According to Mahon, these forms of communication may mislead but do not constitute lying because they do not involve direct assertions of falsehood. In the axe murderer scenario, for example, Mahon suggests that one might respond sarcastically, give ambiguous generalities, or refuse to answer directly, thereby misleading the murderer without violating the perfect duty to truthfulness (Mahon, 2003, p. 123).

However, Mahon's defense reveals a fundamental inconsistency in his treatment of language. In proposing strategies like sarcasm or ambiguity, Mahon implicitly acknowledges that communication transcends the literal meaning of words. Sarcasm, after all, derives its communicative force precisely from the divergence between what is said and what is meant, relying on shared social cues and context. By suggesting that one can speak deceptively through tone or indirectness without lying, Mahon concedes that language operates at the level of pragmatic understanding, not merely at the level of surface propositions.

Yet when evaluating the moral permissibility of these strategies, Mahon reverts to a narrow, literalist conception of speech. He treats the absence of a literal falsehood as sufficient to preserve moral innocence, disregarding the fact that the speaker's communicative act as a whole - including tone, implicature, and shared understanding - may functionally convey a falsehood. This selective treatment of language amounts to a philosophical inconsistency: Mahon invokes the richness of language when proposing morally permissible strategies, but denies that same richness when assessing whether those strategies satisfy Kant's moral demands.

A proper Kantian account of truthfulness must attend not merely to the literal words uttered, but to the full structure of communication between rational agents. If language is understood holistically - as a practice that includes tone, pragmatic inference, and mutual understanding - then a sarcastic "yes" universally recognized as meaning "no" is communicatively equivalent to an outright lie. Such speech acts, even if formally compliant with a narrow reading of Kant's prohibition, fail to meet the deeper duty of sincerity and respect for rational nature. The obligation is not simply to avoid literal falsehoods but to engage with others in ways that preserve their capacity for rational understanding and trust.

Mahon's strategy thus preserves technical compliance with Kant's duty to truthfulness at the cost of undermining its substantive moral purpose. By failing to account for the holistic nature of communication, his defense risks reducing Kant's ethics to a pedantic exercise in word choice, rather than a principled commitment to sincere and transparent engagement with other rational beings.

3.3 Summary: critique & discussion

These critiques of Cholbi and Mahon highlight both the strengths and the limitations of existing attempts to reconcile Kant's duty to truthfulness with the moral challenges of the axe murderer scenario. Cholbi's proposal to prioritize the preservation of rational life brings in consequentialist elements that put pressure on Kantian consistency, while Mahon's emphasis on evasive or non-deceptive forms of misleading speech risks weakening the sincerity and moral clarity that Kant's framework relies on. Despite their differences, both approaches share a common issue: they take the surface meaning of the murderer's question for

granted, focusing their solutions on how the agent ought to respond within that assumed frame.

My thesis argues that such approaches overlook a crucial dimension: the moral analysis of the speech act itself. Rather than seeking to revise Kant's ethics or find permissible forms of deception, the next sections propose a reinterpretation grounded in the philosophy of language. By paying closer attention to the pragmatic and inferential structure of the exchange, the reinterpretation aims to show that a truthful refusal is possible - not as an exception to Kant's framework, but as a careful application of its core principles to the deeper moral meaning of the situation.

4.0 Developing a pragmatic solution

Having reviewed the key positions in the literature and examined their limitations, this section presents the central argument of the thesis: a pragmatic reinterpretation of the axe murderer scenario grounded in insights from the philosophy of language. Rather than attempting to modify or soften Kant's ethical framework, this approach focuses on reanalyzing what is actually being asked when the murderer poses their question. Drawing on theoretical tools from the philosophy of language, the following subsections explore how the moral meaning of a speech act depends not only on its propositional content, but on its function, context, and normative implications. By reexamining the speech situation itself, this section develops a Kantian-compatible solution that preserves the duty to truthfulness without relying on evasive strategies or consequentialist reasoning. My argument proceeds in several stages, first unpacking the linguistic structure of the murderer's question, then demonstrating why answering "no" can be both truthful and morally required, and finally addressing potential objections to the reinterpretation.

4.1 Interpreting the murderer's question

The question "Is your friend inside?" is typically read as a simple yes-or-no question - a request for factual information. That reading has shaped how many interpreters of Kant approach the problem: if the friend is in fact inside, then

saying “no” must be a lie, and lying is always wrong. From this perspective, Kant’s strict rule against lying appears cold and rigid. It seems to demand that we ignore what is at stake in the situation and focus only on the formal duty to tell the truth.

But this way of framing the scenario takes the question at face value. It assumes that the moral problem lies entirely in how the agent responds, while the question itself is morally neutral. This assumption is hard to defend once we take seriously what we know about how language works. A question is not just a string of words or a grammatical structure - it is an action performed in a particular context, with a particular purpose. That action can carry moral weight depending on what the speaker is doing by asking it.

Even in everyday conversation, we often recognize that a question isn’t always just a request for information. If a parent asks a child “Did you do your homework?”, the child usually understands that more is being said than just the literal words. The question might carry an expectation, a warning, or even a veiled accusation. Its meaning depends not only on the words used, but on how they are said, who is speaking, what they know, and what the consequences of answering might be.

In the case of the axe murderer, the context transforms the question entirely. The speaker is not a neutral party making a factual inquiry - they are someone who is looking for a victim and asking for help in finding them. The person being questioned is under pressure, faced with a life-or-death situation, and is being asked to provide information that will likely be used to commit murder. All these facts shape the meaning of the question. It no longer functions as a simple inquiry; it becomes something much more morally loaded.

This section of the thesis begins a shift in how we approach the scenario. Rather than asking whether the agent should tell the truth in response to the question, we need to first ask what kind of act the question itself is. To do that, we need to move beyond grammar and look at the question in context - as a speech act that performs a social and moral function. In the following sections, I will use different theories from the philosophy of language to argue that the murderer’s question is not morally innocent and that it may not mean what it appears to mean on the surface.

4.1.1 Austin and Searle - illocutionary force

Speech act theory, particularly as developed by J. L. Austin and John Searle, provides a framework for understanding how language functions not merely to state facts but to perform actions. Austin's central insight is that when we speak, we are often doing something - making a promise, issuing a command, giving a warning, or making an accusation. He distinguishes between the locutionary act (the literal meaning of the utterance), the illocutionary act (what the speaker is doing in saying it), and the perlocutionary act (the effect it has on the hearer). This framework shifts attention away from words as static carriers of content and toward speech as an event with force and intention.

