Philosophical Studies (2023) 180:2291–2307
https://doi.org/10.1007/s11098-023-01979-7

A puzzle about moral responsibility

Fabio Lampert[1] · **John William Waldrop**[2]

Accepted: 25 April 2023 / Published online: 22 May 2023

Abstract
We present a new puzzle about logical truth, necessity, and moral responsibility. We defend one solution to the puzzle. A corollary of our preferred solution is that prominent arguments for the incompatibility of determinism and moral responsibility are invalid.

Keywords Moral responsibility · Determinism · Logical truth · Necessity · Incompatibilism

1 A puzzle

Here are three initially plausible thoughts. First, no one is even partly morally responsible for logical truths; second, the truths for which no one is (even partly) morally responsible are closed under entailment; and third, there is at least one truth for which someone is (at least partly) morally responsible. Here's a puzzle: you can't consistently have all three.

To see why, let Nα abbreviate 'α and no one is even partly morally responsible for the fact that α', and let the operator $\Box$ express (metaphysical) necessity. The three initially plausible thoughts above can now be regimented thus:

(A) $\alpha \vdash \mathrm{N}\alpha$, if α is a logical truth
(B) $\mathrm{N}\alpha, \Box(\alpha \rightarrow \beta) \vdash \mathrm{N}\beta$
(C) $\exists p(p \,\&\, \neg \mathrm{N}p)$

We include the operator @ (read: 'actually') in our modal language, and we assume standard principles of the modal logic governing $\Box$ and @—hereafter,

✉ Fabio Lampert
fabio.lampert@uni-greifswald.de

John William Waldrop
jwaldrop@nd.edu

[1] Universität Greifswald, Greifswald, Germany

[2] University of Notre Dame, Notre Dame, USA

'standard modal logic'.[1] The following argument establishes that (A)-(C) generate a contradiction:

(1) $p \,\&\, \neg \mathrm{N}p$
(2) $\mathrm{N}(@p \rightarrow p)$
(3) $\Box((@p \rightarrow p) \rightarrow p)$

---

(4) $\mathrm{N}p \,\&\, \neg \mathrm{N}p$

(1) comes from (C) by quantifier elimination. Since $@p \rightarrow p$ is a logical truth according to standard modal logic, (2) follows from (A); since $p \rightarrow \Box((@p \rightarrow p) \rightarrow p)$ is a logical truth according to standard modal logic,[2] (3) follows from (1). (For more on these points, see the following section.) Finally, in (4), $\mathrm{N}p$ follows from (2) and (3) given (B), and $\neg \mathrm{N}p$ follows from (1). The argument is valid.

The puzzle can be stated alternatively, without quantifying into sentence position in (C), as the startling claim that (A) and (B) together force the operator N to collapse in the following sense:

COLLAPSE: $\alpha \dashv\vdash \mathrm{N}\alpha$

In other words, nobody is morally responsible for anything. The right-to-left side of COLLAPSE holds by design, since N is factive. The converse derivation from (A) and (B) is implicit in our argument above: assuming p, (2) and (3) follow as before, and so $\mathrm{N}p$ follows by (B). Therefore, given (A) and (B), no one is morally responsible for any truth. Since people *are* sometimes morally responsible for things, COLLAPSE is totally unacceptable. We assume as much in what follows. So, in evaluating the puzzle generated from (A), (B), and (C) above, we assume that abandoning (C) is out of the question. The puzzle on display here, then, forces a choice between two antecedently plausible principles, (A) and (B); at least one of them must be abandoned.

2 The status of standard modal logic

Before moving on to evaluating the contest between (A) and (B), let us consider what is perhaps a tempting response to this puzzle. The puzzle results not simply from accepting the principles (A), (B), and (C), but also from accepting what we have here called standard modal logic. One resolution of the puzzle—which strikes

[1] The natural deduction system in Hazen (1978) is an example of a system of standard modal logic.

[2] This is provable in the aforementioned system developed in Hazen (1978), and in any similar system for which logical validity is defined as *truth in the actual world in every model* for the modal language of $\Box$ and @, which is what Crossley and Humberstone (1977) call 'real-world validity'. This is also the notion of validity featuring in Kripke's (1959, 1963) seminal work in modal semantics, albeit for languages not containing the operator @.

us as drastic—involves rejecting "enough of" standard modal logic so that no contradiction results from accepting (A), (B), and (C). Though in this essay we are officially assuming standard modal logic, let us briefly give some motivation for holding on to it in light of the present puzzle. For the sake of space, we will focus only on the least familiar part of standard modal logic, namely, what it mandates about the behavior of the actuality operator @.

