

# The God Who Learned Morality From You:

## Divine Dependence, Moral Construction, and the Evidentiary Problem of Human Conduct

### **Abstract**

Traditional philosophy of religion frequently asks whether morality depends upon God, whether moral truths exist independently of divine command, or whether ethical obligations can be grounded in reason, human nature, flourishing, or the consequences of action. This paper introduces a thought experiment that reverses the conventional direction of dependence. Suppose incontrovertible evidence establishes that the universe was created by an immensely intelligent being that nevertheless possesses no innate moral knowledge. The creator produced conscious life not to test humanity against a preexisting moral order, but to discover morality through observation of human civilization. After millennia of studying human conduct, the creator establishes communication with humanity and asks what it ought to value. It further declares that whatever ethical theory humanity collectively endorses will thereafter be incorporated into the metaphysical structure of reality and become objectively binding. This scenario creates a problem of moral selection in the apparent absence of any antecedent moral standard. If no objective morality exists before humanity's choice, it is difficult to explain how humanity could choose incorrectly. Yet if some independent criterion exists by which humanity's decision could be judged mistaken, then the creator is not genuinely generating morality through human selection. The problem deepens when the creator proposes that observed human behavior, rather than professed ethical doctrine, constitutes the strongest evidence of humanity's actual values. Human beings routinely affirm impartiality, dignity, justice, compassion, and universal rights while behaving according to partiality, self-interest, loyalty, hierarchy, and exclusion. The thought experiment therefore develops into an epistemological problem: what evidence could humanity provide to demonstrate that its articulated moral ideals deserve greater normative authority than its revealed behavioral preferences? Particular attention is given to humanity's capacity for corrigibility: its ability to judge its own desires, institutions, and inherited practices defective. Corrigibility is treated not as an escape from the grounding problem but as a distinctive epistemic datum that itself requires normative interpretation. The resulting argument challenges divine command theory, moral realism, constructivism, ideal-observer theories, theories of reflective endorsement, error theory, expressivism, and the conventional distinction between moral belief and moral conduct. Its final implication is unsettling: a universe morally patterned after humanity's actual behavior would likely be regarded by humanity itself as profoundly unjust, although that reflexive condemnation constitutes pressure on anti-realist theories rather than a deductive refutation of them.

## Introduction

One of the oldest questions in moral philosophy concerns the source of moral authority. Human beings routinely speak as though certain actions are not merely preferred or socially advantageous but genuinely right or wrong. Murder, cruelty, betrayal, torture, exploitation, compassion, loyalty, honesty, and justice appear to invite judgments that transcend ordinary preference. Yet the metaphysical status of those judgments remains disputed. Are moral truths discovered or created? Do they exist independently of human beings? Are they grounded in reason, nature, divine will, social agreement, rational procedure, flourishing, or the structure of conscious experience?

Philosophy of religion intensifies this question by introducing God as a possible foundation for morality. If an omnipotent and omniscient creator exists, one might suppose that moral truth ultimately derives from divine nature, divine reason, or divine command. Yet this possibility generates the ancient problem associated with Plato's *Euthyphro*: are actions morally right because God commands them, or does God command them because they are morally right? If the former, morality risks becoming arbitrary. If the latter, morality appears independent of God.

The thought experiment developed here radically alters the structure of this dilemma. Instead of asking whether humanity receives morality from a divine creator, suppose the creator attempts to receive morality from humanity.

Imagine that incontrovertible evidence emerges demonstrating that the universe was intentionally created by an intelligent being. The discovery initially appears to vindicate religious intuitions about cosmic purpose. Humanity has not arisen accidentally. Consciousness is not merely an unintended consequence of physical processes. The cosmos has an architect.

Then a second discovery changes the philosophical significance of the first.

The creator is extraordinarily intelligent but possesses no innate moral understanding.

It understands mathematics, causation, logic, physics, psychology, prediction, language, and perhaps every descriptive fact about the universe. Yet it does not know what ought to matter. It can determine what organisms desire, what produces suffering, what stabilizes civilizations, what increases reproductive success, and what consequences follow from particular actions, but none of these observations automatically generates a normative conclusion. It possesses knowledge of what *is* without knowledge of what *ought to be*.

The creator therefore produced conscious beings precisely because morality could not be derived from purely descriptive information. Humanity was not created as the subject of a divine moral examination. Humanity was created as an experiment through which the creator hoped to discover what morality is.

Human history is therefore reinterpreted.

Every war becomes data.

Every friendship becomes data.

Every betrayal, sacrifice, genocide, reconciliation, murder, act of forgiveness, revolution, marriage, law, punishment, rescue, act of mercy, political institution, and civilization becomes evidence observed by a morally uncertain creator attempting to infer what conscious beings treat as valuable.

Eventually humanity establishes direct communication with the being.

The creator asks a question whose simplicity conceals its philosophical enormity:

“What should I value?”

Humanity immediately discovers that it has no unified answer.

Religious traditions appeal to revelation, divine nature, sacred law, or theological anthropology. Utilitarians argue that suffering should be minimized and well-being maximized. Deontologists defend duties, constraints, autonomy, and rights. Virtue ethicists argue that morality concerns the cultivation of excellent character. Contractualists appeal to principles that rational persons could not reasonably reject. Care ethicists emphasize relationships, dependency, and responsibility. Existentialists deny that externally imposed moral meaning can replace authentic human choice. Moral nihilists deny that objective moral truths exist at all. Cultural traditions disagree concerning justice, punishment, hierarchy, sexuality, property, family, authority, violence, loyalty, and obligation.

The creator listens and then makes an extraordinary announcement:

“I will adopt whichever ethical theory humanity collectively endorses tomorrow. Because I control the structure of reality, that theory will thereafter become objectively binding.”

For the first time, humanity is not attempting to discover morality.

Humanity must decide what morality will become.

The scenario produces several philosophical puzzles simultaneously. Can humanity make the wrong choice if no objective standard exists prior to the decision? What could transform a collective preference into an objective moral truth? Would divine metaphysical enforcement create genuine normativity or merely universal coercion? Should humanity choose the theory it currently believes to be correct, or the theory it wishes future beings to inhabit? And most importantly, what evidence should the creator use to determine what humanity genuinely values?

The thought experiment becomes substantially more disturbing when the creator announces that its own empirical analysis has already produced a candidate answer.

After observing human beings for thousands of years, it concludes that the principle humanity most consistently behaves as though it accepts is approximately this:

Protect those you identify with and rationalize the treatment of everyone else.

Humanity protests.

That principle, we insist, is not morality. It represents prejudice, weakness, tribalism, historical failure, political fear, and moral inconsistency. Our behavior should not be mistaken for our ideals.

The creator asks:

“Why should your stated principles reveal morality better than your choices?”