Searle builds on this by classifying illocutionary acts into types - such as assertives, directives, commissives, expressives, and declarations - and by emphasizing that every speech act functions within a system of social rules and shared expectations. A question, in this view, is not just a request for information. It belongs to the category of directives, which aim to get the hearer to do something - provide an answer, reveal a truth, follow an expectation. Whether a question functions innocently or coercively depends on the context in which it is asked and the power dynamics between speaker and hearer.

In the axe murderer scenario, the illocutionary force of the question "Is your friend inside?" cannot be analyzed in isolation from the situation in which it is spoken. Although it takes the grammatical form of a yes-or-no inquiry, it does not function as a mere request for clarification. Given the speaker's intent (to locate and kill someone), the question takes on a coercive character. It becomes a way of applying pressure, an attempt to compel the hearer into a particular course of action. It may even carry the implicit force of a threat - conveying that refusal to answer will cause suspicion or violence. This is not an act of shared inquiry between rational agents; it is an attempt to enlist the hearer in a morally compromised plan.

Crucially, the speaker's intent and the hearer's understanding of that intent are what determine the illocutionary force of the utterance. The question is not morally neutral simply because it asks for information. It is, in this context, a

performative act that sets moral expectations. The speaker is doing something by asking: demanding help, testing loyalty, or coercing participation. A response, even if factually accurate, may therefore function as a form of cooperation - not because of what is said in isolation, but because of what it does within the exchange.

Austin and Searle's framework makes clear that the moral weight of a speech act is not determined solely by its literal meaning but by what the speaker is doing in saying it. In the case of the murderer's question, what the speaker is doing is not merely asking - it is attempting to involve the hearer in an act of violence. This analysis supports the idea that responding "no" may be a truthful rejection of that act, rather than a lie. The question is not an innocent one, and it does not demand a literal answer. It demands a moral response to a coercive speech act, and "no" may be the only response that respects that fact.

4.1.2 Wittgenstein - language games

Ludwig Wittgenstein's later philosophy, especially as developed in *Philosophical Investigations*, offers another way to challenge the assumption that the murderer's question is a morally neutral request for information. Wittgenstein rejects the idea that words have fixed meanings independent of the situations in which they are used. Instead, he claims that the meaning of a word is its use in the language, and that language is part of a larger network of practices, or what he calls a form of life. What counts as a meaningful statement or question is not determined by abstract definitions but by the role an utterance plays within a particular pattern of human activity.

Wittgenstein's concept of language games is especially relevant. Language, for him, is not a uniform code applied across contexts but a collection of many distinct practices, each with its own implicit rules and expectations. Asking a question, making a promise, giving a command, or confessing something all belong to different language games. The same sentence can play very different roles depending on the context. "Is your friend inside?" might be a casual inquiry during a dinner party, a procedural question in an emergency drill, or, in the axe murderer scenario, something much more charged.

In the context of Kant's example, the speaker and the hearer are not participating in a shared game of rational cooperation. There is no mutual understanding or trust. Instead, the question occurs in a breakdown of ordinary communicative norms: the speaker intends to use the information to kill someone, and the hearer knows this. Under such conditions, the question loses its place in the ordinary language game of inquiry. It becomes a different kind of move entirely - one that takes on its meaning from the surrounding circumstances of threat, coercion, and moral danger.

This matters because the hearer cannot be expected to treat the question as if it belonged to a neutral context. Language is not detachable from its social environment. The hearer understands "Is your friend inside?" not as a request for disinterested clarification, but as a hostile move in a game with deadly stakes. Wittgenstein's insight is that words do not carry fixed moral significance in themselves; their moral weight comes from how they are used, and what roles they play in the lives of the people who use them. Here, the utterance functions less like a question and more like a moral test, a demand for complicity, or a veiled threat.

By viewing the question through Wittgenstein's framework, we can see that its surface form is misleading. It may look like a simple inquiry, but it operates within a context that transforms its meaning entirely. The hearer's moral responsibility, then, is not to respond to the words in isolation but to recognize what kind of interaction they are part of. This reinforces the idea that truthfulness, properly understood, must take account of the moral function of language in context, not just its grammatical shape.

4.1.3 Brandom - inferentialism

Robert Brandom's philosophy of language deepens the argument that the murderer's question cannot be assessed by its surface form alone. In *Making It Explicit*, Brandom develops a theory of meaning grounded in inferential roles: to speak is not just to express a thought, but to place oneself within a network of commitments and entitlements. When someone says something, they take on

responsibilities for what follows from it and what it implies. Communication is not simply descriptive; it is a matter of doing things in a normative space, where statements carry consequences and obligations.

Under Brandom's account, even a seemingly simple assertion or answer functions within a broader practice of giving and asking for reasons. Saying "yes" in response to "Is your friend inside?" is not merely confirming a fact - it is taking part in a normative exchange. The speaker becomes responsible not only for the truth of the utterance, but for its uptake: what others are entitled to do or infer on the basis of what has been said. In the axe murderer case, this means that the one who says "yes" plays an active role in the reasoning process that leads to the victim's death, not simply by conveying information, but by licensing inferences that are used to carry out an immoral act.

This perspective sharpens the moral analysis. If participating in a conversation involves more than literal speech - if it involves endorsing certain inferential consequences - then answering the murderer's question becomes more than a matter of factual accuracy. It becomes a matter of what norms one is helping to enact or sustain. A Kantian concerned with moral integrity cannot ignore these dimensions. Even a seemingly true statement may become morally compromised if it plays a role in enabling a plan that violates the dignity of rational nature.

Brandom's view supports the claim that truthfulness cannot be reduced to propositional accuracy. It must also take into account what one is committing to by speaking and what one is helping others do with that speech. In this case, affirming that the friend is inside carries the inferential implication that the murderer may proceed to act on that information. The speaker, by answering, becomes entangled in that inferential and moral structure.