In defending steps (2) and (3) of the argument above, it suffices to note that standard modal logic validates the following principles:

(i) $p \leftrightarrow @p$
(ii) $@p \rightarrow \Box @p$

Justifying (2) given (A) requires only the left-to-right direction of (i). Our justification for (3) given (1) appealed to the fact that $p \rightarrow \Box((@p \rightarrow p) \rightarrow p)$ is a logical truth of standard modal logic. This can be shown by appeal to the right-to-left direction of (i) as well as (ii), given some uncontroversial principles governing $\Box$,[3] the overturning of which, recall, we are not considering here.

It seems clear to us that abandoning either (i) or (ii) in response to this puzzle is not a promising way forward. For one thing, the standard behavior of the actuality operator codified in (i) and (ii) is assumed in many contexts throughout philosophy, as well as in contemporary logic and formal semantics. This fruitful understanding of the actuality operator also does theoretical work in many other areas.[4] By contrast, once the *prima facie* plausibility of principles like (A) and (B) is accounted for, they have no wider theoretical importance to recommend them.[5] The puzzle discussed in this paper can be addressed by simply rejecting some initially attractive principles about moral responsibility, and doing so appears to be a theoretical free lunch. Given this, the alternative option—modifying standard modal logic and thereby hamstringing fruitful programs and theories in other domains—strikes us as an immoderate response to the puzzle.

Finally, we note two ways in which rejecting standard modal logic simply postpones addressing the kind of problem our puzzle brings out. First, the inconsistency

[3] To see this, note that an instance of the K axiom for $\Box$ is the following conditional:

$$\Box(@p \rightarrow ((@p \rightarrow p) \rightarrow p)) \rightarrow (\Box @p \rightarrow \Box((@p \rightarrow p) \rightarrow p))$$

We then apply necessitation to the tautology $@p \rightarrow ((@p \rightarrow p) \rightarrow p)$, which together with the above conditional allows us to derive $\Box @p \rightarrow \Box((@p \rightarrow p) \rightarrow p)$ by modus ponens. (Note that even though in standard modal logic necessitation cannot be applied with full generality over more than the @-free fragment of our language, this particular application of necessitation is licit, since what is necessitated is a tautology.) From this, $p \rightarrow \Box((@p \rightarrow p) \rightarrow p)$ follows given the right-to-left direction of (i) together with (ii) (that is, given $p \rightarrow @p$ and $@p \rightarrow \Box @p$).

[4] Besides the other works cited in this paper (for instance, see footnote 10), we mention the logical developments in Davies and Humberstone (1980) distinguishing different notions of necessity, shedding light on the view of the contingent *a priori* developed in Evans (1979), and on the distinction between indicative and subjunctive conditionals defended in Weatherson (2001); consider also the investigation into formal languages containing implicit versus explicit parameters carried out in Köpping and Zimmermann (2018).

[5] The one exception we know of concerns appeals to (B) in arguments for the incompatibilty of physical determinism and moral responsibility, which we discuss in §7 below.

argument from (A)-(C) does not come with a prescription about the intended interpretation of the actuality operator @. Instead, all we need is the existence of (as it were) a property of propositions the operator associated with which obeys, as a matter of logic, the analogues of (i) and (ii). As it happens, formal arguments for the existence of such a property of propositions are available.[6] Second, even if no modal operator in our language obeys the analogues of (i) and (ii) as a matter of logic, similar puzzles arise given only a modest modal logic for the necessity operator $\Box$. For example, let T be a set collecting all and only the true propositions.[7] Where '$[p]$' denotes the proposition that p, the following two principles are plausible:

(T-i) $p \leftrightarrow [p] \in \mathrm{T}$
(T-ii) $[p] \in \mathrm{T} \rightarrow \Box[p] \in \mathrm{T}$

Since only particular instances of (T-i) are ever needed to generate the puzzle, we are content to recommend (T-i) as obvious enough; (T-ii) follows from the rigidity of set membership. If we assume that for some p, both $p \wedge \neg \mathrm{N}p$ and $\mathrm{N}([p] \in \mathrm{T} \rightarrow p)$, then modal logics considerably weaker than standard modal logic allow us to derive a contradiction given (B), in a way that parallels the inconsistency argument from (A)-(C) above.

