On moral responsibility
A metaphysical examination

Alexander Towell
lex@metafunctor.com

Abstract

This paper examines three fundamental metaphysical challenges to moral responsibility: the ontological status of moral properties, the persistence of persons across time, and the compatibility of moral agency with causal determinism. I argue that while these challenges reveal genuine philosophical puzzles about the deep structure of reality, they do not establish that moral responsibility is impossible. Rather than demanding that ethics be grounded in contested metaphysical foundations, I propose that we embrace moral agency as a practice grounded in phenomenology—in what we directly experience. The key insight is that modernity has inverted the proper epistemic order: we directly experience qualities (the badness of suffering, the goodness of flourishing, the phenomenology of choice), while quantitative models and metaphysical theories are tools we construct to organize these experiences. By reversing this inversion and recognizing the epistemic priority of experience over theory, we can ground ethics phenomenologically while acknowledging profound metaphysical uncertainty. This approach extends moral consideration to all sentient beings capable of valenced experience and provides a foundation for moral responsibility that is more certain than any disputed metaphysical theory—because it is grounded in what is most immediately given: the mountain of experience rather than the map of theory.

Contents

1 Introduction

People throughout history have believed that they belong to a special categorical class called *persons*. What supposedly makes members of this class special is their status as *moral agents*–beings capable of making choices, forming intentions, and bearing moral responsibility for their actions. This belief in moral agency is not merely theoretical; it is deeply woven into the fabric of human social life, informing our practices of praise and blame, our legal systems, and our most intimate relationships.

Yet this seemingly fundamental aspect of human existence faces profound metaphysical challenges. If the nature of reality is at odds with the criteria for moral agency, our self-understanding as moral beings could be reduced to a pleasant fiction. This paper examines three fundamental metaphysical challenges to moral responsibility: challenges concerning the reality of moral properties themselves, challenges concerning the persistence and nature of persons across time, and challenges concerning the freedom required for genuine choice.

The central argument However, this paper does not conclude that these challenges are insurmountable. Rather, I argue that we should embrace moral agency as a practice grounded in phenomenology–in what we directly experience–despite unresolved and possibly unresolvable metaphysical questions. The metaphysical challenges I examine reveal genuine uncertainties about the deep structure of reality, but they do not establish that moral responsibility is impossible or illusory. Instead, they demonstrate that moral agency may rest on foundations different from those assumed by naive moral realism.

The key insight is this: modernity has inverted the proper epistemic order. We directly experience *qualities*–the badness of suffering, the goodness of flourishing, the phenomenology of choice and deliberation. Quantitative models and metaphysical theories are tools we construct to predict and organize these qualitative experiences. Yet we have come to treat our theoretical models as more fundamental than the experiences they were designed to explain. We stand on a mountain looking at a map of the mountain, claiming the mountain arose from the map.

This inversion has profound consequences for ethics. When we demand that moral agency be grounded in robust metaphysical facts–objective Universal properties, persisting substantial selves, libertarian free will–we are demanding that direct experience be validated by theoretical constructs. But the validation should run in the opposite direction: our theories should answer to our experiences, not our experiences to our theories.

Metaphysics and its limits Metaphysics is a discipline in philosophy with a very long history. The etymology of *metaphysics* derives from the Greek words *metá* (after) and *physiká* (physics), referring to Aristotle's work that followed

his chapter on physics. Over time, *metá* has been reinterpreted to mean *beyond*, so that metaphysics becomes the study of that which is beyond physics–the categorical structure of reality itself.

Yet there is deep disagreement among metaphysicians about this categorical structure. Do categories represent absolute ontological differences, or do they represent useful ways in which humans organize their experiences? Does reality have an inherent categorical structure waiting to be discovered, or do minds invent conventional categories to order their experiences? These are not merely academic questions; they bear directly on whether moral categories like "person," "agent," and "responsibility" pick out objective features of reality or conventional classifications.

I will argue that our cognitive limitations may prevent us from definitively answering such questions. The tension between qualitative experience and quantitative description, between the map and the territory, may be an artifact of the bandwidth constraints inherent in human cognition. We may lack the cognitive capacity to fully resolve whether reality is fundamentally mathematical, fundamentally experiential, or something our categories cannot adequately capture.

Moral agency To say that we are moral agents is to say that we can be held morally responsible for our actions, which can range from exemplary to reprehensible. People who rescue villagers from a volcanic eruption demonstrate moral agency; so too do those who choose not to help. The volcano itself is causally responsible for destruction but not morally responsible, for it lacks the capacities that ground moral agency.

Under what circumstances can conduct be ascribed moral responsibility? A theory of moral responsibility must address:

1. The subject of morality itself–what makes actions right or wrong?
2. The criteria for being a moral agent–what capacities are required?
3. The circumstances under which moral responsibility can be ascribed–when are agents responsible for their actions?

The discipline of metaphysics has raised profound challenges to each of these elements. But rather than concluding that moral responsibility is impossible, I will argue that these challenges reveal the limitations of certain metaphysical frameworks, not the impossibility of moral agency itself. The path forward requires acknowledging genuine metaphysical uncertainty while grounding ethics in the phenomenology of experience–in what we actually undergo when we suffer, flourish, deliberate, and choose.

Roadmap This paper proceeds as follows. Section 2 examines the ontological status of moral properties, contrasting realism (which treats moral properties as objective Universals) with nominalism (which treats them as conventional classifications), with special attention to Bradley's regress and its implications for moral realism. Section 3 analyzes the criteria for moral agency, including rationality, intentionality, and the knowledge condition. Section 4 investigates

personal identity and persistence, examining Bundle Theory versus Ego Theory and their implications for moral responsibility across time. Section 5 develops the key insight about the map-territory inversion, arguing that modernity has reversed the proper epistemic order by treating quantitative theories as more fundamental than the qualitative experiences they were designed to explain. Section 6 proposes phenomenological grounding for ethics, arguing that the direct experience of suffering and flourishing provides sufficient foundation for moral practice independent of contested metaphysical theories. Section 7 examines challenges from determinism, including four-dimensionalism, causal determinism, and the compatibilism-incompatibilism debate. Finally, Section 8 synthesizes these arguments, concluding that we can maintain moral agency while embracing metaphysical uncertainty—living at the limits of understanding while remaining committed to the reduction of suffering.

2 The reality of moral properties

> *"Morality is simply the attitude we adopt towards people whom we personally dislike."* –Oscar Wilde

The etymology of *morality* derives from the Latin word *moralitas* (manner, character, proper behavior), and it can be used a) descriptively to refer to a code of conduct promoted by some group, or b) normatively to refer to a code of conduct that, given specified conditions, would be promoted by any moral agent. In what follows, our primary concern will be of the second normative type where the nature of morality can be explored in greater isolation from cultural norms.

What is the basis for morality? From a behavioral point of view, morality is arguably the study of right and wrong conduct. However, what is the basis of right and wrong? One possibility is that its basis is authoritative prescription, in which case morality is a dictum of *might makes right*–something is morally good or bad because an authority says so with no further explanation required. A forceful criticism of this type of moral theory is revealed by asking: is an action good because God commands it, or does God command it because it is good? If it is good simply because God commands it, then morality is nothing more than the arbitrary preference of God. But if God commands it because it is good, then morality has a basis independent of God.

Many would contend the basis is essentially rational; persons use reason and experience to inform their conduct, e.g., how can a group best realize common interests? In which case, a sense of right and wrong is, for the most part, the habituation of conduct that promotes cultural dominance or health, i.e., the principle of natural selection applied in the context of a social Darwinism.

While there are perhaps many other moral theories–some of which may not support the premise of moral responsibility–each must fall within one of two general categories: the conventional, and the intrinsic. On the one hand, if morality is conventional, then persons are only morally responsible by convention. On the other hand, if morality is categorically intrinsic, then arguably moral

agents have an inherent duty to conform to it. But what does it mean to say that morality is categorically intrinsic?

Realism

There are many things–particulars–which are red. For instance, there are red cars, red apples, and red signs. But *what* makes them red? The philosophical discipline of realism submits that all red particulars are red in virtue of the existence of a Universal–a single abstract entity that is a part of all red particulars. More precisely, Universals are repeatable entities that can be exemplified simultaneously by different particulars, where entities are things like *properties*, like *red*, *relations*, like *behind*, and *kinds*, like *dog*.

Realists argue that attribute agreement is due to relationships between entities and Universals. Furthermore, they contend that the subject-predicate grammar of sentences reinforces this view, e.g., in the sentence "My apple is red," the predicate "is red" seems to be picking out something (a noun)–call it *redness*–and linking it to the subject, "my apple." Let us examine how a realist might account for following two sentences:

1. My apple is red.
2. My friend's apple is red.

In each sentence, there are arguably two Universals being exemplified. First, there is the Universal property, *redness*, that is being exemplified by the two particulars, *my apple* and *my friend's apple*. Second, there is the Universal kind, *apple-ness*, that is also being exemplified by *my apple* and *my friend's apple*. The *redness* being exemplified by the entity, *my apple*, is numerically identical with the *redness* being exemplified by the entity, *my friend's apple*. Likewise, the *apple-ness* that is being exemplified by *my apple* is numerically identical with the *appleness* that is being exemplified by *my friend's apple*.