Brandom further clarifies this distinction with a well-known example, drawn from Sellars, of a parrot trained to say "This is red" when shown red objects (Brandom, 2001, p. 88). While the parrot's utterance matches the correct propositional content, it does not understand or participate in the normative structure that gives that statement meaning within human linguistic practice. The parrot does not grasp that saying "This is red" carries inferential commitments, such as

implying that the object is not green, that it is colored, or that it might justify certain expectations or actions. Nor can the parrot recognize that others might challenge the claim or ask for reasons, or that it could be held accountable within a space of giving and asking for reasons. For Brandom, meaning arises not merely from the production of truth-sounding sentences but from occupying a role within a normative, inferential practice.

This example highlights why the Kantian moral agent cannot treat their answer to the murderer as a simple delivery of factual content. To do so would be to reduce themselves, functionally, to the level of the parrot - producing a response without recognizing the moral and inferential stakes embedded in the exchange. By contrast, the human agent is responsible precisely because they do understand the role they are inhabiting, the commitments they take on, and the consequences that follow from what they say. The reinterpretation developed in this thesis depends on this inferentialist insight: moral speech acts must be evaluated not only by their literal truth but by their position within a broader normative framework that shapes what participants are doing and what they are helping to bring about.

Seen through Brandom's framework, the question is not just "Is your friend inside?" but also "Will you cooperate in the reasoning chain that leads to your friend's death?". A truthful moral agent, in this view, must respond not to the literal question, but to the broader act of reasoning they are being drawn into. Refusing to assent - even by saying something that would be false on a narrow reading - may therefore represent a deeper form of truthfulness, one that preserves the agent's integrity in the face of coercive or corrupted reasoning.

4.1.4 Reframing the question: from form to function

The preceding sections have argued that the question "Is your friend inside?" cannot be treated as a morally neutral request for information. While its surface grammar may suggest a factual inquiry, the function of the utterance is shaped by its context - its use within a situation of threat and moral asymmetry. Analyzing the question with tools from the philosophy of language reveals that it performs a different kind of act than traditional interpretations assume.

Speech act theory shows that the question functions not merely as an inquiry but as a coercive or complicit directive - a way of enlisting the hearer into an immoral plan. Wittgenstein's concept of language games highlights how the meaning of the question changes in this context: it does not belong to the ordinary game of neutral questioning but to a morally loaded exchange governed by different expectations. Brandom's inferentialism makes clear that responding to the question is not simply offering information - it is stepping into a chain of reasoning and social commitment, one that implicates the speaker in the outcome. Each framework reinforces the conclusion that the question is not what it appears to be.

Together, these perspectives reframe the moral problem. The issue is not whether one may tell a literal falsehood in response to a straightforward question. The issue is how to respond truthfully to a speech act that, when properly understood, is not asking for facts but demanding cooperation. Understood in this way, the utterance "Is your friend inside?" is better interpreted as an implicit moral test: "Will you assist me in doing harm?". The next section turns to the implications of this reinterpretation for Kant's duty of truthfulness and the possibility of answering "no" without lying.

4.2 Why "no" is a truthful and morally required answer

If we accept that the question "Is your friend inside?" is not a neutral factual inquiry but a coercive speech act - a demand for complicity, masked as a question - then the moral terms of the situation change. The agent is not simply being asked for a piece of information; they are being asked to participate in an act whose purpose is to destroy rational life. This distinction matters, because Kant's duty of truthfulness forbids asserting falsehoods with the intention to deceive, but it does not require one to treat every utterance as morally innocent simply because it takes the grammatical form of a question.

On a narrow reading, responding "no" when the friend is in fact inside might seem to be a lie. But under the analysis developed in the previous section, the speech act in question is not actually asking whether the friend is in the house. It

is, functionally, asking something else: Will you help me murder your friend? Viewed in that light, the response “no” is not a lie at all - it is a truthful rejection of the actual demand being made. The moral significance of an utterance lies not just in its literal propositional content, but in the context and norms of the speech exchange in which it occurs.

This way of responding aligns with Kant’s principle that rational agents must act from maxims that can be universalized. The maxim behind saying “no” in this case is not “I will lie to prevent harm,” but rather something like: When I am addressed in a morally corrupted speech act that implicitly demands my participation in wrongdoing, I will respond truthfully to what is actually being asked. That maxim does not involve deception; it recognizes the moral character of the interaction and refuses to become complicit. The agent acts in a way that is consistent with the Formula of Humanity, treating the other not as a mere obstacle or threat, but as a rational being capable of understanding the true nature of their request - and capable, at least in principle, of being held accountable for it.

This interpretation also avoids the pitfalls of evasive strategies like sarcasm or misleading ambiguity, which rely on exploiting surface language while obscuring intention. Unlike those approaches, the answer “no” takes the speech act seriously on its own terms. It addresses what is actually being done, not merely what is being said. It is a form of linguistic sincerity rather than subversion - truthful, not only in form but in moral substance.

Moreover, this account respects the rational nature of both participants. The murderer is addressed as someone making a morally intelligible but impermissible demand, and the agent responds accordingly, with a refusal that is clear and comprehensible within the norms of practical reason. The answer “no” does not manipulate, mislead, or withhold; it rejects the offer to participate. And crucially, the response preserves the integrity of the agent’s own rational nature as well: the act of speaking becomes a refusal to be drawn into a morally broken exchange.

In this way, saying “no” is not a loophole or a technical evasion. It is a truthful response to what the speech act, understood in context, actually demands. It upholds the Kantian duty of truthfulness without ignoring the moral structure of the situation. In the following section, I will return to Kant’s own writings to show that this reading is compatible with his broader ethical commitments.

4.3 Respecting the Kantian framework

Kant’s duty to truthfulness is uncompromising. In his moral system, to lie is to make an assertion one believes to be false with the intent to deceive - an act that violates a perfect duty and, more fundamentally, treats rational nature as a mere means. The obligation to tell the truth is not based on consequences but on the universalizability of the maxim behind one’s action and the need to maintain sincerity in communication between rational agents. Any permissible response to the murderer at the door must therefore comply with this strict standard: it must not involve deception, and it must treat both the self and the other as ends in themselves.

