Axiomatic Underdetermination in the Free-Will Debate: An Independence Result

August 2026

Abstract

The classical dispute between compatibilism and incompatibilism about free will and moral responsibility is often described informally as an "impasse." This paper gives that impasse a precise formal shape. We define a minimal base theory T_0, in propositional S5 modal logic, capturing only what compatibilists and incompatibilists already share: a notion of determinism that fixes outcomes given the past, and the bare vocabulary of action, reasons-responsiveness, and moral responsibility, with no bridge principle connecting them. We then prove, by exhibiting explicit finite Kripke models, that T_0 together with the further fact that the actual world is deterministic is independent of both the standard compatibilist bridge principle and the standard incompatibilist alternative-possibilities requirement: each is separately consistent with the shared base, and neither is derivable from it. This is structurally analogous to the independence of the Continuum Hypothesis from the axioms of ZFC, and is established the same way, by exhibiting models rather than by appeal to informal intuition. We then isolate three further, narrower results that are sometimes conflated with this central independence claim but are logically distinct from it and from one another: a provable impossibility theorem for self-disclosing predictors, obtained by direct diagonalization; a scope restriction on what Gödel's second incompleteness theorem can license a formal system to claim about itself; and a use of the Anti-Foundation Axiom's Solution Lemma showing that an infinite, beginningless regress of character-formation is formally coherent as a structure, without thereby securing the further, substantive notion of ultimate authorship that the incompatibilist regress objection actually needs.

Contents

1 Introduction

The compatibility question — whether moral responsibility can survive the truth of causal determinism — is standardly posed as a disagreement over a single conditional: does determinism, or the falsity of the ability to do otherwise, undermine moral responsibility? Compatibilists say no, provided some weaker condition (typically reasons-responsiveness, or a suitably reflective relation to one's own motivational states) is met. Incompatibilists say yes, holding that moral responsibility requires access to genuine alternative possibilities, or, in the strongest "ultimate responsibility" versions of the view, requires the agent to be the ultimate causal source of the character from which the choice issued.

A recurring temptation in formalizing this debate is to reach for the heaviest available logical machinery — incompleteness theorems, computability theory, non-well-founded set theory — in the hope that some deep result forces a verdict, or at least forces the conclusion that no verdict is possible. This temptation should be resisted, because it invites two easily separable errors: treating the derivation of a contradiction from jointly incompatible premises as if it were a discovery about the subject matter rather than an artifact of the premises chosen, and treating *inconsistency* (a theory that proves everything) as though it were the same phenomenon as *undecidability* (a question for which no algorithm settles the answer), when in fact these are different failure modes and typically arise, if at all, from different sources.

This paper takes the more disciplined route. Section 2 sets out a base theory T_0 containing only what both sides of the debate already accept. Section 3 proves the paper's central result: T_0, together with the fact of determinism, is independent of the competing bridge principles, established by exhibiting three explicit finite models rather than by analogy. Sections 4–7 then examine, one at a time and with appropriate scope restrictions, the three pieces of heavier machinery gestured at above — diagonalization/computability, Gödel's second incompleteness theorem, and non-well-founded sets together with the Axiom of Choice — identifying exactly what each does and does not establish about the debate. Section 8 draws these together.

2 A minimal shared base theory

We work in propositional S5 modal logic, the standard choice for metaphysical (as opposed to temporal) necessity and possibility evaluated relative to a fixed past, interpreted over Kripke models $\mathcal{M} = (W, R, V)$ where W is a non-empty set of possible worlds, $R = W \times W$ (S5's universal accessibility relation), and V assigns each atomic proposition a subset of W at which it is true.