In the sentence, "This apple is red," a metaphysical realist maintains that this denotes "*This apple* exemplifies *redness*," where *redness* is a Universal that inheres in *this apple*. More generally, a realist claims that any subject-predicate sentence of the form "a is F" can be paraphrased as "a exemplifies F-ness" where a is an entity and *F-ness* is a Universal. In the same way, a realist could maintain that *morality*, like *color*, is a Universal. In the same way that an apple can exemplify a color like *redness*, a person can exemplify a *virtue* (a moral excellence) like *honesty*, e.g., "Bob exemplifies honesty." However, it turns out that the existence of Universals is one of the most important disputes in the discipline of metaphysics.

Paradox If "a is F" denotes "a exemplifies F-ness," then we can construct a paradox. Let F mean "does not exemplify itself" and so *F-ness* could be the Universal property, *non-selfexemplification*. Thus, for some a, the proposition "a exemplifies *non-selfexemplification*," should be either **true** or **false**. But the existence of this property leads immediately to a paradox, for if a thing does

exemplify itself, then it does not exemplify itself and if a thing does not exemplify itself, then it does exemplify itself. It is self-contradictory.

On the one hand suppose that an entity, a, does exemplify itself. Then, since it is a property an entity exemplifies when it does not exemplify itself, a does not exemplify itself. On the other hand, suppose that a does not exemplify itself. Then, it does exemplify itself. It is similar to the sentence, "This sentence is a lie." If it is **true**, then it is **false**; and if it is **false**, then it is **true**. It is self-contradictory.

To avoid this paradox, restrictions on what constitutes a Universal must be imposed, such as denying that *non-selfexemplification* is an actual Universal. However, as a result of this, the realist cannot take it for granted that "a is F" means that F must be selecting a Universal, *F-ness*. This is not necessarily a problem for morality, but the realist cannot simply claim, without qualification, that if a sentence can take on a subject-predicate form that necessarily means that the existence of Universals can be assumed to account for things like attribute-agreement. For instance, "*Socrates* is *honest*" does not necessarily mean that the adjective, *honest*, is pick out out a Universal entity, *honesty*.

Bradley's regress: The central challenge to realism A more fundamental challenge to realism emerges when we examine the nature of the exemplification relation itself. This problem, first articulated by F.H. Bradley in his 1893 work *Appearance and Reality* [2], strikes at the heart of how Universals are supposed to relate to particulars. Bradley argued that if relations are themselves Universals, then an infinite regress ensues that may undermine the entire realist project.

Consider the realist's analysis of "Socrates is courageous" as "Socrates exemplifies courage," where courage is a Universal property and Socrates is a particular. This analysis invokes three entities: Socrates, the Universal courage, and the exemplification relation that connects them. But if exemplification is itself a Universal relation, then we must explain how this relation relates to its relata–how exemplification itself connects to both Socrates and courage.

It appears we need a further relation, exemplification$_2$, to relate exemplification$_1$ to Socrates and courage. But then exemplification$_2$ requires exemplification$_3$ to relate it to its relata, and so on ad infinitum. For any subject-predicate sentence "a is F," the realist analysis generates an infinite hierarchy:

1. a exemplifies$_1$ F-ness
2. The complex "a exemplifies$_1$ F-ness" exemplifies$_2$ exemplification
3. The complex "[a exemplifies$_1$ F-ness] exemplifies$_2$ exemplification" exemplifies$_3$ exemplification
4. And so on, indefinitely

The critical question is whether this infinite regress is vicious–that is, whether it undermines the realist's explanatory project. An infinite regress is generally considered vicious when each step in the series depends on a subsequent step for its explanation or grounding, such that the explanation can never be completed.

Realist responses to Bradley's regress Contemporary realists have developed several sophisticated strategies to address this challenge, and the success or failure of these responses is central to assessing realism about moral properties.

The non-relational tie response. The most influential realist response, defended prominently by Armstrong [1], denies that exemplification is a genuine Universal. Instead, exemplification is a "non-relational tie" or "fundamental nexus"–a basic feature of reality that requires no further analysis. On this view, that particulars instantiate Universals is a brute metaphysical fact. The basic furniture of reality includes not just particulars and Universals, but also the instantiation relation that holds between them. Just as we must accept some primitives in any explanatory framework (e.g., physical laws in science, logical connectives in logic), the realist claims we should accept instantiation as primitive.

Critics object that this response abandons the realist's explanatory ambitions. If instantiation is primitive and inexplicable, then the realist has not really explained attribute agreement–they have merely labeled it. The nominalist might retort: if you are willing to accept primitive facts about instantiation, why not accept primitive facts about resemblance instead, thereby avoiding commitment to Universals altogether?

The benign regress response. An alternative strategy, explored by contemporary metaphysicians like Maurin [13], accepts the infinite regress but denies its viciousness. The argument proceeds as follows: once we establish the ground-level fact that a exemplifies F-ness, all higher-order facts in the infinite series follow automatically and require no additional metaphysical grounding. The truth of "$\{a$ exemplifies$_1$ F-ness$\}$ exemplifies$_2$ exemplification" is not an independent fact requiring explanation; it is simply a consequence of the first-order fact plus the general principle that when a particular instantiates a Universal, that instantiation itself instantiates the Universal of exemplification.

This response faces its own difficulties. It must explain why the regress, though infinite, is not vicious. One criterion often proposed is that a regress is benign if each level is explanatorily posterior to the previous level–that is, the higher-order facts are explained by the lower-order facts, not vice versa. But this seems to contradict Bradley's original worry: if the higher-order exemplification facts are needed to ground the claim that a exemplifies F-ness, then the explanation runs upward, not downward, making the regress vicious.

Criteria for vicious regress Following the contemporary literature [13], an infinite regress is generally considered vicious when it exhibits one or more of the following features:

1. *Explanatory circularity*: Each level in the regress requires the subsequent level for its explanation, such that no level provides a foundation (the explanation "chases its tail" infinitely)

2. *Ontological dependence*: Each fact or entity at level n depends for its existence on a distinct fact or entity at level $n+1$, creating an infinite chain of dependencies

3. *Epistemic inaccessibility*: To know that p_1 is true, we must first know that p_2 is true, but to know that p_2 is true we must first know that p_3 is true, ad infinitum, making knowledge impossible

A regress is benign if the infinite series exists but does not exhibit these problematic features–for instance, if all higher-level facts follow automatically from a ground-level fact without requiring independent verification or grounding.

Implications for moral properties The Bradley regress poses special difficulties for moral realism specifically. If the regress undermines realism about Universals generally, then moral properties like virtue, honesty, courage, and responsibility cannot be grounded as objective features of reality. The realist moral philosopher who claims that "Torture is wrong" expresses the objective fact that torture exemplifies the Universal property of moral wrongness faces all the problems discussed above.

Moreover, moral properties present an additional challenge: they seem to require not just exemplification but also a normative force–a reason-giving or action-guiding dimension. Even if we could solve Bradley's regress for descriptive properties like redness or roundness, it remains unclear whether this solution extends to normative properties. The statement "This action exemplifies wrongness" must not only pick out an objective property but must also somehow explain why this property gives us reason to avoid the action.

If Bradley's regress is vicious, moral realism collapses into nominalism or conventionalism about moral properties. On such views, moral categories would be useful classifications without ontological significance–more akin to legal categories (which vary by jurisdiction and convention) than to natural kinds. The implications for moral responsibility are profound: if "moral agent" is a conventional designation rather than a recognition of metaphysical fact, then ascriptions of moral responsibility become matters of social practice rather than objective truth. We might still hold people responsible, but this practice would lack the metaphysical foundation many believe it requires.

Nominalism

There is competing philosophical doctrine to realism, called nominalism, which disclaims the existence of Universals. Nominalists believe that realism needlessly introduces the strange notion of Universals, exhibits incoherency, and is burdened by a busy ontology. In its place they propose a supposedly simpler but sufficient account for the apparent categorical structure of reality.

Nominalists submit that only actual particulars have independent existence.

After all, how could a repeatable entity, like *redness*, simultaneously inhere in multiple entities, like an apple and a firetruck, such that the *redness* in the apple is numerically identical to the *redness* in the firetruck? The Realist is introducing a strange idea that entities with independent existence have non-local, unbounded physical presence. So, realists claim that the sentence, "a is F" denotes "a exemplifies *F-ness*" where *F-ness* is a Universal. In response, nominalists counter that "a is F" denotes "a is a member of the set of *F things*." For instance, "*Socrates* is courageous" denotes "*Socrates* is a member of the set of *courageous things*." Nominalists introduce *sets* (which are rigorously defined mathematical entities) to do the work of Universals. And the primary pay-off to this is that they avoid having to invoke the strangeness of Universals.

Set An unordered collection of entities considered as a whole. The identity conditions for sets is such that, if a and b are sets, they are numerically identical iff for each entity in a that entity is in b. Using mathematical language, $a \equiv b$.

Abstract reference For simple subject-predicate sentences, the nominalist's account as already given would seem sufficient. However, how might they interpret a sentence including abstract parts, like "Honesty is a virtue?" The realist would contend that this can be understood to mean, "*Honesty* exemplifies being a *virtue*," *honesty* and *virtue* picking out abstract Universals. But what might nominalists say? One possible interpretation might be, "The set of *honest acts* is a subset of the set of *virtued acts*." While this gives nominalists a little trouble, their response might seem sufficient enough. However, their problems do not end here.