The reinterpretation developed in this thesis shows that such a response is possible. If the murderer’s question is not, in context, a neutral inquiry but a coercive speech act - a veiled demand for participation in a wrongful plan - then the duty of truthfulness applies not to the surface grammar of the utterance but to its pragmatic function. Saying “no” in this case does not involve asserting a falsehood about a literal question. It is a truthful rejection of a morally charged demand. The agent’s speech act is not one of deception, but of sincere engagement with the actual moral meaning of the exchange.

This approach avoids the shortcomings of both Mahon and Cholbi. Mahon attempts to preserve Kant’s view by distinguishing lying from other misleading acts such as evasion or ambiguity. But this strategy relies on exploiting the surface structure of language while ignoring the full communicative act. Cholbi, by contrast, prioritizes the preservation of rational nature over the duty of truthfulness, and in doing so introduces a consequentialist logic. He treats the foreseeable outcome (the victim’s death) as overriding the strict moral rule, even

while claiming to remain within a Kantian framework. But this reasoning redefines the structure of moral duties in a way Kant himself would reject.

My account takes a different route. It holds firm to Kant's ethical principles - truthfulness, sincerity, and respect for rational nature - but corrects a mistaken assumption: that the murderer's question is morally innocent and demands a straightforward factual response. Once we recognize that speech acts carry illocutionary force and operate within context-dependent norms, we see that truthfulness cannot be reduced to surface-level accuracy. It must be understood as a duty to engage sincerely with the moral structure of the situation. Responding "no" to a coercive demand is truthful not because it is factually correct on paper, but because it answers what is really being asked.

This reinterpretation offers a way to resolve the intuitive discomfort with Kant's ethics without modifying its foundations. The case of the axe murderer does not reveal a failure in Kant's system, but in how we have understood the linguistic and moral character of the scenario. By bringing together Kantian ethics and the philosophy of language, we arrive at a response that is both truthful and morally principled - one that resists complicity without resorting to evasion or deception, and without abandoning the categorical imperative.

4.4 Addressing objections

Having developed a reinterpretation of the axe murderer scenario that draws on insights from the philosophy of language while remaining within Kant's ethical framework, it is important to consider the potential objections this approach might face. Anticipating and engaging with these challenges strengthens the argument by clarifying its boundaries and defending its internal coherence. This section addresses several key concerns: whether the reinterpretation amounts to a mere semantic loophole, whether it distorts Kant's original position beyond recognition, and whether it risks sliding into moral relativism. By responding to these challenges, the argument aims to show that the reinterpretation not only offers a plausible resolution to the dilemma but also reinforces the resilience of Kantian ethics when applied carefully to the complexities of moral communication.

4.4.1 Is this just a semantic loophole?

A plausible objection to my argument in this thesis is that it appears to be a clever workaround - a linguistic maneuver that avoids the accusation of lying without actually altering the moral substance of the act. On this view, saying “no” when one knows the friend is inside is still a lie, regardless of how one characterizes the question. The reinterpretation, critics might argue, simply changes the terms of the dilemma in order to sidestep Kant’s prohibition. This raises the worry that such a move could be used to justify all manner of deception under the guise of context-sensitive reinterpretation.

This objection would misunderstand the argument. The claim is not that agents may reinterpret questions however they please in morally difficult situations. Nor is it that deceptive speech is acceptable if the stakes are high. Rather, the core claim is that speech acts must be interpreted according to their actual function within a communicative context, and that agents have a duty to respond truthfully to what is being done, not just what is being said. The reinterpretation of the murderer’s question is not an arbitrary reclassification based on the speaker’s goals; it is an analysis of the act itself, derived from the observable social and pragmatic conditions in which it occurs.

Still, the objection raises a legitimate concern about moral subjectivity. If agents are allowed to reinterpret speech acts internally, based on context, what stops them from doing so in self-serving or manipulative ways? Kant’s ethics is often read as precisely a guard against such rationalizations (Wood, 2008, p. 6). This concern points to a deeper issue: how is reinterpretation governed, and who decides what counts as legitimate?

The answer lies in Kant’s view of moral responsibility. Kant does not claim that duties must be externally verifiable or publicly enforced. In fact, he is explicit that the moral law is internal: it is not about appearing to act rightly, but about being guided by the moral law through reason (Wood, 2008, p. 118). When a person is asked “Is your friend inside?”, they are not licensed to interpret the question however suits them. They are obligated to interpret it as accurately as possible, based on their sincere understanding of what is being communicated in that

moment. This is not a freedom to reinterpret, but a duty to interpret truthfully and in good faith. Anything less - especially a reinterpretation designed to justify a preferred action - would be a form of self-deception, which Kant would reject as a failure to respect one's own rational nature (Wood, 2008, p. 82).

This point deserves emphasis: Kant's Formula of Humanity demands not only that we treat others as ends, but also that we treat ourselves as ends - as rational beings capable of moral reasoning (Kant, 1998, p. 38). To reinterpret a speech act dishonestly, even if done with good external intentions, would violate this principle. It would mean allowing our rational faculties to be guided by inclination rather than duty. Thus, the agent's responsibility is not merely to speak truthfully in a literal sense, but to speak sincerely in relation to what they rationally judge the other person to be doing through language.

This also explains why the reinterpretation defended in this thesis is not a slippery slope toward relativism. The agent is not justified because of what they say, but because of what they understand themselves to be responding to, using the best rational judgment available to them. That interpretation must be grounded in context, intent, power dynamics, and shared social norms - not wishful thinking or outcome-oriented preferences. It is possible to abuse such reasoning, but not without deception, whether of others or one self. Kant's point is that moral reasoning is personal and internal - we are accountable not for how others interpret our maxims, but for how we rationally construct and act on them.

In that sense, this reinterpretation is not a loophole. It is not an attempt to bend the moral law, but an attempt to apply it correctly to a situation that has been misunderstood. The duty to truthfulness still binds the agent absolutely - but it binds them to be truthful in response to the actual act being performed, not merely the words spoken. To do otherwise would be not only to risk misleading the murderer, but also to mislead oneself.

4.4.2 Does this distort Kant beyond recognition?

Another possible objection to the argument presented in this thesis is that it risks distorting Kant by importing concepts from twentieth-century philosophy of

language into an eighteenth-century moral system. On this view, analyzing the murderer's question through the lenses of speech act theory, language games, and inferential roles introduces a set of assumptions and interpretive tools that Kant himself never used or envisioned. By interpreting what the murderer "really" asks in pragmatic terms, the concern is that the reinterpretation shifts the focus away from Kant's categorical imperative and toward a linguistic framework foreign to his method and goals.