In our judgment, though the argument we present in the previous section depends on standard modal logic, the *kind* of puzzle we are interested in arises more generally. To address the puzzle, our scrutiny is best directed towards the principles about moral responsibility embodied in (A) and (B).[8]

3 Responding to the puzzle

Principles similar to (B) are much debated in the literature on free will and moral responsibility, though surprisingly little has been written on principles like (A). As the puzzle above brings out, this neglect is unwarranted: together with (B), (A) delivers the result that no one is morally responsible for anything.

Despite its wider coverage relative to (A), the support available for (B) is modest. On the one hand, (B) is occasionally defended on intuitive grounds alone. Warfield (1996), for example, says this much on its behalf:

> I have no argument for the validity of [(B)]. I can think of no example which demonstrates that it is not valid nor have I found anyone who can produce such an example. (1996: 216)

[6] For arguments of this kind, see the discussion in Yli-Vakkuri and Hawthorne (2021, §3.2); see also Goodsell and Yli-Vakkuri (forthcoming).

[7] Some views about the granularity of propositions generate cardinality worries for the existence of such a set (cf. Grim (1984)); for our immediate purposes, propositions can be identified with sets of possible worlds, or sets of centered worlds, or some other appropriate coarsening of the notion of a proposition, which suffices to avoid such cardinality worries. See also Uzquiano (2015) and Lampert and Merlussi (2021a) for discussion.

[8] Thanks to an anonymous reviewer for pressing us to consider some of these issues.

On the other hand, (B) follows from two closure principles widely endorsed for the N operator:

NECESSITY: $\Box\alpha \vdash \mathrm{N}\alpha$
DISTRIBUTION: $\mathrm{N}\alpha, \mathrm{N}(\alpha \rightarrow \beta) \vdash \mathrm{N}\beta$[9]

These two principles form the logical spine of a standard and popular argument against the compatibility of moral responsibility and determinism (more on this in §7 below). So there is something to be said on behalf of (B). Nonetheless, in a competition between (A) and (B), we think that (A) is the clear winner.

(A) says that no one is even partly morally responsible for any logical truth. This is at first blush very plausible—no less plausible than (B), we take it. It is initially doubtful that one could ever be morally responsible for logical truths. If something is true as a matter of logic alone, then it is true regardless of what we do. Logical truths are, in some good sense of the phrase, truths that are "true no matter what". This is not to say that all logical truths are necessary—some, like those we are interested in here, are contingent,[10] It is a consequence of standard modal logic that if there are any contingent truths, some logical truths will be among them. It is still nonetheless plausible that no one could be *morally responsible* for a logical truth, even a contingent one.

At this point, an important caveat about what counts as *logic* is in order. For some extended uses of the term 'logic', it is perhaps plausible that one could be morally responsible for some logical truths, so-called. If our logical vocabulary includes the indexical 'I', for example, as in Kaplan (1989) logic of demonstratives, then non-defective occurrences of 'I exist' or 'I am here now' may count as logical truths,[11] and these may well express truths for which one could be (morally) responsible. But we will set aside these grounds for challenging (A), since abandoning (A) on these terms does not motivate a principled response to the puzzle presented above. If standard modal logic is part of logic, then (A) and (B) together generate the puzzle; but so long as nobody is ever morally responsible for the truths of *standard modal*

[9] Assume $\mathrm{N}\alpha$ and $\Box(\alpha \rightarrow \beta)$. $\mathrm{N}(\alpha \rightarrow \beta)$ follows from the second assumption by NECESSITY and $\mathrm{N}\beta$ follows from this and the first assumption by DISTRIBUTION. Principles such as NECESSITY and DISTRIBUTION are suggested in van Inwagen (1983: 184).

[10] That there are contingent logical truths is endorsed widely, both explicitly and implicitly. For instance, see Hazen (1978), Zalta (1988), Kaplan (1989), Williamson (2007: 64–65), Nelson and Zalta (2012), and Salmon (2019: 653). For some dissenting views, see Crossley and Humberstone (1977), and also in Hanson (2006, 2014).