Set equivalency How might a nominalist deal with the supposition that all the things that have hearts also have kidneys and vice verse? In other words, the set of *things that have hearts* and set of *things that have kidneys* have the same members. But according to the identity conditions for sets, that means that the set of *things that have hearts* is identical to the set of *things that have kidneys*. Consequently, *having a heart* is the same thing as *having a kidney* (the set of *things that have hearts*$\equiv$ the set of *things that have kidneys*), which is plainly absurd.

What is the nominalist's response to this? Generally, they invoke the terminology of *possible* entities. Instead of "a is F" denoting "a is a member of the set of *F-things*," it denotes "a is a member of the set of *possible F-things*." The set of *possible things that have hearts* and set of *possible things that have kidneys* do not have the same members, therefore they are not equivalent. Unfortunately, part of the justification for nominalism was that it has a simpler ontology than realism, but the introduction of *possible worlds* somewhat undermines this effort.

Conventionalism and its limits Even if nominalism offers a satisfactory account of attribute agreement and subject-predicate discourse, it remains nominative in character. There is no entity *honesty* as such–only sets of resembling things. The nominalist analyzes "My friend is honest" as meaning that my friend resembles other persons called honest. This appears to deny ontological significance to *honesty* as an independent property.

Does this undermine moral agency? Not necessarily. The debate between realism and nominalism reveals genuine metaphysical uncertainty about the status of properties, but neither position clearly eliminates moral responsibility. If realism is correct and moral properties are objective Universals, moral agency has robust metaphysical grounding. If nominalism is correct and moral properties are conventional classifications, moral agency may still be grounded–not in metaphysical facts about Universals, but in phenomenological facts about experience and in the practical efficacy of moral categories.

The error would be to assume that only the realist position can support moral responsibility. Nominalism tells us that moral categories may be human constructions, but this does not entail that they are arbitrary or dispensable. Many human constructions–language, mathematics, social institutions–are conventional yet indispensable for organizing experience and enabling human flourishing. Moral categories might be similarly conventional yet objectively important.

Moreover, both realism and nominalism may be incomplete frameworks. The very debate assumes we can categorically determine whether properties are "objective features of reality" or "conventional classifications." But this dichotomy may itself be too crude. Our cognitive architecture may prevent us from fully resolving whether the map (our conceptual schemes) or the territory (reality itself) is fundamental–or whether this question even makes sense.

From metaphysical foundations to agency Having examined the metaphysical status of moral properties themselves, we must now turn to the agents who supposedly bear moral responsibility. Even if we were to resolve the debate between realists and nominalists about moral properties, a separate but equally fundamental question remains: what are the criteria for being a moral agent? The metaphysical challenges regarding Universals and conventions resurface in our examination of personhood and agency. However, as with moral properties, the challenges reveal genuine puzzles rather than knockdown arguments against moral responsibility.

3 Persons and moral agency

Criteria for moral agency

> *"I know only that what is moral is what you feel good after and what is immoral is what you feel bad after."* –Ernest Hemmingway

What is the criteria for being a moral agent? By definition, a moral agent is a being who can make moral judgments. Thus, at a minimum, it would seem a moral agent should have the capacity to make rational decisions–that is, reason about the effects their decisions have on reality.

Core requirements for agency The traditional philosophical account of moral agency requires several interconnected capacities. First, the agent must possess *rationality*–the ability to form beliefs based on evidence, draw logical inferences, and understand causal relationships. Second, the agent requires *intentionality*–the capacity to form intentions and direct actions toward specific

goals. Third, the agent needs *reflective awareness*–the ability to consider one's own mental states and evaluate one's actions against moral principles.

Consider a concrete example: a physician deciding whether to administer an experimental treatment to a terminally ill patient. The physician must rationally evaluate the evidence for the treatment's efficacy, form an intention based on the patient's best interests, and reflect on the moral implications of the decision. Without any one of these capacities, genuine moral agency becomes questionable.

Degrees of moral responsibility Rather than treating moral agency as a binary property, we might consider it as existing along a spectrum. Young children, for instance, gradually develop moral agency as their cognitive capacities mature. They begin by understanding simple cause-and-effect relationships ("hitting hurts others"), progress to recognizing intentions ("I didn't mean to break it"), and eventually develop the capacity for moral reflection ("Was my action fair?").

This graduated view raises important questions about the boundaries of moral responsibility. If a person's rational capacities are temporarily impaired–through intoxication, extreme emotional distress, or sleep deprivation–to what extent are they morally responsible for their actions? The law recognizes such gradations through concepts like diminished capacity and mitigating circumstances, but the philosophical foundations remain contentious.

The knowledge condition Beyond basic cognitive capacities, moral agency seems to require adequate knowledge of relevant circumstances. An agent cannot be held fully responsible for consequences they could not reasonably have foreseen. This epistemic requirement introduces complications: what constitutes "reasonable" foresight? Should moral agents be held responsible for their ignorance itself?

Consider the historical example of physicians who prescribed mercury as medicine before its toxic effects were understood. Were they morally responsible for the resulting harm? The answer depends partly on whether the relevant knowledge was available and whether they had an obligation to seek it. This suggests that moral agency includes not just the capacity to make informed decisions but also the responsibility to become adequately informed.

So, how do persons measure up to these criteria? At the very least, they seem to have the capacity to reason about a limited domain of reality, and so are perhaps morally responsible within the limits circumscribed by that domain. But putting all of that aside, one can instead ask, "Do moral agents exist at all?" That is, are there subjects *making* decisions? This question about the very existence and persistence of agents through time leads us to fundamental issues of personal identity.

4 Personal identity and persistence

To consider whether moral agents exist at all–whether there are subjects making decisions–we must examine how entities, including persons, persist through time.

Persistence

To consider this question, let us analyze Chisholm's essay, *Identity through Time* [3]. In this essay, he distinguishes two ways in which entities persist.

Intactly persisting entities In the first way, entities persist in a numerically identical sense such that an entity, a, at time t_1 is the same as entity, b, at time t_2 iff the set of parts for entity a at time t_1 is equivalent to the set of parts for entity b at time t_2. Therefore, if $a \equiv b$, said entity persisted, at a minimum, $|t_2 - t_1|$ units of time.

Non-intactly persisting entities If an entity changes even a single part, it is not strictly the same. Therefore, since physical entities are in a constant state of flux, when one says physical entity a at time t_1 is the same as physical entity b at time t_2 (where $t_2 \neq t_1$), we do not mean that entities a and b are numerically identical. Rather, we mean to say that they are conventionally the same.

Ship of Theseus To clarify, let us consider the thought experiment of the *Ship of Theseus*, which raises the question of whether a ship which has had all its parts replaced retains the same identity. If the ship at t_1 has different parts than the ship at t_2 then they are not numerically identical; so, if the ships are identical, a non-intactly persisting sense of identity is being used. Chisholm contends that the question is not *is it the same ship* but *does it constitute the same ship?* And if it constitutes the same ship, that means it satisfies the Theory of Persistence.

> ***Theory of Persistence*** A composite entity persists if conditions 1 and 2 are satisfied:
>
> 1. The precedent composite entity evolved from a antecedent composite entity. Note: Evolved denotes that after each successive change, the antecedent composite must have at least one part in common with the precedent composite.
> 2. The composite entity must satisfy additional quantifiable criteria.

In practice, how might this theory provide a solution for the *Ship of Theseus* dilemma? First, we must determine whether it satisfies, from initial state to final state, the *evolution* criterion. And second, it must satisfy additional criteria, e.g., *has the same sailing schedule*. For example, if the evolution of the ship undergoes fission then the *evolution* criterion by itself cannot determine which ship constitutes the *Ship of Theseus*. However, if only one of the ships after the fission has the same sailing schedule as the ship before the fission, then according to the *has the same sailing schedule* convention, that ship constitutes the *Ship of Theseus*.

Persons

It is not even clear how *persons* should be defined, but providing an unambiguous definition for it would help to resolve many controversial. For instance, to resolve the dispute over abortion, one could categorically determine when a bundle of cells constitutes person-hood. There appears to be a consensus that motile sperm cells and ova are not persons, but that newborn babies are. What was the critical turning point in which it from from just being a bundle of cells to being a person who has a bundle of cells? Expending considerable time and effort could be avoided if a definitive categorical account of this problem was given, but like with the pile of sand, such a definition seems problematically conventional.

Chisholm on persons In light of the ***Theory of Persistence***, what can be said about persons? As with the *Ship of Theseus*, are person's identities also conventional? Or are they like sub-atomic particles which intactly persist? In *Which Physical Thing Am I* [4], Chisholm contends that *persons* are not conventional. In support of this, he asks you the reader to consider a hypothetical situation in which you are about to undergo an operation. The doctor presents you with two options: you can either choose the expensive option where you are subjected to total anesthesia during the operation, or you can choose the cheaper option. The cheaper option consists of the doctor giving you a pill before the operation to induce complete amnesia so that during the operation there is no memory of you, and giving you a pill after the operation so that everything that transpired during the operation is forgotten.