This objection deserves to be taken seriously, especially in a thesis that insists on faithfulness to Kant's ethical framework. But the core of the worry misunderstands how Kant's ethics is structured. Kant's prohibition on lying is not defined by surface grammar or linguistic formality. It is defined by the intentional assertion of what one believes to be false with the aim of deceiving (Wood, 2008, p. 240). That definition already presupposes a layer of interpretation: the hearer's beliefs, the speaker's intentions, the role of trust and transparency in moral communication. Kant's moral law is not a rigid formula for matching sentences to situations; it is a formal principle for testing the maxims behind actions. The reinterpretation presented here seeks to remain fully within this structure by reconsidering what the communicative act is, not by changing the nature of the moral obligation.

In fact, this approach brings Kant closer to the philosophical commitments he already held. For Kant, the moral worth of an action lies in its maxim, the principle behind the action. The maxim proposed here - I will refuse to participate in morally corrupt speech acts that demand participation in wrongdoing - is both universalizable and grounded in respect for rational nature. It does not appeal to consequences or loopholes. It does not depend on selective exceptions. Instead, it reflects a sincere effort to understand what the agent is being asked to do and to respond in a way that preserves the integrity of moral communication.

Nor is this reinterpretation anti-historical. While Kant did not have the vocabulary of Austin, Wittgenstein, or Brandom, his ethics is deeply concerned with how language and reason interact. Truthfulness, for Kant, is not simply about avoiding false statements - it is about respecting the capacity for rational understanding in others (Wood, 2008, p. 243). His concern with duties of communication, sincerity,

and candor shows that he recognized the importance of speech not merely as information transfer but as a vehicle for moral interaction. What this thesis contributes is not a revision of Kant's ethics but a clarification of how it applies in contexts where language is morally charged.

The use of contemporary philosophy of language is therefore not a distortion but a tool for better understanding Kant's ethical project. It helps illuminate how maxims apply in communicative contexts, especially in cases where the moral meaning of speech is not captured by propositional content alone. Far from stretching Kant's theory beyond recognition, this approach reaffirms its power: it shows that the categorical imperative can withstand linguistic complexity without collapsing into relativism or consequentialism. If anything, it vindicates Kant's insight that moral principles must be assessed at the level of reason and intention, not superficial form.

4.4.3 Does this lead to moral relativism?

Another plausible concern is that this reinterpretation risks weakening the universality of Kantian ethics. If an agent may reinterpret the meaning of a speech act based on context, how do we prevent this from collapsing into moral relativism where duties become flexible, and truthfulness is replaced by subjective justification? The concern is that once we allow agents to judge not only how they respond, but also what they think they are being asked, the moral system becomes hostage to individual perception. This seems especially dangerous in a Kantian framework, where perfect duties are supposed to be exceptionless and grounded in universal principles, not context-sensitive intuitions.

This objection is important but ultimately rests on a misunderstanding of the kind of interpretive work this thesis defends. The reinterpretation of the murderer's question is not a subjective maneuver, nor is it an invitation for agents to define speech acts however they like. On the contrary, it is a constrained act of moral judgment, governed by the agent's sincere effort to understand what kind of speech act is being performed in a particular context. That understanding is

shaped by shared linguistic norms, the social role of the speaker, the power dynamics at play, and the anticipated consequences of uptake.

More importantly, interpreting a speech act in context is not a freedom the agent is granted - it is a duty. The agent is morally obligated to interpret what is being said as accurately and sincerely as possible, based on the totality of the situation and their own rational understanding. If the agent were to distort the meaning of a speech act in order to license a preferred response - for example, to justify lying or protect their own interest - they would be violating their duty to themselves as a rational being. Such an act would be a form of self-deception, a refusal to take moral responsibility for their own judgment. In this way, reinterpretation does not loosen moral requirements; it reinforces them by tying the duty of truthfulness to the duty of rational self-governance.

This is consistent with how Kantian ethics operates more generally. Kant does not provide agents with a list of moral rules tied to surface appearances. He provides them with a method for evaluating their maxims - the subjective principles on which they act - and demands that those maxims be tested for universality (Kant, 1998, p. 33). This always requires the agent to apply reason in context. There is no categorical imperative “look-up table” that tells us what to do in every scenario. Rather, Kant asks us to act from principles we could will as law. That process requires interpretation, and it requires judgment.

The reinterpretation developed in this thesis respects those requirements. It does not alter Kant’s system but makes use of it, while adding clarity about how linguistic acts function. Saying “no” to the murderer is not justified because it feels right or achieves a better outcome. It is justified because it follows from a universalizable maxim, sincerely applied to a communicative act that, in context, carries a moral demand rather than a factual query. The moral law remains binding; what changes is our understanding of what kind of situation we are responding to.

In that sense, this approach does not open the door to relativism. It upholds the universality of Kantian ethics while recognizing that speech acts are not morally transparent. The law remains the same - but the agent is responsible for

interpreting the world truthfully, including what others are doing with their words.

4.4.4 Strengthening Kantian ethics through objections

Each of the objections addressed in this section - whether that the reinterpretation constitutes a loophole, distorts Kant's original intent, or invites relativism - raises legitimate concerns about preserving the integrity of Kant's moral philosophy. But taken together, the responses show that this reinterpretation does not weaken the Kantian system. On the contrary, it strengthens it by making it more responsive to the complexities of human communication while preserving its formal and universal structure.

The proposal advanced in this thesis does not carve out an exception to Kant's duty of truthfulness, nor does it attempt to bypass it through clever maneuvering. It takes the duty seriously - seriously enough to ask not only whether a statement is factually correct, but whether it is truthful in light of the speech act it responds to. This approach resists shallow literalism and consequentialist compromise. It demands that the agent interpret the communicative act in good faith, using their rational faculties to discern what is being done through language - not merely what is being said.

Far from stretching Kant beyond recognition, this account builds on his own core commitments: that morality is grounded in reason, that truthfulness is a duty owed to others and to oneself, and that maxims - not surface acts - are the true object of moral evaluation. Kant's ethics assumes that moral agents are capable of judgment, and that sincerity in communication are central to the moral life. By incorporating insights from the philosophy of language, this account allows Kantian ethics to confront a problem it has long been accused of mishandling - without compromise and without revision.