[11] A different but related issue has to do with whether logic, for the standard interpretation of the language of standard modal logic, is given by what Crossley and Humberstone (1977) call "global validity" as opposed to what they call "real-world validity". The challenge does not have to do with whether (A) should be accepted, but with whether truths such as $@p \rightarrow p$ are *logical* truths. What we say about the challenge posed by Kaplan's liberal conception of logic in what follows applies just as well to these challenge posed by more conservative conceptions of logic.

logic, the upshot of the puzzle remains, regardless of the general status of (A).[12] Arguments against (A) from considerations about the extent of logic, in other words, miss the point. For these reasons, we will continue to assume that logic includes standard modal logic, and we will also stipulatively understand (A) as the claim that no one ever is morally responsible for the truths of standard modal logic.

There is something to be said too for the related claim that no one is morally responsible for *a priori* truths. The unrestricted version of this claim may be false, since which truths are *a priori* can come to depend on the actions of language users, as when one standardizes the extension of the predicate 'is one meter long' or fixes the reference of the name 'Julius' by the definite description 'the sole inventor of the zip'.[13] But the cases we have in mind—the truths of standard modal logic—are not like this. To pick one paradigm case, *a priori* truths such as what is expressed by 'p, if actually p' are not the sort of truths for which one can be morally responsible.

That is all by way of arguing in favor of (A). We think that (A) has quite a bit going for it, and we know of no objections to it other than those which, like our puzzle above, explicitly or implicitly assume something like (B). But we cannot say the same for (B). As we aim to show in what follows, there are several reasons, not themselves depending on (A), and so independent of the present puzzle, for rejecting (B).

4 Two more puzzles

The most damning point against (B) has to do with puzzles similar to the one we discuss above: (B) generates comparable puzzles even if we abandon (A), given other independently plausible assumptions. For example, it is plausible that if someone is morally responsible for the fact that p, someone is likewise responsible for the fact that *actually* p. Equivalently, if nobody is responsible for the fact that *actually* p, no one is morally responsible for the fact that p. This idea is captured by the following general principle:

(D) $\mathrm{N}@\alpha \vdash \mathrm{N}\alpha$

But given that there is at least one truth for which no one is morally responsible, we get that no one is morally responsible for anything—an unacceptable conclusion.

For let p be any truth. Since p is true, so is $\Box@p$, by standard modal logic. Now let q be some truth such that $\mathrm{N}q$ holds. Since a necessary truth is entailed by any proposition, it follows that $\Box(q \rightarrow @p)$ is true. But then $\mathrm{N}@p$ follows from this and $\mathrm{N}q$, by (B), and so $\mathrm{N}p$ follows by (D). Therefore, no one is morally responsible for the fact that p. But p was arbitrary. Thus, given both (B) and (D), we cannot affirm the unimpeachable thesis that some but not all truths are truths for which no one is morally responsible.

[12] More narrowly still, what matters for our purposes is simply that nobody is morally responsible for conditionals such as $@p \rightarrow p$, and moreover that these conditionals are *true* Even if one does not think such conditionals are *logical* truths—perhaps, because some of them are only contingently true—one might still think no one is morally responsible for them.

[13] See Kripke (1980: 54–63), Evans (1979) and Lampert and Merlussi (2021a).

As before, something has to give if we are to avoid this paradoxical conclusion. We have seen that the positive rationale for (B) is less than impressive. On the other hand we know of no reason to doubt (D) that is independent of (B); otherwise, (D) strikes us as very plausible. Abandoning (B) is a single, unified solution to this puzzle and to the previous one; the alternative requires us to jettison both (A) and (D)—two principles which are independently plausible. Weighing the theoretical costs, it is (B) that ought to be rejected.

It is somewhat tempting to say that parallel considerations tell against (A): there are similar puzzles involving (A) that make no use of (B), and so perhaps rejecting (B) is not a promising response to our puzzle after all. We think this putative parity between the two arguments is illusory. Let us explain.