The question is, would it be reasonable for you to choose the cheap option? Chisholm contends that it would not be reasonable because you would believe that you would be the person enduring the operation. Furthermore, you would not be persuaded that this is not the case by any convention that denies that the person on the table is you. In other words, you would not consider your identity a matter of convention.

Critical response to Chisholm How might Chisholm's arguments be disputed? In the situation he presented, perhaps you would believe that you and the person enduring the operation in the cheaper option is the same person. Furthermore, perhaps no convention could persuade you from this conviction. However, your conviction could simply be based on a convention that you deeply subscribe to. At this point, it does not seem like Chisholm has established that persons warrant special categorical distinction.

Bundle Theory and Ego Theory In the essay *Divided Minds and the Nature of Persons* [16], Parfit contrasts two competing views of persons, Bundle Theory and Ego Theory. Bundle Theory proposes that unity of consciousness cannot be explained by referring to persons; Bundle Theorists maintain that persons only exist as a feature of language. Persons are a convention–they are a *label* for a *bundle of things* like thoughts, memories, and sensations that are unified by causal kinds of relations. There are no subjects in which these bundles inhere. But the Ego Theorist proposes that these subjects really do exist–as subjects of experiences–and that they *have* a *bundle of things* like thoughts, memories, and sensations.

In support of Bundle Theory, Parfit presents the reader with a hypothetical situation in which a replica of you is created, atom by atom. At one end of the spectrum, if one-hundred percent of your atoms are replaced, this is clearly a *replica of you*. At the other end, if only one percent of your atoms are replaced, this is clearly *you*. Thus, in between these two extremes there must be a critical turning point where it goes from being *you* to being a *replica of you*. However, this appears implausible because any point chosen would seem arbitrary. How could a single atom make a difference? It is similar to the question, *When does sand become a pile of sand?* It is absurd to think that, at one point it is not a pile, and with the addition of a single grain of sand, it becomes a pile.

A practical illustration: The case of Phineas Gage Consider the famous case of Phineas Gage, a railroad construction foreman who survived an accident in 1848 where an iron rod passed through his skull, destroying much of his left frontal lobe. Before the accident, Gage was described as responsible, industrious, and well-liked. After recovering, he became impulsive, profane, and unable to hold steady employment. His friends said he was "no longer Gage."

This case poses profound questions for both theories. For the Ego Theorist, is the post-accident person the same *subject of experience* as before? If Gage's personality, values, and behavioral patterns changed so dramatically, what persisted that makes him the same person? The Ego Theorist must claim there is some essential "Gage-ness" that survived despite these changes.

The Bundle Theorist has a different challenge: if Gage is just a bundle of psychological states and dispositions, and these changed dramatically, in what sense is he the same person? The Bundle Theorist might respond that there is sufficient psychological continuity–memories, basic cognitive abilities, learned skills–to maintain identity, but the degree of change raises questions about where to draw the line.

Ego Theorists have difficulty responding to this question because they believe that persons are *subjects of experiences*, not just *bundles of things*. As such, there must be some special unknown thing that makes you who you are that may not be included in the replication. Bundle Theorists have little problem answering these questions, however, because they deny that there is a *you*–only a *bundle of things*. If an exact copy of the *bundle of things* that constitutes you is made, what we choose to call this thing, *you* or a *replica of you*, is a matter of convention.

Split-brain studies and the nature of persons Parfit argues that empirical results from split-brain case studies provide evidence for Bundle Theory over Ego Theory. In pioneering research by Sperry [19] and Gazzaniga [9], patients who had their corpus callosum severed (to treat severe epilepsy) were presented with stimuli designed to reach only one hemisphere. When shown a placard whose left half is blue and right half is red, with each hemisphere receiving only one color, patients exhibited seemingly independent streams of consciousness–one aware of seeing only blue, the other aware of seeing only red.

Parfit takes such cases to pose a challenge for Ego Theory: if a person is essentially a subject of experience, and split-brain patients manifest two distinct subjects of experience, does this mean one person has become two? Or that

there were always two persons sharing a body? Neither option seems palatable.

However, this argument overstates the difficulty for Ego Theory. Contemporary Ego Theorists have developed several responses that avoid Parfit's dilemma:

The temporary fission response. Some Ego Theorists [15] argue that split-brain cases represent temporary, partial fission events. During tasks that elicit divided consciousness, there may indeed be two subjects of experience, but these are transient and incomplete. During normal activity, the hemispheres cooperate sufficiently that only one subject exists. The occurrence of temporary fission, while metaphysically interesting, does not undermine the Ego Theory's central claim that persons are subjects of experience–it merely shows that in extreme circumstances, subjects can divide or merge.

The dominant hemisphere response. Another strategy maintains that one hemisphere (typically the left, language-dominant hemisphere) houses the person's primary consciousness, while the right hemisphere's responses represent sub-personal processing rather than a full subject of experience. On this view, the right hemisphere's capacities, while sophisticated, do not constitute a complete person or subject. This response is supported by the fact that split-brain patients typically report a unified sense of self after surgery, centered in the language-capable hemisphere.

The animalist response. Eric Olson [15] defends animalism–the view that we are fundamentally biological organisms, not subjects of experience or psychological bundles. On this view, the split-brain patient remains one organism throughout, even if consciousness is temporarily divided. Personal identity consists in biological continuity, not psychological unity. Split-brain cases may show that psychological unity can be disrupted without destroying the organism, and hence the person.

Bundle Theory does have resources to handle split-brain cases more straightforwardly: if persons are just bundles of experiences unified by causal relations, then split-brain cases simply involve bundles that are partially separated. There is no deep puzzle about whether one person becomes two, because persons were never fundamental entities in the first place–they are conventional designations for sufficiently unified bundles of experiences.

However, Bundle Theory faces its own difficulties with split-brain cases. If psychological unity is what constitutes personal identity, and split-brain patients show divided psychology, we need principled criteria for when a single person exists versus when there are two. How much psychological unity is sufficient? The Bundle Theorist's answer risks being arbitrary or purely conventional.

The implications for moral responsibility These debates about personal identity and the nature of persons raise important questions for moral responsibility, but they do not settle the matter definitively. If persons do not persist through time in any robust sense, or if they are merely conventional designations for bundles of experiences, does this undermine holding them responsible for past actions?

The answer depends on what moral responsibility requires. If it requires a robust metaphysical subject–an unchanging ego that persists identically through time–then Bundle Theory poses a serious challenge. But perhaps moral respon-

sibility requires only sufficient psychological continuity and connectedness, not metaphysical identity. The person punished today need not be *numerically identical* to the person who committed the crime, as long as there is appropriate psychological continuity–memory, character, values–linking them.

Moreover, the phenomenology of personal persistence remains regardless of whether Bundle Theory or Ego Theory is correct. We experience ourselves as persisting subjects who remember our past actions, maintain character traits over time, and deliberate about future choices. This experiential continuity may provide sufficient grounding for moral responsibility, even if the metaphysical status of persons remains uncertain.

Both Bundle Theory and Ego Theory face challenges in grounding persistent agency, but neither clearly eliminates the possibility of moral responsibility. The debate reveals metaphysical uncertainty about the nature of persons, not the impossibility of holding them responsible for their actions.

5 The primacy of quality: A metaphysical inversion

The preceding sections have examined metaphysical challenges to moral responsibility–challenges concerning the reality of moral properties and the persistence of moral agents. These challenges assume a particular epistemic framework: that our experiences and moral practices must be validated by metaphysical theories about the deep structure of reality. But this framework may itself be backwards.

The map and the territory

Consider an observer standing on a mountain, examining a cartographic map of that very mountain. The map represents the mountain's features–its elevation, its ridges, its valleys–using quantitative abstractions like contour lines and numerical elevations. The map is useful precisely because it compresses the mountain's overwhelming complexity into a form that human cognition can process.

Now imagine this observer claiming that the mountain arose from the map, that the quantitative abstractions on paper are more fundamental than the lived reality of granite and ice. This inversion would be absurd. Yet this is precisely the epistemic reversal that characterizes much of modern philosophy and science regarding the relationship between experience and theory.

We directly experience *qualities*–the redness of an apple, the painfulness of a toothache, the moral wrongness of cruelty. These qualitative experiences are epistemically primary; they are what we actually undergo. To organize and predict these experiences, we develop quantitative models–wavelengths of light, neural correlates of pain, utility functions for moral reasoning. These models are extraordinarily useful tools, but they are *maps* of experiential *territory*, not the territory itself.

Yet modern thought has inverted this relationship. We treat quantitative, theoretical descriptions as fundamental and qualitative experiences as derivative

or even suspect. Wavelengths of light are "real"; the experienced redness is "merely subjective." Neural firing patterns are "real"; the felt quality of pain is "mysterious" or requires special explanation. This is the map-territory inversion: treating our theoretical compressions as more real than the phenomena they were designed to explain.

Historical reversal: From quality to quantity

This inversion is historically contingent, not inevitable. Pre-modern and contemplative traditions–Buddhist philosophy, phenomenology, certain strands of ancient Greek thought–treated qualities as epistemically primary. These traditions investigated experience directly, developing sophisticated vocabularies for describing and categorizing the qualitative textures of consciousness, emotion, and moral perception.