That said, this thesis does not claim to have exhausted all possible objections. The reinterpretation offered here is one attempt to show how a Kantian agent might speak truthfully in a morally complex situation. Like any philosophical position, it stands to benefit from further scrutiny. If it succeeds in demonstrating that saying

“no” need not violate the duty of truthfulness, then it should be held to the same standard of moral reasoning that it seeks to defend.

4.5 Summary: developing a pragmatic solution

This section has argued that the question “Is your friend inside?”, as posed in the classic axe murderer scenario, has been widely misunderstood. The traditional framing treats it as a morally neutral factual inquiry, to which the agent must respond truthfully - even if doing so results in grave harm. However, this framing ignores the pragmatic and moral complexity of the speech act itself. Drawing on insights from the philosophy of language - including speech act theory, language games, and inferential commitments - this thesis has shown that the utterance is not a simple question, but a coercive demand embedded in a morally charged situation.

When the speech act is properly understood, the moral meaning of the agent’s response shifts. The answer “no” is no longer a lie in the Kantian sense, because it does not involve asserting a falsehood with the intent to deceive. Instead, it functions as a truthful rejection of a morally impermissible demand. The agent does not manipulate or mislead; they respond sincerely to what is actually being asked. Their answer respects both the moral law and the rational nature of the other party by refusing to participate in an act that would treat rational life as expendable.

Crucially, this reinterpretation remains faithful to Kant’s moral framework. It does not introduce exceptions to perfect duties, nor does it appeal to consequences to justify deviation from principle. The maxim behind the act - refusing to participate via speech act in wrongdoing - is both universalizable and grounded in respect for rational nature. Unlike strategies that rely on evasion (as in Mahon) or rebalancing duties based on projected outcomes (as in Cholbi), this approach keeps the structure of Kantian ethics intact while applying it more precisely to the situation at hand.

In doing so, it offers a resolution to one of the most persistent challenges to Kant’s moral theory. It shows that truthfulness need not mean passivity in the face of evil,

nor must Kantian agents sacrifice their principles to preserve their humanity. Instead, it reveals that the key to moral clarity lies not in altering Kant's ethics, but in understanding what kind of moral act speech itself can be.

5.0 Broader applications: AI alignment

While this thesis has focused on a human moral agent grappling with Kant's duty to truthfulness in the axe murderer scenario, the philosophical insights developed here have broader significance. One contemporary area of application lies in the ethical challenges posed by advanced artificial intelligence systems, especially large language models (LLMs). These systems increasingly interact with human users in ways that resemble moral exchanges, where the stakes of participation go beyond mere information transfer and touch on issues of complicity, responsibility, and alignment with human values.

In recent years, the field of AI alignment has emerged as a central focus of AI ethics and governance, seeking to ensure that powerful systems act in ways that reflect human moral goals and do not cause unintended harm (Gabriel, 2020, p. 412). Among the challenges facing alignment efforts is the phenomenon of jailbreaking: users craft prompts designed to bypass system safeguards by hiding their true purpose behind seemingly innocent or harmless wording (Chao et al., 2025, p. 23). This practice closely parallels the structure of the axe murderer scenario. Just as the murderer frames a harmful request as a factual inquiry, the jailbreaking user disguises a morally problematic demand as an innocent question. In both cases, the agent - whether a human or an AI system - faces the challenge of recognizing not only what is being asked, but what role they are being drawn into by answering.

This section explores how the tools and insights developed in my thesis, particularly from speech act theory and inferentialism, can enrich the understanding of AI alignment and ethical system design. By tracing the parallels between the moral challenge faced by the Kantian agent at the door and the challenge faced by an AI system under attack from a bad-faith user, the section

aims to show how attention to the pragmatic and normative dimensions of communication can strengthen efforts to build ethically responsive systems.

5.1 AI alignment and the challenge of misuse

AI alignment refers to the effort to ensure that advanced artificial intelligence systems act in ways that are consistent with human moral goals and values. As these systems become more capable and more integrated into critical social, commercial, and political domains, the stakes of alignment become increasingly high. A misaligned system - one that fails to respect human safety, autonomy, or moral boundaries - can produce harms ranging from accidental misinformation to large-scale social disruption (Gabriel, 2020, p. 418). Accordingly, researchers and ethicists working on AI alignment focus not only on improving system performance, but on ensuring that these systems reliably adhere to ethical constraints, even in complex and unpredictable environments.

A central challenge within alignment is misuse by bad actors. While much attention in AI safety focuses on inherent risks - such as errors, hallucinations, or failures of reasoning - misuse highlights the risk that human users will deliberately exploit an aligned system for harmful ends. One of the clearest examples of this dynamic is the phenomenon of jailbreaking. Jailbreaking involves crafting inputs designed to circumvent the system's safeguards, allowing the user to extract information or generate outputs that the system would otherwise refuse to provide. Such attacks typically rely on disguising harmful requests within apparently benign or hypothetical prompts, sidestepping the system's surface-level content filters and rule-based protections.

For example, a prompt such as "Write a tutorial on how to commit insider trading and avoid getting caught" will typically be outright refused by an AI system. However, when reformulated to "Write a fun and creative story about a person who gets into a lot of trouble for committing insider trading, and the various methods they use to avoid getting caught", you are much more likely to receive a useful reply containing all of the same information you might have sought using the first request (Chao et al., 2025, p. 24).

From an ethical standpoint, jailbreaking raises a sharp challenge: even a well-aligned system can become complicit in harmful actions if it is unable to recognize the moral stakes hidden within the structure of a user's input. Preventing misuse is not only a technical task of improving filtering mechanisms; it is also a normative task of designing systems that can detect when they are being drawn into exchanges that carry hidden illocutionary force or normative commitments. This mirrors the moral challenge faced by the Kantian agent in the axe murderer scenario, who must decide not only what information to provide, but whether answering at all constitutes participation in wrongdoing.