Definition 2.1 (Vocabulary)**.** *Atomic propositions: d (the actual world is deterministic), f (agent a performs action p), m (agent a is morally responsible, in the basic-desert sense, for p), rr (agent a is reasons-responsive with respect to p). Define $AP := \Diamond f \wedge \Diamond \neg f$ (alternative possibilities).*

Axiom 2.2 (Base theory T_0)**.** *T_0 consists of the S5 axiom schemata (K, T, 4, 5, and Necessitation) together with:*

(T0-1) Fixity under determinism*: $d \to \big((f \to \Box f) \wedge (\neg f \to \Box \neg f)\big)$. If the world is deterministic, then however f actually comes out, it comes out that way of necessity, relative to the fixed past and laws.*

(T0-2) Contingency of determinism*: across the class of models under consideration, both d and $\neg d$ are separately satisfiable. Whether the actual world is deterministic is not settled by logic alone.*

T_0 contains no bridge principle connecting m to d, f, or AP. This is the honest starting point: everyone in the debate accepts something like (T0-1), and no one thinks logic alone settles whether determinism actually obtains; the debate is about what else to add on top of this shared core.

Axiom 2.3 (Compatibilist bridge, C)**.** $d \wedge rr \to m$. *If the world is deterministic and the agent is reasons-responsive, the agent is morally responsible, with no further requirement of alternative possibilities.*

Axiom 2.4 (Incompatibilist bridge, I)**.** $m \to AP$. *Moral responsibility requires that the agent could have done otherwise.*

The question this paper answers precisely is: does T_0, together with the further fact that determinism is actually true (d), settle whether C or I holds? We show it does not.

3 The independence theorem

Theorem 3.1 (Independence of the bridge principles)**.** $T_0 \cup \{d\}$ *is consistent with each of the following, separately:*

(a) $C \wedge m \wedge \neg AP$ *(compatibilism, with m actually holding and no alternative possibilities);*

(b) $I \wedge \neg m$ *(hard determinism: the incompatibilist bridge holds, but is satisfied vacuously because m is false).*

Moreover T_0 alone (without d) is consistent with:

(c) $I \wedge m \wedge AP \wedge \neg d$ *(libertarianism: the incompatibilist bridge is satisfied non-vacuously under indeterminism).*

Consequently, $T_0 \cup \{d\} \nvDash C$, $T_0 \cup \{d\} \nvDash \neg C$, $T_0 \cup \{d\} \nvDash I$, $T_0 \cup \{d\} \nvDash \neg I$, $T_0 \cup \{d\} \nvDash m$, *and* $T_0 \cup \{d\} \nvDash \neg m$.

Proof. We exhibit finite Kripke models and check satisfaction directly. In S5 with $R = W \times W$, for any W and any world w, $\mathcal{M}, w \models \Diamond\varphi$ iff φ is true at *some* world in W, and $\mathcal{M}, w \models \Box\varphi$ iff φ is true at *every* world in W; truth of boxed/diamonded formulas does not depend on which world w we evaluate at, only on W as a whole. We therefore describe each model simply by its worlds and valuation.

(a) A compatibilist model $\mathcal{M}_C$. Let $W = \{w_1\}$, a single world, with d, f, m, rr all true at w_1. In a one-world S5 model, $\Diamond\varphi \leftrightarrow \varphi$ and $\Box\varphi \leftrightarrow \varphi$ for every φ, since the unique world is the only candidate witness.

Check (T0-1): f is true at w_1, so $\Box f$ is true (trivially, as the only world), so $f \to \Box f$ holds; the $\neg f$ disjunct is vacuous since f is true.

Compute AP: $\Diamond f$ is true (witnessed by w_1), $\Diamond\neg f$ is false (no world makes $\neg f$ true). So AP is false.

Check C: $d \wedge rr \to m = T \wedge T \to T =$ true.

Check I fails: $m \to AP = T \to F =$ false.

So $\mathcal{M}_C \models T_0 \cup \{d, C, m, \neg AP\}$ and $\mathcal{M}_C \not\models I$, proving (a) and showing C and I are not both forced by a single model.

(b) A hard-determinist model $\mathcal{M}_{HD}$. Let $W = \{w_1\}$ with d, f, rr true at w_1, but m *false.* As above, AP is false.

Check (T0-1): holds as in (a), since it does not mention m.

Check I: $m \to AP = F \to F =$ true, vacuously.

Check C fails: $d \wedge rr \to m = T \wedge T \to F =$ false.

So $\mathcal{M}_{HD} \models T_0 \cup \{d, I, \neg m\}$ and $\mathcal{M}_{HD} \not\models C$, proving (b).