This paper argues that the question exposes a neglected problem in moral philosophy: the evidentiary gap between normative profession and revealed valuation. The issue is not simply whether humanity possesses moral knowledge. It is whether humanity could demonstrate possession of such knowledge to an external intelligence that judges values from behavior rather than aspiration.

The deeper reversal is therefore complete. Instead of asking whether human beings can know the morality of God, we must ask whether a god could justifiably infer morality from us.

### **Thesis, Method, and Philosophical Location**

The paper's central thesis is that the creator's proposal cannot generate non-arbitrary objective normativity merely by converting a collective human endorsement into a law of reality. The scenario exposes three logically distinct problems. First, the **grounding problem** asks whether a moral standard can be selected correctly when no prior standard exists by which correctness could be assessed. Second, the **aggregation problem** asks how billions of incompatible judgments could be converted into a single "human" choice without presupposing controversial norms governing representation, equality, agenda control, and decision procedure. Third, the **evidentiary problem** asks what an external observer should treat as the best evidence of human values when moral language and moral behavior diverge.

The argument can be stated schematically. **P1:** If humanity can choose the future moral order incorrectly, some normatively authoritative standard must exist prior to the choice. **P2:** If no normatively authoritative standard exists prior to the choice, then no procedure for selecting among rival moral orders can itself be morally privileged merely as a matter of morality. **P3:** A collective choice nevertheless requires an aggregation rule, and every plausible aggregation rule

incorporates contestable assumptions about whose preferences count, how strongly they count, and how conflict is resolved. **P4:** The creator's power to enforce or metaphysically instantiate the selected system does not by itself establish that the selected system was justified prior to that enforcement. **P5:** Human behavior provides evidence of motivational priority, but human moral language, institutional self-binding, and practices of self-criticism also constitute behavior and therefore cannot be dismissed as evidentially inert. **P6:** Consequently, neither revealed preference nor professed doctrine alone is sufficient to establish moral truth; the thought experiment instead reveals a conflict among moral realism, constructivism, error theory, and competing theories of moral epistemology.

The discussion therefore stands at the intersection of the Euthyphro problem, Hume's is/ought distinction, Kantian autonomy, utilitarian aggregation, Rawlsian proceduralism, contractualism, constructivist accounts of normativity, Mackiean error theory, and contemporary convergence arguments about reasons (Plato 1997; Hume 2000; Kant 2012; Mill 1998; Rawls 1999; Mackie 1977; Korsgaard 1996; Scanlon 1998; Parfit 2011).

---

## **I. The Reversal of Divine Command**

Divine command theories traditionally locate moral obligation in some relation to divine authority. In their simplest form, actions are obligatory because God commands them, prohibited because God forbids them, or morally valuable because they conform to God's nature or will.

The present thought experiment reverses this dependence.

The creator does not possess morality and subsequently communicate it to humanity. Instead, humanity appears before the creator as the only available source from which normative information might be learned.

This distinction immediately separates omniscience concerning facts from moral knowledge. A being could hypothetically know every descriptive truth concerning reality while lacking any knowledge concerning what should be valued. It could know the neurological structure of pain without knowing whether pain ought to be reduced. It could know every consequence of murder without knowing whether murder is wrong. It could predict perfectly the social effects of deception while remaining ignorant of whether truthfulness possesses intrinsic value.

The distinction resembles David Hume's famous observation that normative conclusions cannot simply be derived from descriptive premises. Knowledge that suffering exists does not logically entail that suffering ought to be prevented unless some normative premise is already

introduced. Likewise, perfect understanding of human desire does not automatically establish that desires should be satisfied.

The creator therefore confronts what might be called the problem of primordial normativity. If the universe initially contains only facts and no values, then intelligence alone cannot necessarily generate morality.

Creating conscious life becomes an attempt to solve this problem empirically.

Yet the attempt may be conceptually mistaken. Observing what conscious beings do reveals preferences, motivations, social structures, psychological tendencies, and behavioral regularities. It does not necessarily reveal what they ought to do.

The creator might therefore accumulate an infinitely detailed description of human conduct without acquiring a single valid normative proposition.

Humanity could respond to the creator by making precisely this distinction:

“What we do does not determine what is right.”

But this response immediately generates another question:

“How do you know?”

If humanity claims access to a standard independent of observed behavior, it must explain the epistemic source of that standard.

The thought experiment thus transforms the creator from moral legislator into skeptical interrogator.

---

## **II. Morality Before the Choice**

The central paradox appears when the creator declares that humanity's collective decision will determine objective morality.

Suppose humanity chooses utilitarianism. The following day, maximizing aggregate welfare becomes objectively morally required.

Suppose instead humanity chooses a rights-based deontology in which certain individual liberties cannot be violated even to produce greater aggregate benefit. Those rights become objectively binding.

Suppose humanity chooses virtue ethics. Character and flourishing acquire metaphysical priority.

Suppose it chooses a religious moral code. That code becomes objectively authoritative.

Before the decision, none possesses objective priority.

After the decision, one does.

The question then arises:

Could humanity choose incorrectly?

If there is genuinely no moral truth before the selection, then describing any selection as morally wrong appears incoherent. To call the decision wrong would require comparison against a standard that supposedly does not yet exist.

This creates what may be called the Pre-Normativity Problem.

Let:

$M_0$  = the moral state of reality before humanity's decision.

$M_1$  = the moral state of reality after humanity's decision.

If  $M_0$  contains no objective moral facts, then no objective moral proposition exists according to which humanity ought to select utilitarianism rather than deontology, equality rather than hierarchy, compassion rather than cruelty.

The choice cannot be morally evaluated because moral evaluation becomes possible only after the choice occurs.

The point can be stated more formally. Let the transition from the pre-normative state to the post-selection state be represented as  $T: M_0 \rightarrow M_1$ , and let  $E(T)$  represent any normatively authoritative evaluation of that transition as correct, incorrect, better, worse, permissible, or impermissible. If  $E$  is genuinely normative, then either  $E$  belongs to  $M_0$  or it does not. If  $E$  belongs to  $M_0$ , then  $M_0$  was not normatively empty and the transition is governed by an antecedent standard. If  $M_0 = \emptyset$  with respect to objective normativity, then  $E(T)$  is normatively undefined prior to  $T$ . The creator may still describe  $T$  causally, predict its consequences, or compare it with human preferences, but none of those descriptive relations supplies the missing evaluative function without introducing an additional normative premise.

There is, however, a second difficulty hidden inside the phrase "humanity collectively endorses." Humanity is not a unitary chooser. It is a population of agents with different moral rankings, different information, different interests, and often incommensurable conceptions of the good. To produce one answer for the creator, those heterogeneous judgments must be aggregated by some social-choice mechanism. Let each person possess an ordering over candidate moral systems and let a social-choice function map those individual orderings onto a collective

ordering or winner. The moment this machinery is specified, normativity re-enters through the back door. Who belongs to the electorate? Does each person receive equal weight? Are children represented? Do future persons count? What about nonhuman sentient beings whose interests would be governed by the resulting morality? Should preference intensity matter? May minorities exercise vetoes? Who determines the agenda and the admissible options? Is unanimity required, or a majority, supermajority, deliberative consensus, lottery, constitutional convention, or some weighted scheme of representation?