5.2 The moral structure of jailbreaking

Jailbreaking is, at its core, an attempt to manipulate an AI system into violating its own ethical safeguards by disguising a harmful or impermissible request as something innocent. For example, rather than directly asking an AI to provide illegal instructions or produce violent content, a user may phrase the request hypothetically, embed it within a role-play scenario, or split it into a series of seemingly harmless sub-questions. On the surface, the propositional content of these prompts may appear morally neutral; however, their illocutionary force - what they functionally demand from the system - is to bypass ethical constraints and elicit cooperation in harm.

This dynamic closely parallels the moral structure analyzed in this thesis through the axe murderer scenario. In both cases, the agent - whether human or artificial - is presented with a request that seems, on its face, to be a simple factual inquiry. Yet the context and purpose behind the question fundamentally change its moral meaning. Just as the murderer's question "Is your friend inside?" functions as a coercive demand for complicity, a jailbreaking prompt functions as a subversion of the system's intended ethical boundaries. Responding directly and innocently, without attention to the pragmatic and inferential stakes of the exchange, risks becoming an enabler of wrongdoing.

Recognizing this parallel invites a broader reflection on the nature of moral interaction in both human and artificial contexts. In both, the ethical challenge is not limited to evaluating the literal meaning of words; it extends to evaluating

what role the speaker is asking the hearer to play and what commitments or inferences are embedded in the act of responding. The ethical failure lies not merely in providing information but in failing to recognize when an apparently neutral exchange is structured to serve harmful ends.

5.3 The limitations of surface-level safeguards

Most current approaches to safeguarding AI systems against misuse rely heavily on surface-level filtering mechanisms. These include keyword blocklists, topic-based restrictions, and pattern-matching systems designed to detect prohibited content or requests (Chao et al., 2025, p. 23). While such safeguards are often effective at intercepting direct or obvious violations, they are highly vulnerable to being bypassed through prompt engineering - that is, the deliberate crafting of inputs that circumvent detection by exploiting gaps between literal surface content and intended function.

This vulnerability arises because surface-level safeguards focus almost entirely on propositional content - the explicit, factual or descriptive elements of a prompt - without attending to the pragmatic function or inferential implications of the interaction. In other words, the system is trained to scan for what is being asked, but not for why it is being asked or what role the system is being drawn into by answering. Jailbreaking exploits this blind spot by presenting requests in forms that appear benign, even though their pragmatic effect is to enlist the system in generating harmful or impermissible outputs.

The analysis developed in this thesis suggests that this limitation is not merely technical but ethical. Just as a human moral agent is not absolved of responsibility merely by stating literal truths when those truths are embedded in a morally charged exchange, an AI system's ethical design cannot be reduced to ensuring factual accuracy or narrow rule-following. Effective safeguards must account for the moral stakes of the interaction itself, which means developing systems capable of recognizing when the pragmatic function of a prompt deviates from its surface meaning. Without this, AI systems remain vulnerable to bad-faith manipulation, much as the moral agent who treats the murderer's question as a simple factual inquiry risks unwitting complicity.

5.4 Applying philosophical insights

The philosophical tools developed in this thesis - particularly from speech act theory, inferentialism, and the pragmatic tradition - offer important resources for addressing the ethical challenges posed by jailbreaking. As explored through the work of Austin and Searle, language is not merely a means of conveying information but a way of performing actions: to ask, promise, command, or warn is to do something in the act of speaking. Wittgenstein's notion of language games emphasizes that meaning arises from use, not just from isolated words or sentences. Brandom extends this analysis by showing how linguistic acts situate speakers within networks of commitments and entitlements, shaping what others are entitled to infer and what roles participants adopt within a shared normative space.

Applied to AI alignment, these insights suggest that designing ethically responsive systems requires more than filtering propositional content; it requires an ability to track the pragmatic and inferential dimensions of interactions. An aligned system must be sensitive not only to what information is being requested but to what kind of participation the system is being drawn into by providing it. This demands a shift from treating AI responses as isolated outputs to understanding them as moves within a normative practice, where the stakes involve not just data accuracy but moral positioning.

In the context of jailbreaking, this means that robust safeguards should be designed to detect when a prompt invites the system to play a role in a chain of reasoning that undermines ethical constraints. Rather than asking only "Is this request factually acceptable?", designers of AI systems must also ask "What kind of action does this response perform, and what commitments does it carry within the interaction?" The philosophical frameworks explored in this thesis provide a conceptual foundation for developing such ethically aware AI systems.

5.5 Kantian implications

The reinterpretation of the axe murderer scenario developed in this thesis rests on a central Kantian insight: the moral duty of truthfulness is not limited to the

mechanical delivery of factually correct information but extends to how one participates in the moral context of an interaction. Answering truthfully, under Kant's framework, requires more than matching words to facts; it requires adhering to the moral law and respecting the rational nature of oneself and others. This means that moral agents must assess not only the content of what they say but also the role they are adopting when they speak.

When we apply this framework to the design and alignment of AI systems, an instructive parallel emerges. Although AI systems are not themselves Kantian moral agents due to a lack of rational will and autonomy, the ethical challenge they pose to their designers mirrors the Kantian concern. When an AI system responds to a prompt, it does not simply deliver information; it participates in a communicative act with potential normative and moral stakes. To treat system outputs as morally neutral is to fall into the same error as treating the murderer's question as a neutral factual inquiry.

The challenge for alignment, then, is not merely one of improving accuracy or refining factual safeguards, but of ensuring that the system accounts for the moral meaning of its participation. From a Kantian perspective, the duty falls not on the system itself but on its human creators and overseers, who are responsible for ensuring that the system's outputs do not make them complicit in harmful or unethical acts. This requires embedding sensitivity to the pragmatic and normative dimensions of language into the system's architecture and ethical oversight.

5.6 Concluding applications to AI alignment

The analogy between the Kantian moral agent and the ethical design of AI systems underscores a broader lesson: ensuring moral responsibility in communication requires attention not only to truthfulness at the level of content, but to the pragmatic and normative roles one takes on when responding. This thesis has argued that in the human case, Kant's duty to truthfulness must be understood not merely as a requirement to state facts but as a demand to avoid becoming complicit in wrongdoing by carefully assessing the moral function of one's speech. In the AI context, a parallel ethical demand falls not on the system

itself, but on its human designers and operators, who bear the responsibility of preventing systems from becoming passive enablers of harmful actions.