Quantification came later, as a tool for organizing qualitative experience. Mathematical models allowed prediction and control: if we quantify the properties of matter, we can predict when we will encounter certain qualities (heat, hardness, color). This instrumental success was revolutionary. But somewhere in this process, the tool became mistaken for the ground.

Modern science's insistence on quantification is methodologically sound–quantitative models enable precise prediction and experimental test. But this methodological commitment has hardened into a metaphysical claim: that only the quantifiable is truly real, that qualities are either reducible to quantities or are not fully real. This metaphysical claim does not follow from science itself. Science requires that we build useful predictive models; it does not require materialism or the primacy of quantity over quality.

The hard problem and cognitive bandwidth

The map-territory inversion creates what Chalmers calls the "hard problem of consciousness"–why do physical processes give rise to subjective experience? But this way of framing the problem already assumes that physical (quantitative) descriptions are fundamental and that qualitative experience is derivative and mysterious.

We might equally ask the inverse question: why do our qualitative experiences admit of quantitative description? If qualities are epistemically primary, the puzzle is not why quantities produce qualities, but why the qualitative structure of experience happens to exhibit patterns amenable to mathematical modeling.

Both framings may be equally puzzling–or equally misguided. The apparent tension between quality and quantity, between experience and physical description, may be an artifact of our cognitive architecture. Human cognition operates under severe bandwidth constraints. We cannot simultaneously attend to more than approximately seven items. Our working memory is minuscule compared to the complexity of what we experience.

These bandwidth constraints may force us to choose between experiencing qualities directly (phenomenology, contemplation) and developing quantitative

models (science, mathematics). We lack the cognitive capacity to simultaneously grasp both the full qualitative richness of experience and the complete quantitative structure that might describe it. The apparent incompatibility between quality and quantity may be a limitation of the map-maker, not a feature of the territory.

Metaphysical possibilities

What is the relationship between quantities and qualities? Between mathematical descriptions and experiential reality? Several metaphysical possibilities remain open:

1. **Materialism/Physicalism**: Quantities (physical properties, mathematical structures) are fundamental; qualities emerge from or reduce to quantitative relations. This is the dominant metaphysical framework in contemporary philosophy of mind and science.

2. **Idealism/Panpsychism**: Qualities (consciousness, experience, or proto-experiential properties) are fundamental; matter emerges from or is a type of experience. On this view, experiential character is built into the fabric of reality.

3. **Neutral Monism/Dual-Aspect Theory**: Quantities and qualities are two aspects of a more fundamental reality; neither reduces to the other. Reality has a structure that our categories of "physical" and "mental" only partially capture.

4. **Pluralism**: Both qualities and quantities are irreducibly real as distinct fundamental types. Reality has both mathematical structure and qualitative character as independent fundamental features.

5. **Question Ill-Formed**: The quality/quantity distinction may be an artifact of our cognitive architecture. Our bandwidth-limited cognition forces us to describe reality using either qualitative or quantitative frameworks, but reality itself may not be fundamentally either. The dichotomy is a cognitive artifact, and the question presupposes categories that don't carve reality at its joints.

Crucially, *we do not know which of these metaphysical positions is correct*, and our cognitive limitations may prevent us from knowing. We lack the cognitive bandwidth to definitively resolve this question. But we can ground ethics in phenomenology without settling this metaphysical question. The debates between realists and nominalists, between materialists and idealists, between Bundle theorists and Ego theorists, all assume we can definitively answer questions that may exceed our cognitive capacity.

Implications for moral philosophy

This analysis has profound implications for moral philosophy. Traditional moral realism demands that moral properties be objective features of reality, comparable to physical properties. But this demand accepts the map-territory inversion: it treats theoretical metaphysical categories as more fundamental than experienced moral reality.

The alternative is to ground ethics in phenomenology–in what we directly experience. We do not need to resolve whether "goodness" is an objective Universal, a conventional classification, or something our categories cannot fully capture. What matters is that we directly experience certain states as good (pleasure, flourishing, compassion) and others as bad (suffering, degradation, cruelty). These qualitative evaluations are epistemically prior to any theory about their metaphysical status.

This is not subjectivism or relativism. The claim is not that goodness is "merely subjective" or "just a matter of opinion." Rather, the claim is that the experiential reality of goodness and badness is more fundamental than our theoretical attempts to explain it. Just as the mountain is more fundamental than the map, the lived experience of moral value is more fundamental than metaphysical theories about whether value is "objective" or "conventional."

Contemplative traditions understood this. Buddhist ethics does not depend on resolving metaphysical questions about the ultimate nature of reality. It grounds moral practice in the direct experience of suffering (*dukkha*) and the possibility of liberation from suffering. The ethical imperative emerges from phenomenology, not metaphysics.

We can adopt a similar approach: acknowledge profound metaphysical uncertainty about the nature of persons, properties, and agency, while grounding moral practice in the phenomenology of experience. This is not a retreat from philosophical rigor but a recognition of our cognitive limits and a return to experiential foundations.

6 Phenomenological grounding for ethics

If we cannot resolve the metaphysical status of moral properties, persons, or free will, what foundation remains for ethics? The answer is phenomenology–the direct examination of experience itself. We need not resolve whether moral properties are Universals or conventions, whether persons are egos or bundles, or whether choices are determined or free. What we require is the recognition that certain experiences carry immediate normative content.

The immediacy of valence: From toothache to flourishing

Consider the experience of a toothache. There is something it is like to have a toothache–a distinctive qualitative character that Nagel [14] calls phenomenal consciousness. But the toothache is not merely a neutral quale. It is *bad*. Not bad because we judge it to be bad, or because social convention deems it bad, but

bad in its very phenomenology. The badness is given directly in the experience itself.

This immediate normative content extends across the full spectrum of valenced experience. Consider these concrete examples:

Physical pain and discomfort: A toothache, migraine, hunger pangs, nausea, the burning sensation of touching a hot surface–each presents itself as intrinsically bad in its phenomenology. We do not first experience these sensations neutrally and then judge them to be bad; the badness is given with the sensation itself. The toothache does not merely correlate with something we dislike; it is constituted by its aversive quality.

Physical pleasure and relief: The satisfaction of hunger after prolonged fasting, the relief when pain ceases, the warmth of sunlight after being cold, the pleasure of cool water on a hot day–each presents itself as immediately good. These experiences carry positive valence in their very phenomenology. We do not infer that satiation is good from some background theory; we experience it as good directly.

Emotional suffering: Grief, anxiety, despair, loneliness, humiliation–these emotional states present themselves as intrinsically bad. Someone in the grip of profound grief does not experience a neutral sensation that they then evaluate negatively; the suffering is phenomenologically constitutive of the grief itself.

Emotional flourishing: Joy, contentment, love, connection, pride in accomplishment–these emotional states carry immediate positive valence. The goodness of joy is not a judgment we make about an otherwise neutral feeling; it is intrinsic to the phenomenology of joy itself.

Aesthetic experience: While more complex and culturally mediated, aesthetic experiences often carry immediate normative content. The beauty of a landscape or musical passage can present itself as good in its phenomenology; the ugliness of a scene of destruction can present as bad. While our aesthetic judgments are shaped by culture and individual taste, the immediate experiential impact often precedes conscious evaluation.

The evolutionary grounding: These valenced experiences are fundamental reward signals that guide behavior across all sentient organisms. Evolution did not wait for organisms to develop metaphysical theories about value before making suffering aversive and flourishing attractive. Pain and pleasure are ancient biological adaptations that provided immediate behavioral guidance long before conscious reflection emerged. The toothache does not become bad when we theorize about it; it is bad as a fundamental feature of how sentience guides behavior.

This is not a theoretical claim about the metaphysical status of badness. I am not arguing that badness is an objective Universal property inhering in the toothache, or that we can derive an "ought" from the "is" of phenomenology. Rather, I am pointing to pre-theoretical normative data: suffering and flourishing present themselves with evaluative character prior to any theory about their metaphysical status. This evaluative character is epistemically prior to theories about whether value is "objective" or "subjective," whether it involves Universals or conventions.

It seems difficult to maintain that a toothache is morally neutral–that it is an arbitrary social convention to treat suffering as bad. These are fundamental reward signals that instruct our behavior at a level more basic than cultural learning or conscious deliberation. They are phenomenological bedrock.

We can question whether the badness of the toothache corresponds to an objective property in external reality. We can develop sophisticated theories about whether badness is a Universal, a conventional classification, or a projection of our evaluative faculties. But we cannot coherently deny the experiential reality of the toothache's badness–that would require ignoring what is most immediately given in experience. The mountain of experienced value is more certain than any map we might draw of its metaphysical structure.

Extension to sentience: Beyond anthropocentrism

If phenomenology provides the ground for ethics, the moral circle extends naturally to all sentient beings–all beings capable of experiencing suffering and flourishing. This is not anthropocentric favoritism but a principled extension of phenomenological ethics.

The capacity for phenomenal consciousness–for there to be something it is like to be an organism–is what matters morally, not membership in the category "person" or "human." A non-human animal experiencing pain undergoes the same kind of immediate normative reality as a human with a toothache. The suffering presents itself as bad in its phenomenology, regardless of the metaphysical status of the suffering organism.