To understand the objection we have in mind here, consider that just as (A), (B), and (C) generate an inconsistency given standard modal logic, so too does the combination of (A) and (C) with (E):

(E) $\mathrm{N}\alpha, \Box(\alpha \leftrightarrow \beta) \vdash \mathrm{N}\beta$

To see this, note that the following inconsistency argument,[14] a slight modification of our original argument, goes through given (A), (C), (E), and standard modal logic:

(1) $p \,\&\, \neg\mathrm{N}p$
(2*) $\mathrm{N}(@p \leftrightarrow p)$
(3*) $\Box((@p \leftrightarrow p) \leftrightarrow p)$
(4) $\mathrm{N}p \,\&\, \neg\mathrm{N}p$

The conclusion, (4), is derivable just as before. Here, (2*) follows from (A) given standard modal logic; (3*) follows from (1) given that $p \rightarrow \Box((@p \leftrightarrow p) \leftrightarrow p)$ is a truth of standard modal logic.[15]

Given uncontroversial assumptions about the behavior of $\Box$, any counterexample to (E) is a counterexample to (B), but the converse is not guaranteed by standard modal logic alone. For this reason, there is some pressure to regard (E) as a genuine weakening of (B), and accordingly some pressure to regard the puzzle formulated in terms (E) as more basic than the original puzzle formulated in terms of (B). Or, more to the point, considering the two puzzles appears to tell against (A) when

[14] Many thanks to Brian Cutter for suggesting this argument and for suggesting some of its significance in the present context.

[15] An argument analogous to that given in footnote 4 above for the truth of $p \rightarrow \Box((@p \rightarrow p) \rightarrow p)$ suffices to show this:

1. $\Box(@p \rightarrow ((@p \leftrightarrow p) \leftrightarrow p))$
2. $\Box(@p \rightarrow ((@p \leftrightarrow p) \leftrightarrow p)) \rightarrow (\Box@p \rightarrow \Box((@p \leftrightarrow p) \leftrightarrow p))$
3. $\Box@p \rightarrow \Box((@p \leftrightarrow p) \leftrightarrow p)$
4. $p \rightarrow \Box((@p \leftrightarrow p) \leftrightarrow p)$

Step 1 is a necessitated tautology; 2 is an instance of the K axiom; 3 follows from 1 and 2; and 4 follows from (i), (ii), and 3.

evaluating the original puzzle, since rejecting (A) is a cogent response to both puzzles, but rejecting (B) appears to be a cogent response only to the first.

But a case can be made that (E) is not a genuine weakening of (B); granted, (E) is strictly weaker than (B) against a background of standard modal logic *alone*. But against a richer background which includes other plausible principles governing our *non-logical* vocabulary—in particular, governing the no-responsibility operator N—there are compelling arguments for taking (E) and (B) to be equivalent. In particular, consider the following plausible distribution principle:

(F) $\mathrm{N}(\alpha \,\&\, \beta) \vdash \mathrm{N}\alpha \,\&\, \mathrm{N}\beta$

Given (F), it is easy to show that (B) and (E) are interderivable. (E) is straightforwardly derivable given (B), since necessary equivalence is just two-way entailment. Showing that (B) is derivable from (E) given (F) is also straightforward. Suppose Np and suppose $\Box(p \rightarrow q)$. From the second it follows that $\Box(p \leftrightarrow (p \,\&\, q))$, and thus from the first it follows that N$(p \,\&\, q)$, given (E). Then by (F) we derive Nq, as desired.

So we doubt that the puzzle formulated in terms of (E) really does direct our attention away from (B): given (F), rejecting (B) amounts to rejecting (E), and vice versa. In this, there is an asymmetry between the two puzzles as regards the contest between (A) and (B). The first puzzle, in generating paradoxical conclusions by appeal to (B) but not (A), tells against (B); the second, though it generates an inconsistency by appeal to (A) and without explicit appeal to (B), does not plausibly tell against (A).

5 Counterexamples to (B)

Though principle (B) is explicitly defended in the literature only sparingly, a growing literature tells against it and against closely related principles.[16] Instead of surveying this literature, we will focus only on the most plausible and systematic counterexamples to (B) on offer. In the main, these have to do with the initially surprising suggestion that someone can be morally responsible for *necessary* truths.

Just as some logical truths are contingent, not all necessary truths are logical truths. And just as there are contingent logical truths for which no one can be morally responsible, so too there are non-logical necessities for which one *can* be morally responsible. A paradigmatic example is suggested by Kearns (2011,309):

> Stephen murders someone. Furthermore, it is completely uncontroversial that Stephen is morally responsible for the fact that he murders someone [...] He does so knowingly and intentionally, he could have done otherwise, he is aware of the wrongness of his action, etc. Thus Stephen is responsible for the fact that he murders someone. This being so, it is also clear that he is responsible for the fact that he *actually* murders someone. However, the fact that he actually murders someone is necessarily true. It is true in every possible world that, in the actual world, Stephen murders someone. Therefore, Stephen is (partly) morally responsible for a necessary truth.