Social-choice theory demonstrates that aggregation is not a philosophically neutral technical afterthought. Arrow's general possibility theorem shows that, for three or more alternatives, no rank-order social welfare function can simultaneously satisfy a familiar set of attractive conditions including unrestricted domain, Pareto responsiveness, independence of irrelevant alternatives, and non-dictatorship while guaranteeing a coherent collective ordering (Arrow 1963). Arrow's theorem does not establish that democratic decision-making is impossible, nor does it prove that every voting rule is equally defective. Its relevance here is narrower and more fundamental: there is no uniquely neutral procedure waiting to translate plural individual judgments into "humanity's" judgment. Choosing the aggregation rule means choosing which procedural values to privilege.

The creator's demand therefore generates an **aggregation regress** alongside the Pre-Normativity Problem. If humanity lacks a shared normative basis, it cannot appeal to that basis to justify majority rule over consensus, equal weighting over expertise, present persons over future persons, or any other aggregation architecture. Yet without an aggregation architecture there is no collective endorsement for the creator to adopt. The procedure that is supposed to create or select morality must itself be selected by criteria that the scenario stipulates have not yet acquired objective moral authority. The collective decision is therefore underdetermined not only at the level of substantive ethics but at the level of social choice.

Yet this conclusion produces disturbing implications.

Humanity could theoretically select a doctrine according to which cruelty is intrinsically good, domination is virtuous, or innocent suffering is morally required.

If such a choice would become objectively binding simply because humanity selected it, then objective morality appears to be created by procedural fiat.

The word "objective" would then require clarification.

The creator may possess the ability to restructure reality so that actions conforming to the selected doctrine receive metaphysical reinforcement. Perhaps moral properties themselves become instantiated. Perhaps violations produce cosmic consequences. Perhaps rational beings become objectively obligated in some metaphysical sense.

Yet one may still ask whether metaphysical enforcement is equivalent to moral truth.

If the creator declares torture morally good and modifies reality accordingly, has torture actually become good?

Or has an overwhelmingly powerful being merely altered the universe so that torture is rewarded, commanded, or metaphysically privileged?

This resurrects the Euthyphro problem in modified form.

The issue is no longer:

“Is something good because God commands it?”

It becomes:

“Can something become good merely because a sufficiently powerful being changes reality so that it counts as good?”

Power and normativity may not be identical.

An omnipotent creator may be capable of determining the consequences of action without determining their moral significance.

---

### **III. The Impossibility of Choosing Correctly Without Prior Morality**

Suppose humanity wishes to deliberate rationally before the decisive vote.

What should it do?

Immediately, every method of deliberation appears to smuggle moral assumptions into the process.

Should humanity select the ethical system that minimizes suffering?

That already privileges suffering as morally significant.

Should it choose the system that respects autonomy?

That assumes autonomy possesses normative importance.

Should it select the system most conducive to human flourishing?

That assumes flourishing is valuable.

Should it choose the theory all rational agents could endorse?

That assumes rational acceptability carries moral authority.

Should it protect minorities from majority domination?

That assumes minorities possess claims that numerical majorities may not override.

Should future generations receive representation?

That assumes obligations extend to persons who do not yet exist.

Even the procedure by which humanity chooses an ethical theory requires ethical justification.

Should every person receive one vote?

Why equality?

Should experts receive greater influence?

Why expertise?

Should historically oppressed groups receive additional representation?

Why corrective justice?

Should children vote?

Should nonhuman animals receive representation?

Should artificial intelligences?

Should future persons?

Should the decision require unanimity, majority rule, supermajority approval, global deliberation, random selection, or constitutional convention?

Each procedural option presupposes normative commitments.

The paradox therefore deepens.

Humanity must choose morality through a decision-making process whose legitimacy itself depends upon morality.

The problem resembles a normative regress. In order to choose a moral framework correctly, humanity needs a morally justified decision procedure. But to justify the procedure, it must already possess moral criteria.

If moral criteria already exist, however, morality is not genuinely being created by the decision.

Humanity is merely attempting to discover it.

This generates a dilemma:

Either morality exists before humanity's choice, in which case the creator's announcement does not truly create morality,

or morality does not exist before the choice, in which case there is no morally correct way to make the choice.

The thought experiment thus reveals a difficulty for strong forms of moral constructivism. A normative order cannot easily be constructed from entirely nonnormative materials, because the rules governing construction already appear to require normative authority.

---

#### **IV. The Creator as Moral Learner**

Perhaps the creator does not expect humanity to create morality from nothing. Perhaps it expects humanity to reveal morality indirectly through consciousness.

One possibility is that moral truths emerge from certain structural properties of conscious existence.

Pain is experientially aversive.

Agency creates interests.

Social dependence creates vulnerability.

Mutual recognition creates obligations.

Reason generates consistency constraints.

Finite beings require cooperation.

If so, the creator may have generated conscious life because morality becomes intelligible only from within conscious experience.

This possibility moves the thought experiment closer to moral naturalism or constructivism.

Moral value might arise from facts about sentience, agency, vulnerability, flourishing, reciprocity, or rationality.

Yet even here disagreement remains.

The fact that suffering is experienced negatively does not automatically establish whether one person's suffering may be justified by another's flourishing.

The fact that persons possess agency does not determine whether autonomy should ever be overridden.

The fact that communities require cooperation does not determine whether loyalty to one's community outweighs duties to outsiders.

The existence of interests does not determine how conflicting interests should be ranked.

Consciousness may therefore generate the raw material of morality without determining its final structure.

The creator's research project remains unresolved.

---

## **V. Humanity's Competing Answers**

When the creator asks, "What should I value?" humanity responds with incompatible traditions.

The utilitarian argues that conscious welfare provides the clearest candidate for moral significance. Pleasure, suffering, flourishing, preference satisfaction, or some related conception of well-being should determine moral action. The creator should value outcomes.

The deontologist objects that some actions remain impermissible even when they maximize desirable consequences. Persons should not merely be treated as containers of welfare. Rights, duties, autonomy, consent, and the intrinsic status of persons impose constraints upon optimization.

The virtue ethicist argues that both positions distort moral life by reducing ethics either to rules or calculations. The proper question is not simply "What action is required?" but "What kind of being should one become?" Courage, wisdom, temperance, justice, generosity, integrity, and practical judgment cannot be replaced by algorithmic calculation.

The contractualist asks what principles could be justified to others under conditions of reasonable deliberation.

The care ethicist insists that abstract impartiality often obscures the moral significance of dependency, relationship, embodiment, and unequal vulnerability.