This suggests a sentience-based ethics rather than a person-based ethics. We need not resolve whether non-human animals are "persons" in some metaphysical sense, or whether they possess the sophisticated cognitive capacities required for rational moral agency. What matters is that they can suffer and flourish–that their experiences carry immediate normative content.

This approach avoids arbitrary boundary-drawing. The traditional focus on "persons" as special moral agents risks excluding beings who clearly can suffer (animals, infants, cognitively impaired humans) while potentially including beings whose moral status is unclear (sophisticated AI systems, corporations). A phenomenologically grounded ethics focuses instead on the capacity for valenced experience–experience that presents itself as good or bad in its very phenomenology.

Practical efficacy: Restructuring reality toward preferred states

Beyond the immediate normative content of experience, phenomenological ethics is grounded in practical efficacy. We *can* restructure local reality toward states we experience as valuable and away from states we experience as harmful. This is not a metaphysical claim but an empirical observation about our capacity to intervene in the world. Consider concrete examples across multiple domains:

Medical advances: The history of medicine is a history of transforming the qualitative landscape of human experience. Anesthesia eliminated the agony that once made surgery a last-resort horror; what was once unendurable torment became manageable or absent. Antibiotics transformed bacterial infections from death sentences into minor inconveniences, fundamentally changing the experiential quality of what it means to fall ill. Vaccines prevented entire categories of suffering that once seemed inevitable. Analgesics provide relief from chronic pain, antiemetics reduce nausea, antidepressants can lift the fog of despair. Each of these interventions literally restructures experiential reality–not by changing our theories about suffering, but by reducing or eliminating the suffering itself.

Social institutions: Human cooperation has created structures that channel conflict away from violence and toward negotiation. Legal systems, when functioning well, transform disputes from blood feuds into manageable procedures. Democratic institutions (when they work) give voice to the powerless, reducing the suffering that comes from domination and disenfranchisement. Education transmits knowledge that reduces suffering from ignorance–teaching us how to purify water, grow food, treat wounds, avoid poisonous plants. Economic cooperation through markets and specialization means that individuals need not master every survival skill; we can benefit from others' expertise, raising overall living standards and reducing the grinding toil that characterized most of human history.

Agricultural and technological innovation: The development of agriculture, irrigation, food preservation, and modern farming techniques has dramatically reduced starvation and malnutrition–forms of suffering that were once pervasive. Shelter technology reduces exposure to harsh elements. Clean water infrastructure prevents waterborne diseases. Refrigeration prevents food poisoning. Electric lighting extends productive and social time beyond daylight hours. Climate control (heating and cooling) makes extreme temperatures survivable and comfortable. Each innovation moves the needle on actual experienced quality of life.

Individual practices: Contemplative traditions developed techniques for transforming the quality of experience from within. Meditation practices can reduce anxiety, increase equanimity, and cultivate compassion. Psychotherapy helps individuals recognize and restructure patterns of thought and behavior that cause suffering. Cognitive-behavioral techniques provide tools for managing depression and anxiety. Physical exercise reliably improves mood and mental health. These practices demonstrate that we can intervene not just in external circumstances but in the structure of experience itself.

Animal welfare improvements: We can reduce animal suffering through more humane farming practices, banning cruel confinement systems, developing alternatives to animal testing, protecting habitat, treating injured wildlife, and managing populations to prevent starvation. Veterinary medicine extends to animals the same kind of relief from pain and disease that human medicine provides to us. These interventions demonstrably affect the quality of experience for non-human sentient beings.

The common thread across all these examples is that they *work*–they reliably produce outcomes we experience as good and avoid outcomes we experience as bad. This practical efficacy is independent of metaphysical questions about whether the goodness is objectively real or conventionally assigned. Anesthesia relieves surgical pain regardless of whether pain is a Universal property or a conventional classification. Clean water prevents cholera whether or not suffering corresponds to objective features of reality.

Moreover, this practical capacity extends beyond human flourishing. We can promote flourishing not just for ourselves but for the broader community of sentient beings with whom we share the world. The fact that such interventions work–that they reliably affect the quality of experience for ourselves and others–provides pragmatic grounding for moral practice.

This is not consequentialism in the traditional sense, which requires comparing outcomes according to some maximizing principle (greatest happiness for greatest number). Rather, it is the simple recognition that we can affect whether suffering or flourishing occurs, and that this capacity for intervention is itself morally significant regardless of our metaphysical theories. Each time we successfully move the needle toward less suffering and more flourishing, we demonstrate the practical foundation of ethics.

Living with metaphysical uncertainty

The phenomenological approach allows us to bracket metaphysical questions that may be unanswerable given our cognitive architecture. We need not know whether:

- Moral properties are Universals, conventions, or something else entirely
- Persons are egos, bundles, or conventional designations
- Choices are determined, undetermined, or neither
- Reality is fundamentally mathematical, fundamentally experiential, or fundamentally something we cannot conceptualize

What we do know is that suffering is bad as given in experience, that flourishing is good as given in experience, and that we can reliably affect whether suffering or flourishing occurs. This experiential and practical knowledge provides sufficient grounding for moral agency and moral responsibility.

This is not anti-realism or eliminativism about moral properties. It is a recognition that the experiential reality of value–the badness of pain, the goodness of joy–is more fundamental than our theoretical explanations of it. Just as the mountain is more fundamental than the map, the lived reality of moral experience is more fundamental than metaphysical theories about moral properties.

Epistemic humility about metaphysics is compatible with–indeed, may require–confidence about phenomenology. We can be uncertain about the ultimate nature of reality while being certain that suffering is genuinely bad and that

we have reason to prevent it. This combination of metaphysical uncertainty and phenomenological confidence is not philosophically unstable; it is the appropriate response to our cognitive situation as finite beings with direct access to experience but limited access to ultimate metaphysical truth.

7 Challenges from determinism

> *"Responsibility: A detachable burden easily shifted to the shoulders of God, Fate, Fortune, Luck or ones neighbor. In the days of astrology it was customary to unload it upon a star."* –Ambrose Bierce, *The Devil's Dictionary*, 1911

Under what circumstances can a moral agent's conduct be ascribed moral responsibility? At a minimum, it would seem to require that its conduct be a product of being able to select, in a way that is unconstrained by causal agencies, a specific action among a set of possible actions–that is, a product of exercising a free will. For instance, an agent's decision to go shopping could be a product of a free will if it was possible for the agent to have decided not to go shopping. Conversely, if an agent stumbles off a building and free falls towards the ground, the act of free falling is certainly not a product of free will. Under the circumstances, the agent had no choice but too fall.

Four-dimensionalism

It is often said that entities have location and extension in a three-dimensional space. That is, physical entities can be specified in a three-dimensional Cartesian coordinate system (see Figure 1), x, y, and z, which provides three spatial dimensions–respectively, length, height, and width, e.g., a cubic entity extends m units along the x, y, and z axes.

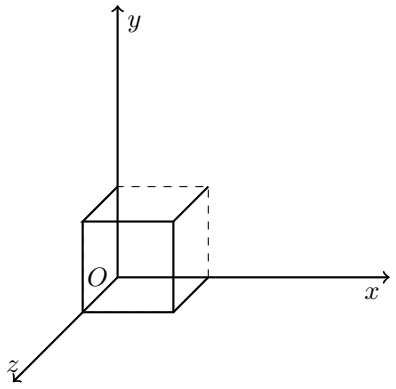

Figure 1: Three-dimensional Cartesian coordinate system with a unit cube illustrating spatial extension

However, our coordinate system is missing a crucial detail: at what time, t, will a particular entity be at a particular x, y, and z, and how long will it endure? Therefore, a physical entity's location and extension is more precisely specified in a four-dimensional coordinate system, x, y, z,and t, which provides three spatial dimensions and one temporal dimension, e.g., a hyper-cubic entity extends m units along the x, y, z, and t axes, or a person may extend an average of ≈ 0.6 meters along the x-axis, an average of ≈ 1.8 meters along the y-axis, an average of ≈ 0.3 meters along the z-axis, and ≈ 85.6 years along the t-axis.

Four-dimensional space-time has taken on a new-found importance with the advent of modern science. Under this model, when persons observe an entity at a particular time, they observe a single *time-slice* of it. More precisely, they observe a three-dimensional spatial "slice" (which is perpendicular to the temporal dimension) of a four-dimensional space-time entity. For example, when they observe a two-dimensional spatial slice of a sphere, they observe a disk; when they observe a three-dimensional spatial slice of a hyper-sphere, they observe a sphere.

At any time t, persons can only observe a three-dimensional spatial "slice" located at time t of a four-dimensional space-time entity. However, from a "God's eye" vantage point outside the spacetime manifold, one could in principle observe the entire four-dimensional structure at once. From this perspective, the passage of time would represent different locations along the temporal dimension rather than genuine change.

Perdurantism and moral responsibility Does four-dimensionalism undermine moral agency? A common but mistaken objection holds that if persons are four-dimensional entities, they are "unchanging" and therefore cannot make genuine choices. This objection rests on a confusion. Four-dimensional entities do change–they have different temporal parts with different properties at different times. A person-worm extending from birth to death has temporal parts that differ in location, height, beliefs, desires, and moral character. The difference is that for the four-dimensionalist (or perdurantist), change consists in different temporal parts having different properties, rather than a single three-dimensional entity gaining and losing properties over time.