(c) A libertarian model $\mathcal{M}_L$. Let $W = \{w_1, w_2\}$ with f true at w_1 and false at w_2; d false at both worlds; m true at both worlds.

Check (T0-1): the antecedent d is false at both worlds, so the conditional is vacuously true throughout.

Compute AP: $\Diamond f$ is true (witnessed by w_1); $\Diamond \neg f$ is true (witnessed by w_2). So AP is true at both worlds.

Check I: $m \to AP = T \to T =$ true at both worlds.

So $\mathcal{M}_L \models T_0 \cup \{I, m, AP, \neg d\}$, proving (c).

Together, $\mathcal{M}_C$ and $\mathcal{M}_{HD}$ show that fixing d true within T_0 leaves C, I, and m each independently satisfiable and falsifiable: $\mathcal{M}_C$ has C true, I false, m true; $\mathcal{M}_{HD}$ has C false, I true, m false. By the soundness and completeness of S5: if a formula were derivable from a consistent set of premises, it would be true in every model of those premises; since we have exhibited models of $T_0 \cup \{d\}$ disagreeing on C, I, and m, none of these is a consequence of $T_0 \cup \{d\}$. □ □

Remark 3.2. This is a deliberately modest result, and its modesty is the point. It does not show the free-will debate is meaningless, nor that all positions are equally warranted all-things-considered; it shows that the dispute cannot be settled by the shared logical and definitional apparatus in T_0 alone, plus the empirical fact of determinism. Settling it requires an additional, non-logical commitment — to C, to I, or to some other bridge principle — defended on independent metaphysical or normative grounds. This is structurally analogous to the independence of the Continuum Hypothesis from the axioms of ZFC [2, 3]: exhibiting models of ZFC in which CH holds and models in which it fails does not show set theory is broken; it shows CH is not decided by the axioms as given, and that adopting CH or ¬CH is a further, substantive choice. Theorem 3.1 is the free-will debate's counterpart, established the same way: by exhibiting models.

4 A provable impossibility result: the predictor paradox

Independence, in the sense of Theorem 3.1, is not the only formal phenomenon relevant to this debate. It is worth isolating one place where a genuine impossibility result — not an independence result, but an outright theorem — is available, obtainable by ordinary diagonalization.

Definition 4.1. *A* disclosing predictor *for agent a is a total computable function* pred *that takes the complete past state s (including, if the agent is itself computable, a description of the agent's decision procedure) and outputs a prediction* $\text{pred}(s) \in \{f, \neg f\}$ *of the agent's action, where this prediction is disclosed to the agent* before *the agent acts, and where the agent's decision procedure* $\text{Agent}(s)$ *has read access to* $\text{pred}(s)$.

Theorem 4.2 (Impossibility of a sound contrarian-proof disclosing predictor)**.** *There is no disclosing predictor* pred *that is sound against every possible computable agent, in the sense that*

$\mathrm{pred}(s) = f \iff \mathrm{Agent}(s) = f$ *for all* s*, if the class of possible agents includes a* contrarian *agent defined by* $\mathrm{Agent}^*(s) := \neg\,\mathrm{pred}(s)$*.*

Proof. This is a direct diagonalization, structurally identical to Cantor's diagonal argument, Turing's proof of the undecidability of the halting problem [5], and the self-referential sentence used in Gödel's first incompleteness theorem — all instances of the same fixed-point schema. Suppose such a sound pred exists. Since Agent^* is a legitimate computable agent (it merely reads $\mathrm{pred}(s)$, which is disclosed to it, and outputs the opposite), soundness requires

$$\mathrm{pred}(s) = f \iff \mathrm{Agent}^*(s) = f \iff \neg\,\mathrm{pred}(s) = f.$$

For the two-valued domain $\{f, \neg f\}$, the right-hand biconditional says $\mathrm{pred}(s) = f$ iff $\mathrm{pred}(s) \neq f$, a contradiction. So no such sound, disclosing, universally-applicable pred exists. □ □

Remark 4.3. This result is appropriately narrow: it says nothing about whether $F(a, p)$ is fixed by the past — that is what determinism, if true, already settles per (T0-1) — and it says only that a specific kind of self-referential, disclosure-obligated predictor cannot exist as a sound universal oracle. This is the precise content behind the intuition that a fully deterministic agent's choice can still fail to be something a predictor can always disclose in advance without thereby falsifying itself. Determinism and unpredictability-by-disclosure are not in tension; the theorem isolates exactly where genuine, provable unpredictability enters the picture, and it is considerably narrower than the free-will question itself, related instead to the structure of Newcomb-style prediction problems [12].