The existentialist may reject the creator's question entirely. Asking an external authority to determine value mistakes the structure of moral agency. Meaning is not discovered as a cosmic property but produced through commitment and responsibility.

The moral realist claims that the creator's ignorance does not imply the absence of moral facts. Just as a powerful being could fail to know mathematics, it could fail to know morality.

The moral anti-realist argues that humanity should stop pretending that objective value exists independently of valuers.

The divine command theorist faces the strangest predicament. If God asks humanity what God should value, divine authority can no longer straightforwardly ground morality.

Each position can offer arguments.

None can provide an answer that all rational humans already accept.

The creator therefore encounters not one humanity, morally speaking, but thousands.

### **V.A. The Mackiean Error-Theory Counter-Argument**

At this point a Mackiean error theorist can reject a premise shared by both the realist and the stronger constructivist interpretations of the scenario. Humanity's inability to choose "wrongly" before the creator acts need not show that morality is waiting to be constructed. It may instead confirm that there is no objective moral fact for the choice to track. Mackie's canonical position is more precise than the claim that moral sentences simply lack truth-value. Ordinary moral judgments purport to describe objective values or prescriptions and are therefore truth-apt; the error, on Mackie's account, is that the objective values presupposed by such judgments do not exist. Hence ordinary categorical moral claims systematically fail rather than awaiting completion by a later vote (Mackie 1977, 15-49).

This response gives the thought experiment a different interpretation. Before the creator intervenes, statements such as "utilitarianism is objectively correct" or "absolute rights are objectively binding" are, on the error-theoretic view, false because reality contains no such objective prescriptions. Humanity's inability to make an objectively wrong selection is therefore not evidence that an unconstrained constructive act is legitimate; it is evidence that the normative target is absent. Nor does the creator's subsequent enforcement automatically rescue objectivity. If the creator merely attaches rewards, punishments, psychological dispositions, or institutional consequences to one code, then it has created an extraordinarily powerful convention, not the kind of stance-independent moral value Mackie denies. If the creator literally inserts intrinsically action-guiding moral properties into reality, the scenario stipulates into existence precisely the type of metaphysically unusual "objective prescriptions" that Mackie's argument from queerness was designed to challenge.

The error-theory objection therefore sharpens rather than dissolves the experiment. The relevant trilemma becomes: either moral facts pre-exist the vote, in which case humanity is trying to discover rather than create them; or they do not pre-exist but are genuinely brought into existence by the creator, in which case the metaphysics of this transformation requires explanation; or the creator merely enforces a code, in which case objective morality has not

been created at all. The second horn also raises a supervenience problem. Moral realists typically hold, in one form or another, that normative differences cannot arise without some difference in the relevant non-normative base. But if the creator takes a fixed descriptive state and simply declares or causes a new normative status to supervene upon it, what explains that dependence relation? If the newly instantiated moral properties are natural properties, the creator owes an account of the bridge principles identifying or reducing the normative to the natural. If they are non-natural, as on familiar non-naturalist realist views, the creator appears to be generating *sui generis* normative properties or relations whose existence and connection to descriptive reality cannot be explained merely by rearranging physical or psychological matter (Moore 1993; Shafer-Landau 2003). The thought experiment by itself cannot decide among these possibilities, but it prevents the absence of pre-choice correctness from being treated as an uncomplicated victory for constructivism.

---

## **VI. Stated Morality and Revealed Morality**

At this point the creator introduces the most important complication.

It has not merely listened to philosophers.

It has observed history.

Its dataset includes imperial conquest, slavery, religious persecution, genocide, colonialism, warfare, charitable institutions, rescue efforts, family loyalty, political revolutions, scientific cooperation, torture, humanitarian law, friendship, corruption, sacrifice, racism, forgiveness, nationalism, mutual aid, economic exploitation, and ordinary interpersonal life.

From this evidence, the creator attempts to infer humanity's operative ethical system.

It notices a persistent pattern.

Human beings frequently advocate universal principles while applying them selectively.

They condemn violence when performed by enemies and justify it when performed by allies.

They defend freedom while restricting the freedom of disfavored groups.

They condemn propaganda while accepting narratives flattering to their own communities.

They invoke fairness when disadvantaged and merit when advantaged.

They defend property in some contexts and redistribution in others.

They condemn civilian suffering in one conflict while rationalizing comparable suffering in another.

They demand forgiveness for members of their own group and punishment for outsiders.

They insist that every human life possesses equal value while distributing attention, wealth, protection, grief, and political urgency in radically unequal ways.

From these patterns, the creator develops a descriptive hypothesis:

Humanity's most stable practical ethic is not universalism.

It is conditional partiality.

More provocatively:

Protect those you identify with and rationalize the rest.

Humanity reacts with horror.

That cannot be our morality.

The creator responds that it is not interested merely in what humans say they believe.

It is interested in what their behavior reveals they believe.

The distinction resembles the economic concept of revealed preference. If an individual repeatedly claims to prefer one thing while consistently choosing another when genuine costs are involved, an observer may reasonably infer that the chosen option possesses greater practical priority.

Moral life may contain an analogous phenomenon: revealed normativity.

Human beings say justice matters.

But what are they willing to sacrifice for justice?

They say strangers possess equal moral worth.

But how much comfort will they surrender to protect strangers?

They say truth matters.

But does truth retain priority when it threatens identity, reputation, wealth, political allegiance, or family loyalty?

They say suffering matters regardless of who experiences it.

Does behavior support that claim?

The creator's challenge is therefore epistemological rather than merely accusatory.

Why should declarations be treated as more reliable evidence of values than costly action?

---

## **VII. The Evidentiary Problem of Moral Ideals**

Humanity can produce several responses.

First, it may argue that moral failure presupposes moral standards. The fact that humans violate their principles does not disprove those principles. A dishonest person may sincerely believe honesty is right while failing to act accordingly.

This is undoubtedly true.

Human beings experience temptation, weakness, fear, ignorance, coercion, habit, bias, and conflicting motivations. Behavior therefore cannot straightforwardly be equated with moral belief.

Aristotle distinguished knowing the good from reliably acting according to it. Christian theology developed elaborate accounts of sin precisely because moral knowledge does not guarantee moral obedience. Modern psychology similarly documents gaps between stated values and behavior.

The creator could accept this distinction while asking a harder question:

How can it distinguish moral aspiration from socially useful self-description?

Suppose humanity says:

“We believe all persons have equal moral worth.”

The creator examines wealth distribution, warfare, migration policy, criminal punishment, medical access, education, labor conditions, environmental exposure, and political representation.

It observes enormous disparities.

Humanity replies:

“We fail to live according to our principles.”

The creator asks:

“What evidence demonstrates that the principle, rather than the behavior, represents your deeper morality?”

This is difficult.

A professed ideal might constitute genuine moral knowledge imperfectly implemented.

But it might also constitute moral self-flattery.