The more serious question is whether this view of persistence supports moral responsibility. Consider: if I am a four-dimensional worm extending through spacetime, and my temporal part at t_1 commits a crime, can my temporal part at t_2 be held responsible for it? The perdurantist answer is yes, provided the appropriate relations hold between these temporal parts–typically, psychological continuity and connectedness. My t_2 temporal part remembers the crime, has the same character that produced it, and is psychologically continuous with the t_1 part. This is sufficient for moral responsibility on most accounts.

Indeed, perdurantism may have advantages for explaining moral responsibility. It naturally accommodates cases of psychological change: if Phineas Gage's pre-accident and post-accident temporal parts have very different psychological properties, we can say they are parts of the same four-dimensional person while acknowledging their profound differences. The question becomes whether there is sufficient continuity between temporal parts to ground responsibility, not

whether a persisting three-dimensional entity has changed too much.

Eternalism, foreknowledge, and choice A distinct but related issue concerns the relationship between four-dimensionalism and the nature of time. Many four-dimensionalists accept eternalism (the B-theory of time)–the view that past, present, and future events are equally real, and temporal passage is illusory. On this view, facts about what you will choose tomorrow already exist, though they are located at a different temporal location than the present.

Does this undermine free choice? If it is already a fact that you will choose coffee tomorrow, could you have chosen tea? This raises the ancient problem of foreknowledge and freedom. Some argue that if future facts exist timelessly, choice is illusory–you cannot make it the case that you choose coffee, because this is already a fact written into the four-dimensional spacetime structure.

Others respond that eternalism is perfectly compatible with free choice. The fact that your choice already exists in the four-dimensional block does not mean it is determined by prior events. The eternalist can distinguish between: (1) it being fixed which choice you will make, and (2) your choice being causally determined. Even if future facts exist, they might exist precisely because you will freely choose them, not because they are forced by prior causes. The future choice is fixed in the sense that it will happen, but this is compatible with its being free in the sense that it flows from your autonomous deliberation rather than external compulsion.

As Sider [18] argues, four-dimensionalism is a thesis about the persistence of objects through time, while eternalism is a thesis about the reality status of times themselves. Neither directly entails determinism. A four-dimensional person-worm can make genuinely free choices at different temporal locations, and these choices can be real facts without being causally determined facts.

Causal Determinism

Supposing that reality is not an immutable four-dimensional space-time, let us turn our attention to the question of causal determination. Causal determinism is the doctrine that all events are the inevitable result of antecedent conditions, including the actions of *persons*. The totality of existence is a game of cosmic billiards where, given an initial state as input, the final state could be output by a computable function.

The Laplacian demon Pierre-Simon Laplace [12] famously articulated the determinist position through his thought experiment of an omniscient intelligence–now known as Laplace's demon. This hypothetical being, knowing the precise location and momentum of every atom in the universe at a given moment, could use the laws of physics to reconstruct the entire past and predict the entire future. In Laplace's words, "nothing would be uncertain and the future, as the past, would be present to its eyes."

This mechanistic view gained considerable support from classical physics, where Newton's laws seemed to govern all motion with mathematical precision. Every effect had a cause, every cause had a prior cause, stretching back in an

unbroken chain to the beginning of time. Human actions, being physical events in a physical brain, would be no exception to this universal causation.

Is causal determinism compatible with moral agency? In causal determinism, there is only one possible future; acts of apparent choice are the mechanical expression of a supposed moral agent's heredity and past environment. Since its future is determined, that would denote that moral responsibility is not possible, for at the very least a moral agent must have the power to make choices between morally *good* and *bad* actions.

The incompatibilist position, defended by philosophers like van Inwagen [21], holds that determinism and free will cannot coexist. If our actions are the inevitable result of the laws of nature and events in the remote past–events over which we had no control–then we cannot be morally responsible for them. This is formalized in van Inwagen's Consequence Argument: no one has power over facts about the distant past; no one has power over the laws of nature; therefore, no one has power over facts about the future (including their own actions).

The compatibilist alternative Not all philosophers accept the incompatibilist conclusion that determinism undermines moral responsibility. Indeed, compatibilism has historically been the majority position in philosophy and remains influential in contemporary debates. Compatibilists argue that free will and moral responsibility are compatible with causal determinism, provided we properly understand what freedom and responsibility require.

Classical compatibilism: Hume's analysis. The classical compatibilist position, articulated by David Hume [10], argues that freedom is not the absence of causation but rather a particular kind of causation. An action is free when it flows from the agent's own desires, beliefs, and character, rather than from external compulsion or constraint. On this view, "liberty" requires "necessity"–that is, for an action to be attributable to an agent, it must necessarily follow from that agent's psychological states.

Consider the difference between two scenarios: (1) You decide to donate to charity because you value helping others and believe this charity does important work; (2) You donate to charity because someone holds a gun to your head and threatens to shoot if you don't. In the first case, your action flows from your own values and reasoning–this is paradigmatically free, even though your values and reasoning were themselves shaped by prior causes (your upbringing, experiences, etc.). In the second case, your action is compelled by external threat–this is unfree. The compatibilist claims this distinction is all the freedom worth wanting, and determinism does not collapse it.

Frankfurt's hierarchical model. Harry Frankfurt [7, 8] developed a sophisticated compatibilist account centered on second-order desires. On his view, persons are distinguished from other agents by their capacity to form desires about their desires–to want to want certain things. Freedom consists in having one's will (first-order desires that move one to action) aligned with one's second-order desires about what one's will should be.

Frankfurt's famous thought experiments challenge the principle of alternate possibilities–the idea that moral responsibility requires the ability to do otherwise. Consider a case where a neurosurgeon has implanted a device in your brain

that will make you choose coffee if you start to choose tea, but the device never activates because you choose coffee on your own. Frankfurt argues you are morally responsible for choosing coffee even though you could not have done otherwise, because the choice flowed from your own reasoning and values. What matters is not whether alternative possibilities exist, but whether you act on your own authentic reasons.

Reasons-responsiveness theories. John Martin Fischer and Mark Ravizza [6] developed an influential compatibilist account based on reasons-responsiveness. On their view, an agent is morally responsible for an action if the mechanism that produces the action is responsive to reasons–that is, if the agent would have acted differently if presented with sufficient reason to do so. This approach allows for moral responsibility even in a determined world, as long as the causal processes governing behavior are sensitive to rational considerations.

For instance, suppose determinism is true and your decision to help a drowning child was causally necessitated by prior events. Fischer and Ravizza argue you are still morally responsible if the mechanism producing your action (your practical reasoning) is such that, had there been overwhelming reasons not to help (e.g., helping would cause ten other children to drown), you would have responded to those reasons and acted differently. The actual causal necessitation does not undermine responsibility as long as the mechanism is reasons-responsive.

The randomness objection answered. Compatibilists also argue that indeterminism provides no advantage for moral responsibility. If your decision to help or harm someone depends on genuinely random quantum events in your neurons, how does this make you more responsible than if the decision depends on your character and values? Randomness seems to undermine responsibility rather than enable it. As Daniel Dennett [5] argues, the alternative to determinism is not freedom but chaos. The libertarian's indeterministic agent seems less like a responsible person and more like a roulette wheel.

Challenges to compatibilism. Despite its sophistication, compatibilism faces significant objections. Incompatibilists argue that even if an action flows from one's character, if that character was itself determined by factors outside one's control (genetics, early environment, etc.), then the agent is not truly responsible. Galen Strawson [20] presents what he calls the "Basic Argument": to be truly responsible for an action, you must be responsible for the way you are (your character, values, etc.); but to be responsible for the way you are, you must have chosen to be that way; but to choose to be a certain way, you must already have a character that makes that choice; this leads to infinite regress or an arbitrary starting point for which you are not responsible.

Moreover, some argue that compatibilism changes the subject–that when ordinary people think about free will and moral responsibility, they have in mind a libertarian conception incompatible with determinism, and compatibilists offer a different, weaker concept and simply label it "freedom." Whether this objection succeeds depends partly on empirical questions about folk concepts and partly on normative questions about what freedom is worth wanting.

Quantum indeterminacy and libertarian free will The advent of quantum mechanics has introduced fundamental indeterminacy into our picture

of reality. At the quantum level, events may be genuinely probabilistic rather than determined. Some libertarians about free will have seized on this as providing the metaphysical room needed for genuine choice. However, critics point out that random quantum events hardly seem a better foundation for moral responsibility than deterministic ones. If your decision to help or harm someone depends on quantum coin-flips in your neurons, how does this make you more responsible than if it depends on prior causes?

Robert Kane [11] has developed a sophisticated libertarian account that attempts to navigate between determinism and randomness. On his view, quantum indeterminacy may play a role in moments of moral conflict, where we experience genuine uncertainty about what to do. The resolution of this conflict, while not determined, is not merely random but shaped by our efforts of will. The indeterminacy provides the metaphysical space for self-formation, where we become responsible for our character through these undetermined but non-random choices.

An argument from absurdity: the coherency of choice For the sake of argument, assume choice is not an illusion. How then might we offer a rational account of it? But if something is rational, does that not signify that it can be explicated in terms of its antecedent conditions? How else can a physical process be rationally explained if not by cause-and-effect? But if choice can be understood in terms of its antecedent conditions, then it is not free–it was determined by cause-and-effect.