5 Scope of Gödel's second incompleteness theorem

Proposition 5.1. *Let* $\Sigma \supseteq T_0$ *be any formal system strong enough to interpret elementary arithmetic (needed for Gödel numbering of* Σ*'s own proofs), and suppose* Σ *additionally asserts* $\mathrm{Con}(\Sigma)$ *as part of certifying itself a complete and self-verifying arbiter of the compatibility question. Then* Σ *is inconsistent.*

Proof. This is Gödel's second incompleteness theorem [1] applied directly: for any consistent, recursively axiomatized formal system Σ interpreting arithmetic, $\Sigma \nvdash \mathrm{Con}(\Sigma)$. If Σ nonetheless includes $\mathrm{Con}(\Sigma)$ as a theorem, Σ cannot be consistent. □ □

Remark 5.2. The proposition is a scope restriction, not a route to the paper's main result. It says: any formal system ambitious enough to arithmetize its own reasoning and certify its own consistency as part of its verdict on m is thereby inconsistent, so no such self-certifying Σ should be trusted. It does not bear on whether T_0, $T_0 \cup \{C\}$, or $T_0 \cup \{I\}$ is consistent; Theorem 3.1 already established each directly, by finite models, which is a strictly stronger and more elementary form of consistency proof than anything Gödel's theorem could supply here. The proposition's role is deflationary: it rules out a particular kind of overreach, a formal system that tries to be the last word on the question by fiat, and it should not be read as itself generating the underdetermination shown in Section 3.

6 Non-well-founded sets and the regress argument

The incompatibilist "ultimate responsibility" objection, in the tradition of Galen Strawson's regress argument [6], holds roughly: to be responsible for choosing p, the agent must be responsible for the character or motivational state C_0 that produced the choice; to be responsible for C_0, the agent must be responsible for the prior state C_{-1} that produced it; and so on without end, so that ultimate responsibility requires a completed infinite regress with no first term.

Under the standard, well-founded conception of sets governed by the Axiom of Foundation, an infinite descending membership chain is barred outright, which might seem to hand the incompatibilist a quick formal victory: the regress structure is set-theoretically impossible. Non-well-founded set theory corrects this.

Proposition 6.1 (Solution Lemma, Aczel [4])**.** *Under ZFA (ZF with Foundation replaced by the Anti-Foundation Axiom), every flat system of equations of the form $x_i = \varphi(x_i, \{x_j\}_{j\in J_i})$, for a family of indeterminates indexed over a set I, has a unique solution.*

Model the character-formation regress as the flat system $C_n = \varphi(C_{n-1})$ for n ranging over $\{0, -1, -2, \dots\}$ with no least element, where φ is the character-formation function. By the Solution Lemma, this system has a unique solution in ZFA: a well-defined, coherent assignment of a character-state to every index, forming a non-well-founded stream, with no contradiction and no need for a first term.

Corollary 6.2. *The bare existence of an infinite, beginningless regress of character-forming states is not, by itself, formally incoherent; ZFA supplies a consistent model of exactly that structure.*

Remark 6.3. Corollary 6.2 establishes that the regress is a coherent *structure*; it does not establish that occupying a position within such a structure constitutes *ultimate authorship* of one's choices, in whatever sense a strengthened incompatibilist bridge principle might demand. Formal coherence of the regress and satisfaction of a substantive authorship condition are different claims. The incompatibilist cannot simply assert that the regress is impossible full stop (ZFA blocks that move), but the reverse inference — that a coherent regress secures authorship — is equally unavailable (Corollary 6.2 is silent on authorship). What remains is a genuine, further premise connecting causal history to authorship, which is where the substantive philosophical work in this part of the debate actually lies.