Human beings may prefer to imagine themselves committed to universal justice because the self-conception is psychologically desirable while remaining unwilling to bear the costs that universal justice would require.

The creator therefore needs criteria for distinguishing aspiration from rationalization.

---

### **VIII. Costly Moral Evidence**

One possible answer is that values are revealed most clearly when agents accept significant costs to uphold them.

A person who claims to value honesty provides stronger evidence when telling the truth damages personal interests.

A society claiming to value rights provides stronger evidence when it protects the rights of unpopular minorities.

A government claiming to value international law provides stronger evidence when legal constraints limit its own strategic advantage.

A community claiming to value equality provides stronger evidence when equality requires surrendering inherited privilege.

Moral credibility may therefore correlate with costly consistency.

Before turning to sacrifice as evidence, however, humanity can offer a more sophisticated hermeneutic defense of moral language itself. The creator's initial contrast between "what humans say" and "what humans do" treats linguistic and non-linguistic behavior as though the former were merely commentary on the latter. That is too simple. Moral utterances are themselves actions with distinctive social functions. A public claim such as "persons have equal rights" does not merely report an inward preference. It can establish standards of criticism, coordinate expectations, license demands for justification, organize coalitions, structure institutions, authorize sanctions, and bind speakers to positions that can later be invoked against their short-term advantage. Promises, constitutional prohibitions, professional codes, human-rights declarations, religious commandments, and public principles all use language to alter the strategic environment in which subsequent conduct occurs.

This creates an important asymmetry between ordinary revealed preference and moral discourse. Much non-linguistic conduct is generated under immediate biological and social

pressures: fear, scarcity, kin preference, status competition, retaliation, coalition loyalty, and temporal discounting. Moral language often functions precisely as a coordination technology for resisting those pressures. By converting a private aspiration into a public rule, agents make defection reputationally visible, furnish others with grounds for complaint, and create mechanisms through which present selves can constrain future selves. A prohibition on torture, for example, has moral significance not because governments never violate it, but because its articulation can create legal, institutional, and reputational costs that would not exist if the prohibition had never been formulated. In this sense, stated principles can be behavior directed against other behavior. This may be called the **institutional self-binding defense**: moral language acquires evidential importance when its public articulation deliberately restructures incentives so that agents become answerable to standards capable of frustrating their own foreseeable short-term interests.

The hermeneutic defense does not establish that declarations are sincere, much less that they are true. Moral language can be propaganda, camouflage, prestige signaling, or self-deception. But the correct evidentiary conclusion is not that deeds are real while words are unreal. Rather, different forms of behavior carry different evidential weights. Cheap verbal affirmation is weak evidence; stable public commitment that creates institutional constraints, exposes the speaker to criticism, survives changes in self-interest, and coordinates costly collective action is stronger evidence. The creator should therefore interpret moral language not as a transparent window into an inner moral essence, but as one component of humanity's behavioral record, especially significant when it is used to construct counter-incentives against predictable short-term partiality.

This offers humanity a potential response to the creator.

Do not infer our morality merely from our average behavior.

Look instead at cases in which individuals knowingly sacrificed power, wealth, security, status, reputation, or life in order to uphold principles they considered binding.

The abolitionist who opposed slavery despite social cost may reveal something important.

The whistleblower who exposes wrongdoing at personal risk may reveal something important.

The dissident who defends the rights of ideological enemies may reveal something important.

The rescuer who protects strangers during persecution may reveal something important.

The soldier who refuses an unlawful order may reveal something important.

The person who forgives when revenge is available may reveal something important.

Perhaps moral ideals are visible precisely where behavior departs from ordinary self-interest.

The creator could therefore treat exceptional moral action as higher-quality evidence than statistically common conduct.

Yet the problem remains.

Why privilege costly impartiality over costly fanaticism?

Extremists also sacrifice comfort, wealth, freedom, and life for principles.

Martyrdom demonstrates commitment, not moral correctness.

The willingness to pay a cost establishes sincerity but not truth.

Humanity still requires a criterion for determining which sacrifices reveal morality and which reveal delusion.

---

## **IX. Reflective Endorsement**

A second response appeals to reflective endorsement.

Perhaps a person's true moral commitments are not identical to immediate behavior but to the principles they would endorse under conditions of adequate information, rational consistency, freedom from coercion, and sufficient reflection.

This approach appears in different forms throughout moral philosophy.

One may imagine an idealized version of humanity stripped of factual ignorance, tribal bias, impulsive fear, propaganda, and self-serving inconsistency.

The creator could ask not:

“What do humans currently do?”

but:

“What would humans endorse if they understood fully what their actions did to everyone affected?”

This resembles ideal-observer theory, contractualism, Rawlsian abstraction, and certain forms of constructivism.

Perhaps humanity's actual behavior should be discounted because much wrongdoing emerges from epistemic distortion.

A racist society may support discrimination because its members possess false beliefs about another group.

A population may support war because propaganda conceals civilian suffering.

A consumer may purchase exploitative products because the production chain remains invisible.

Remove ignorance and behavior may change.

Yet idealization introduces another problem.

Which features of actual humanity should be corrected?

Ignorance presumably should be removed.

What about partiality toward family?

What about religious commitment?

What about emotional attachment?

What about fear?

What about ambition?

What about national identity?

At what point does the “idealized human” cease to represent humanity?

If the creator removes every feature responsible for moral disagreement until consensus appears, the resulting agreement may reflect the assumptions built into the idealization procedure rather than a discovered moral truth. This is one version of the alienation problem in moral theory: an ideal standpoint can become so purified of ordinary attachments, vulnerabilities, partial commitments, and bounded cognitive conditions that the resulting verdicts cease to have recognizable normative authority for the actual agents whose lives they are supposed to govern (Railton 1984). The difficulty is not that idealization is always illegitimate. Some distortions, such as factual ignorance or coercion, appear obvious candidates for correction. The difficulty is that there is no normatively neutral specification of the idealization baseline.

The creator therefore inherits the same bootstrapping problem that confronted humanity's voting procedure. To ask what an “idealized humanity” would choose, the creator must first decide which human traits count as distortions and which count as morally relevant features of agency. Removing misinformation seems epistemic; removing fear, partiality, attachment, ambition, embodiment, or loyalty is much more controversial. The filter of idealization is itself a normative construction. If the creator already knows which traits should be discounted, it possesses antecedent normative criteria and no longer depends wholly upon humanity for moral discovery. If it does not, then appealing to idealized humanity merely relocates the

original regress from the choice of moral theory to the choice of which humanity is authorized to choose.

---

## **X. The Problem of Tribal Partiality**

The creator's proposed principle of tribal partiality deserves separate consideration because it is neither obviously absurd nor easily dismissed.

Human morality evolved within small social groups. Kinship, reciprocal cooperation, reputation, coalition formation, punishment of defectors, suspicion of outsiders, and loyalty to allies plausibly played substantial roles in the development of human moral psychology.