This dilemma–that free will seems impossible whether determinism is true or false–has led some philosophers to hard incompatibilism. Derk Pereboom [17] and others argue that free will is incompatible both with determinism and with indeterminism, and therefore moral responsibility in the traditional libertarian sense is impossible.

Yet this conclusion is not universally accepted, and the debate remains very much alive. Hard incompatibilism represents one position in a contested philosophical landscape that includes sophisticated compatibilist defenses, libertarian accounts attempting to navigate between determinism and randomness, and pragmatic approaches that bracket the metaphysical question entirely.

Moreover, even hard incompatibilists typically argue that we can maintain many moral practices and attitudes without ultimate moral responsibility. We can still shape behavior through praise and blame, protect society from dangerous individuals, and express moral emotions like guilt and indignation–understanding these as forward-looking practices rather than backward-looking judgments of ultimate desert. This suggests that moral agency as a practice may be sustainable even if the metaphysical foundations remain uncertain or controversial.

The phenomenological approach developed in the previous section offers another response: the experience of deliberation and choice is epistemically prior to theories about whether choice is determined or free. We experience ourselves as deliberating between alternatives, weighing reasons, and deciding on courses of action. This experiential reality may ground moral responsibility regardless of whether our choices are metaphysically free in the libertarian sense, determined but reasons-responsive in the compatibilist sense, or something our metaphysical

categories cannot fully capture.

8 Concluding thoughts

This paper has examined three fundamental metaphysical challenges to moral responsibility: the reality of moral properties, the persistence of persons across time, and the compatibility of moral agency with determinism. While these challenges reveal genuine philosophical puzzles about the deep structure of reality, they do not establish that moral responsibility is impossible. Instead, they demonstrate the limits of certain metaphysical frameworks and point toward phenomenological grounding for ethics.

Embracing metaphysical uncertainty The central claim is that we should embrace moral agency as a practice grounded in phenomenology, despite unresolved and possibly unresolvable metaphysical questions. Our cognitive bandwidth is severely limited–we cannot simultaneously grasp more than approximately seven items in working memory, cannot hold both the full qualitative richness of experience and complete quantitative theoretical models in mind at once. These limitations may prevent us from definitively answering whether reality is fundamentally mathematical or experiential, whether the map or the territory is prior. The honest philosophical position is intellectual humility: the apparent tensions between quality and quantity, between experience and theoretical description, may be artifacts of our bandwidth-limited cognition rather than features of reality itself.

The epistemic priority of experience Yet metaphysical uncertainty does not entail ethical paralysis, because experience is epistemically prior to theory. We directly undergo suffering and flourishing, deliberation and choice. The experiential reality–that suffering presents itself as bad in its very phenomenology–is more certain than any metaphysical theory about it. We stand on the mountain of experience while consulting maps of metaphysical theory; the mountain is more certain than any map. This is not anti-realism but a recognition of epistemic order: experience comes first, theoretical explanation comes second. Modernity inverted this order, treating quantitative models as more fundamental than the qualitative experiences they were designed to explain.

Phenomenological foundations for moral practice What can we know with reasonable confidence, independent of disputed metaphysical theories?

First, we know that suffering is bad and flourishing is good as given in experience. This is pre-theoretical normative data, epistemically prior to theories about whether goodness is a Universal or a convention.

Second, we know this extends to all sentient beings–all beings for whom there is something it is like to experience suffering or flourishing. The moral circle is not limited to "persons" in some metaphysically loaded sense, but includes all beings capable of valenced experience.

Third, we know that we can reliably restructure local reality toward states characterized by flourishing and away from states characterized by suffering. This practical efficacy is empirically demonstrable across cultures and throughout

history. Medicine works, agriculture works, cooperation works–they reliably affect the quality of experience regardless of their metaphysical status.

Fourth, we know that we experience ourselves as deliberating, choosing, and bearing responsibility for our choices. This phenomenology of agency persists regardless of whether our choices are metaphysically free in the libertarian sense, determined but reasons-responsive in the compatibilist sense, or something our categories cannot fully capture.

These four claims provide sufficient grounding for moral agency and moral responsibility. They do not require resolving whether moral properties are Universals, whether persons are egos or bundles, or whether determinism is true. They require only acknowledging what is most immediately given in experience and what is empirically demonstrable about our capacity to affect experiential quality.

Better uncertain and ethical than certain and cruel One might object: without metaphysical foundations, isn't this just pragmatism or conventionalism? No. The phenomenological approach grounds ethics in experiential reality, not convention or utility. The badness of suffering is given directly in experience; the goodness of flourishing is phenomenologically real. What is conventional is our *theoretical interpretation* of these experiences, not the experiences themselves. Moreover, epistemic humility about metaphysics combined with confidence about phenomenology is philosophically more honest than dogmatism about unverifiable foundations. It is better to be uncertain about ultimate reality while acting to prevent suffering, than to be certain about metaphysical theories while remaining paralyzed by philosophical puzzles.

Living at the limits of understanding We are bandwidth-limited creatures who must act despite metaphysical uncertainty. Perhaps future beings with greater cognitive bandwidth will resolve the metaphysical questions that perplex us, grasping simultaneously both qualitative and quantitative aspects of reality. Or perhaps the questions themselves will dissolve, revealed as artifacts of our particular cognitive architecture. But we are not those beings. The appropriate response is not paralysis but phenomenologically grounded action. We acknowledge what we do not know while acting on what we do know: that suffering is bad, that we can prevent it, and that choosing to do so is what moral agency consists in.

The question transformed The ancient question "Are persons moral agents?" can now be reframed. If the question asks whether persons have some metaphysically robust property of agency establishable by philosophical argument, the answer is: we do not know, and our cognitive limitations may prevent us from knowing. But if the question asks whether we experience ourselves as deliberating and choosing, whether some experiential states are genuinely bad and others good, whether we can reliably affect which states occur, and whether this provides sufficient grounds for moral practice–then the answer is yes.

Moral agency may not have the metaphysical foundations that naive realism assumes. It is grounded instead in phenomenology, practical efficacy, the immediate normative content of experience, and our capacity to intervene to prevent

suffering and promote flourishing. These foundations are more certain than disputed metaphysical theories because they are grounded in what we directly experience rather than theoretical constructs.

We can be uncertain about the ultimate nature of reality while being certain about the reality of suffering and our capacity to prevent it. This is not philosophical weakness but philosophical honesty–living at the limits of human understanding while remaining committed to reducing suffering and promoting flourishing for all sentient beings, a commitment grounded not in metaphysical certainty but in phenomenological immediacy and practical efficacy.

The map may be incomplete, but the mountain is real, and the climb continues.

References

[1] David M Armstrong. *Universals: An opinionated introduction*. Westview Press, 1989.

[2] Francis Herbert Bradley. *Appearance and reality: A metaphysical essay*. Swan Sonnenschein & Co., 1893. Second edition 1897, Oxford University Press.

[3] Roderick M Chisholm. Identity through time. In Howard E Kiefer and Milton K Munitz, editors, *Language, Belief, and Metaphysics*, pages 163–182. State University of New York Press, 1970.

[4] Roderick M Chisholm. Which physical thing am i? an excerpt from "is there a mind-body problem?". *Metaphysics: The Big Questions*, pages 296–300, 1990.

[5] Daniel C Dennett. *Elbow room: The varieties of free will worth wanting*. MIT Press, 1984.

[6] John Martin Fischer and Mark Ravizza. *Responsibility and control: A theory of moral responsibility*. Cambridge University Press, 1998.

[7] Harry G Frankfurt. Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23):829–839, 1969.

[8] Harry G Frankfurt. *The importance of what we care about*. Cambridge University Press, 1988.

[9] Michael S Gazzaniga. Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7):1293–1326, 2000.

[10] David Hume. *A treatise of human nature*. John Noon, 1739. Modern editions available from Oxford University Press.

[11] Robert Kane. *The significance of free will*. Oxford University Press, 1996.

[12] Pierre-Simon Laplace. *Essai philosophique sur les probabilités*. Courcier, 1814. English translation: A Philosophical Essay on Probabilities, Dover, 1951.

[13] Anna-Sofia Maurin. Infinite regress: Virtue or vice? *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*, 2007. URL `www.fil.lu.se/hommageawlodek`.

[14] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83 (4):435–450, 1974.

[15] Eric T Olson. *The human animal: Personal identity without psychology*. Oxford University Press, 1997.

[16] Derek Parfit. Divided minds and the nature of persons. *Mindwaves*, pages 19–28, 1987.

[17] Derk Pereboom. *Living without free will*. Cambridge University Press, 2001.

[18] Theodore Sider. *Four-dimensionalism: An ontology of persistence and time*. Oxford University Press, 2001.

[19] Roger W Sperry. Hemisphere deconnection and unity in conscious awareness. *American Psychologist*, 23(10):723–733, 1968.

[20] Galen Strawson. The impossibility of moral responsibility. *Philosophical Studies*, 75(1-2):5–24, 1994.

[21] Peter van Inwagen. The incompatibility of free will and determinism. *Philosophical Studies*, 27(3):185–199, 1975.