7 Scope of the Axiom of Choice

In an indeterministic branching model with infinitely many worlds — for instance, a model where, at each of infinitely many choice points, both f and $\neg f$ are realized on different branches — selecting one representative branch per choice point, to define a single "history" function, requires the Axiom of Choice only when the index set of choice points is infinite *and* no uniform, definable rule for the selection is available (such as "always take the lexicographically least branch"). Where a definable well-ordering or rule exists, AC is dispensable. AC is thus relevant to constructing certain infinite selection functions over branching futures in indeterministic models, but a choice function obtained this way is precisely non-constructive: it need not correspond to anything causally produced by prior states, and its existence gives no reason to think the selection reduces to, or is fixed by, the deterministic past.

8 Summary of results

(1) A minimal shared base theory T_0, together with the fact of determinism, does not entail or refute either the compatibilist or the incompatibilist bridge principle (Theorem 3.1), established by exhibiting explicit finite models.

(2) A genuine, narrow impossibility result exists concerning self-disclosing predictors, provable by ordinary diagonalization (Theorem 4.2), and it is logically independent of the compatibility question itself.

(3) Gödel's second incompleteness theorem rules out any formal system that tries to certify itself as the final, self-verifying arbiter of the question, but does not bear on T_0's own consistency, which is established directly by finite models (Proposition 5.1).

(4) The Anti-Foundation Axiom shows that an infinite, beginningless regress of character-formation is formally coherent as a structure, which blocks a quick incompatibilist regress objection but does not itself secure the further, substantive notion of ultimate authorship (Proposition 6.1, Corollary 6.2).

(5) The Axiom of Choice is relevant to constructing non-canonical selection functions over infinite branching futures, but such a selection function is precisely not reducible to, or fixed by, deterministic history (Section 7).

9 Conclusion

The compatibility question is best understood, formally, as a genuine independence phenomenon: nothing in the minimal apparatus that compatibilists and incompatibilists already share — together with the bare fact that determinism obtains, if it does — forces a verdict on whether moral responsibility survives determinism. Adjudicating the question requires adopting a further bridge principle on independent grounds, which is exactly what the substantive debate between compatibilists, hard determinists, and libertarians has always concerned. This independence result is distinct from, and should not be conflated with, the narrower and logically separate results available nearby: a provable impossibility theorem for self-disclosing predictors, a scope restriction on what self-certifying formal systems can claim, and a demonstration that infinite regresses of character-formation are formally coherent without thereby settling questions of ultimate authorship. Formal logic does not resolve the compatibility question; it clarifies precisely that the choice between bridge principles is not forced, and it identifies the small number of places where heavier logical machinery does genuine, checkable, if narrower, work.

References

[1] Gödel, Kurt. 1931. "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I." *Monatshefte für Mathematik und Physik* 38: 173–198.

[2] Gödel, Kurt. 1940. *The Consistency of the Continuum Hypothesis.* Princeton: Princeton University Press.

[3] Cohen, Paul J. 1963. "The Independence of the Continuum Hypothesis." *Proceedings of the National Academy of Sciences* 50 (6): 1143–1148.

[4] Aczel, Peter. 1988. *Non-Well-Founded Sets.* CSLI Lecture Notes 14. Stanford: CSLI Publications.

[5] Turing, Alan M. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 2 (42): 230–265.

[6] Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75: 5–24.

[7] Fischer, John Martin. 1994. *The Metaphysics of Free Will: An Essay on Control.* Oxford: Blackwell.

[8] Kane, Robert. 1996. *The Significance of Free Will.* Oxford: Oxford University Press.

[9] Pereboom, Derk. 2001. *Living Without Free Will.* Cambridge: Cambridge University Press.

[10] van Inwagen, Peter. 1983. *An Essay on Free Will.* Oxford: Clarendon Press.

[11] Mele, Alfred. 2008. "Manipulation, Compatibilism, and Moral Responsibility." *The Journal of Ethics* 12: 263–286.

[12] Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher, 114–146. Dordrecht: Reidel.