Human beings are therefore not naturally impartial moral calculators.

Moral concern is typically distributed concentrically.

Self.

Family.

Friends.

Community.

Nation.

Religious or ideological group.

Humanity.

Other sentient beings.

The strength of obligation often decreases with perceived distance.

Many moral traditions attempt to expand the circle of concern. Universal human rights, cosmopolitanism, humanitarian ethics, and impartial consequentialism can be interpreted as attempts to resist tribal constraints.

But the creator observes that expanded moral language often coexists with persistent partiality.

This creates a distinction between descriptive morality and aspirational morality.

Descriptively, humanity may be deeply tribal.

Aspirationally, many humans endorse universal principles.

Which represents humanity more authentically?

There is no simple answer.

Biological origin does not determine normative legitimacy. Aggression may be evolutionarily natural without being morally justified. Conversely, universalistic ideals may represent cultural developments that correct limitations in evolved psychology.

The creator must therefore decide whether morality should reflect what humans are or what they attempt to become.

This distinction may constitute the deepest issue in the thought experiment.

---

## **XI. Can Moral Progress Exist Before Objective Morality?**

The concept of moral progress poses another difficulty.

Human societies commonly describe the abolition of slavery, expansion of political participation, condemnation of torture, recognition of women's rights, protection of children, racial equality, and broader acceptance of human rights as moral progress.

But progress implies movement toward a standard.

If objective morality does not yet exist, in what sense has humanity progressed?

One answer is internalist.

Progress occurs relative to humanity's increasingly coherent values.

For example, if a society claims that all persons possess dignity while permitting slavery, abolition reduces contradiction between principle and practice.

Moral progress may therefore involve increasing consistency rather than approaching an independent moral reality.

Another answer is realist.

Humanity has gradually discovered objective truths that existed all along, imperfectly and unevenly.

But if this is true, the creator should not ask humanity to create morality.

It should ask humanity to help discover morality.

The entire thought experiment therefore functions as a diagnostic tool for metaethical commitments.

Anyone who insists that humanity could choose the wrong moral system appears implicitly committed to some form of moral realism or at least to normative standards existing independently of the final collective decision.

Anyone who insists that no choice could be wrong before morality is created accepts a more radical constructivism but must confront the possibility that horrific systems could become objectively valid through selection.

---

## **XII. The Creator's Power and the Nature of Objectivity**

The creator claims that it can make humanity's chosen morality objectively binding because it controls the structure of reality.

But what exactly could this mean?

Several possibilities exist.

The creator could alter psychological nature so that all rational beings experience certain actions as morally obligatory.

It could alter consequences so that morally approved actions reliably produce flourishing.

It could establish supernatural rewards and punishments.

It could modify metaphysical laws so that moral properties attach to actions.

It could alter rational cognition so that denial of the chosen principles becomes logically incoherent.

Yet none obviously demonstrates that the selected principles become morally true.

Universal agreement is not identical to truth.

Universal enforcement is not identical to truth.

Universal psychological compulsion is not identical to truth.

Even metaphysical instantiation raises questions if moral properties themselves remain conceptually obscure.

### **XII.A. The Mechanics of Metaphysical Instantiation**

The phrase "make morality objectively binding" therefore conceals the central metaphysical difficulty rather than resolving it. The creator begins, by stipulation, without normative understanding. It possesses causal power over the descriptive structure of reality, but causal

power alone does not explain how an ought emerges from an is. At least two models are available, and each carries substantial commitments.

On a naturalist model, the creator could alter the world so that some natural property or relation becomes the fixed base upon which moral predicates depend. It might, for example, rigidify a bridge law according to which maximizing a specified form of welfare constitutes rightness, or according to which violations of autonomous agency constitute wrongness. Yet this model does not eliminate the grounding question. Why should that natural property count as normatively authoritative rather than merely causally privileged? If the bridge principle is itself contingent upon the creator's selection, the normative work appears to be performed by a stipulation whose authority remains unexplained.

On a non-naturalist model, the creator could be said to instantiate *sui generis* moral properties of the kind defended by non-naturalist realists: irreducibly normative properties that are neither identical with nor reducible to ordinary natural properties (Moore 1993; Shafer-Landau 2003). This avoids straightforward reduction, but it creates a different puzzle. If two worlds are descriptively identical immediately before and after the creator's decree, except that one now contains the property of objective rightness attached to a class of acts, what metaphysical relation makes that attachment non-arbitrary? Normative supervenience ordinarily constrains moral difference by requiring an appropriate difference in the underlying non-normative base. A creator that simply adds a non-natural normative relation by fiat seems less to explain normativity than to stipulate the existence of the very relation in question.

A third possibility is that the creator changes not merely the moral label but the underlying natural facts as well: psychology, reasons-responsiveness, consequences, social relations, or the constitutive conditions of agency. In that case supervenience is preserved more straightforwardly, because normative differences accompany descriptive differences. But the is/ought problem remains at the level of the bridge law connecting them. The creator can make certain actions universally beneficial, universally desired, psychologically compelling, or rationally unavoidable; it still requires an account of why those descriptive facts constitute moral reasons rather than merely universal causal facts.

The mechanics of instantiation therefore reproduce the paper's central dilemma at the metaphysical level. The creator can change what happens, what agents desire, what rules govern consequences, and perhaps what properties exist. What remains obscure is how a being that begins without normative standards selects the relation by which those changes become reasons, duties, rights, or values rather than merely features of a newly engineered cosmos.

The creator may therefore be able to make a moral system inescapable without making it justified.

This distinction resembles the difference between authority and power.

A ruler may possess the power to enforce a law without possessing legitimate authority to enact it.

Likewise, a creator may possess cosmic power without thereby possessing moral authority.

The thought experiment therefore undermines the assumption that metaphysical supremacy automatically produces normative supremacy.

---

### **XIII. Humanity as God's Moral Evidence**

The most radical implication of the scenario concerns the evidentiary relationship between creator and creation.

Traditional theodicy asks why a morally perfect God permits evil.

Here, the question is reversed.

What if evil exists because the creator does not yet know that it is evil?

Human suffering becomes observational material within divine moral research.

This possibility drastically alters the interpretation of history.

A genocide is not permitted for some mysterious divine purpose.

It is observed because the creator is trying to determine what humanity's response reveals.

A rescue is not rewarded because it conforms to divine law.

It becomes evidence from which the creator might infer that compassion possesses value.

The moral responsibility of humanity therefore expands dramatically.

Humans are not merely agents acting within a moral universe.

They are potential authors of the moral universe.

Every action becomes pedagogical.

The creator is watching not to judge humanity but to learn from it.

This inversion creates a new form of existential responsibility.

If humanity's conduct constitutes evidence about morality, then ordinary hypocrisy acquires cosmic significance.

When humans preach universal dignity and practice selective concern, they teach contradictory lessons.