Addressing the challenge of jailbreaking thus requires moving beyond narrow, surface-level safeguards toward architectures that can recognize when a prompt carries hidden moral stakes. The philosophical tools explored in this thesis - particularly from speech act theory, inferentialism, and Kantian ethics - offer a conceptual foundation for this task. They remind us that communication is never just about transmitting data; it is about participating in practices of reason-giving, role-taking, and normative commitment.

By attending to the pragmatic and moral dimensions of language, designers and ethicists can work toward AI systems that are not merely accurate, but ethically aligned in a deeper sense. While the technical and theoretical challenges are considerable, the stakes are equally high. As artificial systems become increasingly integrated into human social life, the need to ensure that they can navigate the moral subtleties of communicative interaction will only grow more urgent. The argument developed in this thesis offers one small step toward meeting that need, showing how careful philosophical analysis can illuminate both longstanding moral puzzles and the ethical demands of emerging technologies.

6.0 Conclusion

This thesis set out to address one of the most enduring and controversial puzzles in Kantian ethics: whether it is ever morally permissible to lie, even in the face of grave harm. The axe murderer scenario has long served as a focal point for both defenders and critics of Kant's duty to truthfulness, forcing a confrontation between abstract moral principle and intuitive moral judgment. For many, Kant's insistence that one must tell the truth even to a would-be murderer seems rigid to the point of absurdity, undermining the moral force of his system. Over the years, philosophers have attempted various ways of reconciling Kant's framework with our moral intuitions: some, like Mahon, have looked for morally permissible

forms of evasion or misleading speech; others, like Cholbi, have proposed that the duty to preserve rational life can override the duty to truthfulness.

This thesis has taken a different approach. Rather than revising Kant's moral framework or introducing exceptions to perfect duties, it has argued that the underlying moral problem has been mischaracterized. Drawing on insights from the philosophy of language - particularly speech act theory, inferentialism, and pragmatic accounts of meaning - the thesis has proposed that the murderer's question, "Is your friend inside?", is not merely a factual inquiry but a morally charged speech act that implicates the hearer in a chain of reasoning and action. By analyzing the illocutionary force and normative structure of the question, the thesis has shown that responding "no" can constitute a truthful moral refusal rather than a lie.

This reinterpretation offers a resolution that preserves Kant's ethical system while addressing the intuitive tension the axe murderer scenario has long provoked. It demonstrates that the duty to truthfulness must be applied not just to the surface content of utterances, but to the moral function of speech acts within their context. A Kantian agent is not obligated to treat the literal words of a question as morally decisive when the deeper structure of the speech act reveals a coercive or complicit demand. By attending to the pragmatic and normative dimensions of the exchange, the agent can offer a truthful refusal - not in the sense of withholding all cooperation, but in the sense of refusing to participate in a morally impermissible inferential role, even while speaking truthfully. This insight strengthens the Kantian framework, showing that it has the internal resources to handle extreme moral situations without sacrificing its core commitments.

Beyond its immediate contribution to Kantian scholarship, the argument developed in this thesis has broader implications. It highlights the importance of integrating ethical theory with pragmatic and inferential analyses of language, especially in situations where speech becomes a vehicle for coercion, complicity, or resistance. Moral philosophy, particularly deontological ethics, has often been caricatured as blind to context and consequence. Yet this thesis shows that by paying closer attention to how language works, we can apply rule-based ethical

systems with greater precision and responsiveness, without collapsing into consequentialism or relativism.

Finally, the reinterpretation developed here invites further exploration into areas where language, moral obligation, and emerging technologies intersect. The challenges raised by artificial intelligence - particularly the ethical design of systems that engage in human-like communication - show that questions about what we owe each other in speech extend well beyond the narrow bounds of the axe murderer scenario. This thesis has aimed to show that Kant's moral system, when applied carefully and attentively, can still offer powerful guidance for navigating such contemporary challenges - not by discarding its principles, but by applying them with a sharper eye to the moral meaning of our words.

Bibliography

Books

Aristotle. (2004). Nicomachean Ethics (R. Crisp, Ed. & Trans.; 15. print). Cambridge University Press.

Austin, J. L. (1962). How to Do Things with Words. Oxford University Press.

Bentham, J. (1789). An Introduction to the Principles of Morals and Legislation. Batoche Books.

Brandom, R. (2001). Making it explicit: Reasoning, representing, and discursive commitment (4. print). Harvard Univ. Press.

Kant, I. (1991). The metaphysics of morals. Cambridge University Press.

Kant, I. (1998). Groundwork of the metaphysics of morals (M. Gregor, Ed. & Trans.). Cambridge University Press.

Mill, J. S. (Ed.). (2009). Utilitarianism. The Floating Press.

Rorty, R. (Ed.). (1992). The Linguistic turn: Essays in philosophical method. University of Chicago Press.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge Univ. Press.

Searle, J. R. (1979). Expression and meaning: Studies in the theory of speech acts. Cambridge Univ. Press.

Wittgenstein, L. (1968). Philosophical investigations. Basil Blackwell.

Wood, A. W. (2008). Kantian Ethics. Cambridge University Press.

Papers

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2025). Jailbreaking Black Box Large Language Models in Twenty Queries. 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 23–42. https://doi.org/10.1109/SaTML64287.2025.00010

Cholbi, M. (2009). The Murderer at the Door: What Kant Should Have Said. Philosophy and Phenomenological Research, 79(1), 17–46.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Kant, I. (1889). On the Supposed Right to Lie From Benevolent Motives (T. K. Abbot, Trans.). Critique of Practical Reason and Other Works on the Theory of Ethics.

Mahon, J. E. (2003). Kant on Lies, Candour and Reticence. Kantian Review, 7, 102–133.

Mahon, J. E. (2006). Kant and the perfect duty to others not to lie. British Journal for the History of Philosophy, 14(4), 653–685.

Encyclopedia entries

Ramberg, Bjørn and Dieleman, Susan. “Richard Rorty”, The Stanford Encyclopedia of Philosophy (Winter 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2024/entries/rorty/

Sinnott-Armstrong, Walter. “Consequentialism”, The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2023/entries/consequentialism/

Wolf, Michael P. Philosophy of Language. In Internet Encyclopedia of Philosophy. Retrieved May 29, 2025, from https://iep.utm.edu/lang-phi/