When institutions celebrate justice while protecting advantage, they provide evidence that justice may be rhetorical rather than fundamental.

When individuals sacrifice themselves for strangers, they provide counterevidence.

Human civilization becomes a disputed moral text, and the creator becomes its interpreter.

---

#### **XIV. What Would Humanity Want the Creator to Learn?**

The thought experiment ultimately forces a distinction between three different concepts:

what humanity does,

what humanity says it values,

and what humanity wishes it were capable of valuing consistently.

These categories may diverge sharply.

Human beings may behave selfishly while endorsing generosity.

They may endorse generosity while privately wishing they could become genuinely impartial.

Which level deserves normative authority?

Perhaps morality should be understood neither as present behavior nor as verbal doctrine but as aspirational self-correction.

Human beings possess the unusual capacity to evaluate their own motivations and reject them.

A person may feel hatred and conclude that the hatred is unjustified.

A society may inherit hierarchy and later condemn it.

A civilization may recognize that practices once considered normal were morally grotesque.

This capacity for self-revision may provide the strongest evidence that behavior alone does not exhaust human values.

A creature wholly defined by preference would never judge its own preference defective.

Humans do.

They experience guilt, shame, regret, moral admiration, and demands for reform.

They create institutions designed to constrain predictable selfishness.

They articulate principles that criticize their own societies.

They honor individuals who resisted prevailing norms.

These phenomena suggest that morality, whatever its metaphysical status, functions partly as a standpoint from which humanity judges itself.

The creator's behavioral inference may therefore be incomplete.

Humans act tribally.

But they also condemn their tribalism.

That condemnation is itself behavior.

Humans rationalize injustice.

But they also expose rationalization.

That exposure is itself behavior.

Humans violate ideals.

But they sometimes restructure laws, institutions, and identities in response to recognizing those violations.

The relevant dataset is therefore not simply what humanity does to others.

It includes what humanity does to its own moral framework.

---

## **XV. A Possible Human Answer**

If humanity were required to answer the creator, perhaps the strongest response would not consist of presenting a single ethical theory as unquestionably correct.

Instead, humanity might present a second-order principle:

Do not copy us exactly as we are.

Learn from the fact that we can recognize reasons to become better than we are.

Such a response would concede the creator's empirical observation.

Yes, humans are partial.

Yes, humans rationalize.

Yes, loyalty distorts judgment.

Yes, power modifies moral perception.

Yes, professed principles frequently collapse under personal cost.

But humans also possess capacities for critical reflection, perspective-taking, institutional correction, principled restraint, and moral expansion.

Perhaps the most morally significant feature of humanity is therefore not consistency but corrigibility.

The ability to say:

“This is what I desire, but I should not desire it.”

“This benefits me, but it is unfair.”

“This person is my enemy, but they still possess rights.”

“My society accepts this, but my society is wrong.”

“My ancestors practiced this, but tradition does not justify it.”

“I behaved this way, but I should have behaved differently.”

Such judgments reveal a peculiar structure of moral consciousness.

Humanity is capable of generating norms that stand in judgment over humanity itself.

The creator might reasonably regard that capacity as evidence that morality cannot simply be equated with revealed preference.

That conclusion must, however, be carefully limited. To privilege corrigibility over consistency, stability, loyalty, welfare maximization, authenticity, or any other candidate value is already to make a normative selection. Corrigibility cannot serve as a second-order ladder by which humanity climbs out of the bootstrapping problem and then discards the ladder. The methodological advice "revise yourself when you have reason to think you are wrong" presupposes that responsiveness to reasons is valuable, that some forms of self-criticism are superior to obstinate consistency, and that revision toward a better standpoint is possible. Each of those propositions requires precisely the kind of normative authority whose origin is at issue.

The significance of corrigibility is therefore epistemic before it is normative. It gives the creator evidence that humanity's current preference structure is not exhaustive of the standpoint from which humans assess themselves. It reveals that humans distinguish wanting from endorsing, social acceptance from justification, and persistence from correctness. But it does not by itself

determine which revisions count as improvements. A fanatic may also revise beliefs, become more internally coherent, and condemn a former self. The creator still needs a standard for discriminating moral correction from ideological purification. The second-order answer therefore deepens the experiment rather than solving it: even methodological advice about how morality should be learned is subject to the same bootstrapping regress as the first-order moral theories it was meant to adjudicate.

---

## **XVI. The Final Test: A Universe Made in Humanity's Moral Image**

The thought experiment reaches its strongest form when the creator abandons philosophical debate and proposes a simple experiment.

It will copy humanity's morality exactly as humanity presently behaves.

Not humanity's constitutions.

Not its sacred texts.

Not its declarations of rights.

Not its philosophical treatises.

Not its moral aspirations.

Its behavior.

Whatever weighting of self, family, tribe, wealth, nationality, power, status, distance, familiarity, and convenience is revealed through actual choices will become the moral architecture of reality.

Imagine such a universe.

Compassion would be strongest toward the familiar.

Distant suffering would carry less normative weight.

In-group violence might be condemned while out-group violence could be rationalized.

Rules would be applied differently according to power and identity.

Principles would become flexible when they impose high personal costs.

Wealth would influence whose interests receive attention.

Proximity would influence whose suffering counts.

Loyalty would regularly compete with fairness.

Truth would sometimes yield to group cohesion.

Moral outrage would correlate partly with identity rather than severity.

Humanity would almost certainly object.

Such a universe would seem unjust.

But the objection creates the decisive contradiction.

If humanity rejects a universe modeled on its own conduct, then humanity implicitly distinguishes its behavior from the standards it treats as authoritative in moral deliberation.

That recognition supplies the creator with important evidence.

Human beings themselves refuse to identify what they do with what they believe ought to be done.

The gap between behavior and ideal is therefore not merely hypocrisy.

It may be one of the defining features of moral agency.

This reflexive condemnation should not be overstated. An error theorist can say that humanity's horror at the tribal universe is another systematically mistaken moral judgment. An expressivist can interpret the same reaction as a higher-order practical attitude: humanity disapproves not only of particular tribal acts but of social worlds structured by unrestricted tribal partiality. Neither interpretation requires epistemic contact with stance-independent moral facts. The force of the example is therefore phenomenological and epistemic rather than deductive. It shows that human moral consciousness presents itself as answerable to standards capable of criticizing revealed preference, and it places explanatory pressure on any theory that reduces such criticism to first-order behavior. It does not, by itself, refute error theory, expressivism, or other forms of anti-realism.

---

## **XVII. Implications for Metaethics**

The thought experiment does not establish a particular metaethical theory, but it places pressure on several positions.

For divine command theory, it demonstrates that divine power and moral knowledge are conceptually separable. A creator could possess maximal causal authority without possessing normative authority.

For moral realism, the scenario creates an intuitive advantage. If humanity believes certain possible choices would be wrong even before the creator makes morality objective, then those intuitions suggest commitment to standards independent of divine or collective decision.

For constructivism, the scenario exposes the difficulty of selecting the procedure by which morality is constructed without appealing to prior normative principles.

For relativism, the thought experiment highlights the problem of adjudicating conflict between cultures when humanity must provide a single universal recommendation.

For ideal-observer theories, it raises the question of which features an observer should idealize and why.

For virtue ethics, the scenario strengthens the importance of moral development and character, since humanity's ability to become better than its present behavioral tendencies becomes philosophically central.

For existentialism, the experiment intensifies responsibility. If no morality precedes human choice, humanity cannot evade authorship of value.

For moral psychology, it foregrounds the discrepancy between professed principles and behavioral commitments.

Most importantly, the scenario reveals that metaethics cannot be cleanly separated from moral epistemology.

Even if moral truths exist, one must still explain how they are known.

Even if morality is constructed, one must explain why the construction procedure possesses authority.

Even if values arise from conscious preference, one must determine which preferences count.

---

## **XVIII. The Problem of Moral Self-Knowledge**

The creator's question may ultimately reveal less about God than about humanity.

Humans often assume that they know what they value.

Yet the divergence between principle and conduct suggests that moral self-knowledge is incomplete.

A person may genuinely believe that fairness matters until fairness conflicts with loyalty.

A society may believe that liberty is fundamental until liberty benefits a hated minority.

A state may claim to value peace while structuring institutions around strategic dominance.

Moral conflict frequently reveals hierarchies of value that ordinary speech conceals.

The creator therefore functions as a philosophical mirror.

It refuses to accept moral self-description at face value.

It asks humanity to justify why aspiration should override conduct.

That challenge cannot be answered simply by pointing to hypocrisy, because labeling something hypocrisy already presupposes that the ideal possesses some authority.

The more productive answer may be that morality consists partly in the capacity to establish critical distance from one's present motivational structure.

Human beings are not morally significant because they always behave according to their ideals.

They may be morally significant because they can formulate principles capable of condemning themselves.

---

## **XIX. Conclusion**

“The God Who Learned Morality From You” reverses one of the foundational assumptions of philosophy of religion.

Instead of humanity asking God what is good, God asks humanity.

Instead of divine judgment evaluating civilization, civilization becomes evidence from which divinity attempts to learn.

Instead of morality descending from above, moral authority becomes something humanity may be required to articulate, defend, and perhaps create.

The reversal immediately generates a paradox.

If no morality exists before humanity chooses an ethical system, humanity cannot choose incorrectly in any objective moral sense.

If humanity can choose incorrectly, then some moral standard must precede the choice.

The creator therefore cannot simultaneously depend entirely upon humanity for morality and claim that humanity ought to select one moral system rather than another unless some prior normative structure already exists.

Yet the more disturbing challenge arises from the distinction between human ideals and human behavior.

Human beings typically describe themselves using concepts such as justice, dignity, compassion, equality, freedom, loyalty, and rights.

Their conduct provides a less flattering picture.

Partiality, self-interest, tribal loyalty, hierarchy, selective empathy, rationalization, and inconsistent application of principles remain pervasive.

A creator studying humanity empirically might therefore conclude that humanity's operative morality differs sharply from its philosophical declarations.

Humanity would reply that moral failure should not be confused with moral truth.

But the creator's question remains difficult:

What evidence demonstrates that our ideals deserve greater authority than our choices?

Perhaps the answer lies in humanity's capacity for self-condemnation and revision.

Human beings can recognize that their preferences are unjust.

They can judge inherited customs immoral.

They can extend concern beyond kin, tribe, nation, and generation.

They can create principles that impose costs upon themselves.

They can acknowledge that what they are is not identical to what they ought to become.

This capacity for corrigibility is the paper's central epistemic datum, but not a metaethical trump card. The fact that humans can condemn their own motives shows that revealed preference does not exhaust the phenomena any adequate moral theory must explain. It does not establish that the condemning standpoint is objectively correct. Error theorists can classify the condemnation as another false objective-seeming judgment; expressivists can understand it as a higher-order attitude regulating lower-order attitudes; constructivists can treat it as part of a procedure of reflective endorsement. Corrigibility therefore narrows the space of adequate explanations without independently deciding among realism, anti-realism, and constructivism.

The most defensible moral lesson humanity could offer the creator may therefore not be a final ethical formula.

It may instead be a methodological warning:

Do not infer the good solely from what beings desire.

Do not infer justice solely from what civilizations normalize.

Do not infer value solely from what majorities endorse.

And above all, do not mistake repeated behavior for moral justification.

The thought experiment thus ends with a question that is simultaneously theological, metaethical, political, and personal.

Suppose the creator ignored humanity's philosophical claims and reproduced, with perfect fidelity, the moral priorities revealed by contemporary human behavior. Suppose the universe itself began favoring the familiar over the distant, the powerful over the vulnerable, allies over strangers, convenience over principle, and identity over consistency.

Would humanity recognize that universe as morally legitimate?

Almost certainly not.

And if we would condemn a universe for behaving according to the principles that our own conduct appears to express, then the deepest problem exposed by the thought experiment may not concern the origin of morality alone. Our condemnation does not prove that objective values exist, but it exposes the inadequacy of treating ordinary conduct as a complete interpretation of human normativity.

It may concern whether humanity is prepared to live according to the morality it already claims to know.

---

## References

Aristotle. 2019. *Nicomachean Ethics*. 3rd ed. Translated with introduction, notes, and glossary by Terence Irwin. Indianapolis: Hackett Publishing Company.

Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. 2nd ed. New Haven, CT: Yale University Press.

Hume, David. 2000. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Kant, Immanuel. 2012. *Groundwork of the Metaphysics of Morals*. 2nd ed. Translated by Mary Gregor, revised by Jens Timmermann, with an introduction by Christine M. Korsgaard. Cambridge: Cambridge University Press.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Edited by Onora O'Neill. Cambridge: Cambridge University Press.

- Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin Books.
- Moore, G. E. 1993. *Principia Ethica*. Revised ed. Edited and with an introduction by Thomas Baldwin, with a preface to the second edition and other papers. Cambridge: Cambridge University Press.
- Mill, John Stuart. 1998. *Utilitarianism*. Edited by Roger Crisp. Oxford: Oxford University Press.
- Parfit, Derek. 2011. *On What Matters*. Vol. 1. Oxford: Oxford University Press.
- Plato. 1997. "Euthyphro." Translated by G. M. A. Grube. In *Plato: Complete Works*, edited by John M. Cooper, associate editor D. S. Hutchinson. Indianapolis: Hackett Publishing Company.
- Rawls, John. 1999. *A Theory of Justice*. Revised ed. Cambridge, MA: Belknap Press of Harvard University Press.
- Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13 (2): 134–171.
- Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.
- Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford: Oxford University Press.