[16] See, for instance, Kearns (2011) and Lampert and Merlussi (2021a, 2021b).

Kearns' verdict is bolstered by the intuitive principle (D), disussed in the previous section. The latter principle and standard modal logic together tell us that as long as someone is morally responsible for something, someone is morally responsible for something necessary. This is not the place to adjudicate the merits of Kearns' paradigm case, nor the merits of the principle (D). What interests us here is the uncontroversial fact that *if*, as Kearns suggests, one can be responsible for necessary truths, then there are counterexamples to (B).[17] For there is something no one is morally responsible for—the fact that Neptune has at least ten moons, for example. Now suppose someone is morally responsible for some necessary truth p. Then p, being necessary, is entailed by the fact that Neptune has at least ten moons. But then someone is morally responsible for something entailed by something for which no one is morally responsible, and hence there are counterexamples to (B).

Lest one conclude that all plausible counterexamples to (B) require the thesis that one can be morally responsible for necessary truths, we note that the puzzle with which this paper began generates counterexamples which require no such thing. Given that Stephen is morally responsible for the contingent fact that he murders someone, and given that no one is morally responsible for the logical truth that Stephen actually murders someone only if Stephen murders someone, we get a counterexample to (B), by now-familiar principles of standard modal logic.

And lest one conclude that all tempting counterexamples to (B) exploit the rigidifying behavior of the actuality operator (and what's wrong with that, after all?), we note that Kearns' paper includes a battery of further candidates not exploiting the logical behavior of 'actually'. We will add to this number the following three exemplary cases:

> **Inventing Bifocals**:[18] Benjamin Franklin is morally responsible for inventing bifocals. Franklin is then at least partly morally responsible for the fact that he is the inventor of bifocals. But then Franklin is at least partly morally responsible for the fact that he is the man who *in fact* invented bifocals. But nobody could *contingently* be the man who *in fact* invented bifocals. So Franklin is morally responsible for a non-contingent truth.[19]

[17] For a recent criticism of Kearns' argument, see Turner and Capes (2018) For a recent rejoinder to Turner and Capes, see Lampert and Merlussi (2021a).

[18] This case modifies one found in Lampert and Merlussi (2021a, 2021b).

[19] If characterizing things semantically is more natural, consider that 'Franklin' is a proper name, and 'the man who in fact invented bifocals' is a rigid definite description. So the equation 'Franklin is the man who in fact invented bifocals' is a true identity composed of two rigid designators, which is therefore necessarily true if true at all, by the necessity of identity. Therefore, it is necessary that Franklin is the man who in fact invented bifocals, and Franklin is morally responsible for the fact that he is the man who in fact invented the bifocals. If, however, the locution 'the man who in fact invented bifocals' sounds too much like 'the man who *actually* invented bifocals', we could just as well replace this rigid definite description with the demonstrative 'that very man', and the arguments would be unaffected. (For a semantics for rigid definite descriptions according to which they are not compositionally related to the actuality operator, see Zalta (1988).) Just as no one could contingently be the man who in fact invented bifocals, so too no one could contingently be *that very man*. Just as 'the man who actually invented bifocals' is a rigid designator, so too is the demonstrative 'that very man' (in our mouths).

Naming Numbers:[20] Before Stephen murders someone, someone introduces the name 'M' by the following metasemantic stipulation:
'M' names the number 1 if Stephen murders someone, and names 0 otherwise.
Since Stephen does murder someone, M = 1. Moreover, Stephen is morally responsible for this fact, since Stephen is responsible for the fact that he murders someone. But it is also *necessary* that M = 1. Indeed, the fact that M = 1 is necessarily equivalent to the fact that 1 = 1, an obvious necessary truth.

Pluralities: Consider the truths, i.e. the plurality of all truths. Since Stephen murders someone, the fact that he murders someone—call it *m*—is one of the truths, and moreover Stephen is morally responsible for the fact *m* is one of the truths. But there is also a specific plurality—namely, the plurality, call it the *tt*s, consisting of all and only the truths—such that Stephen is morally responsible for the fact that *m* is one of the *tt*s. But pluralities have their members necessarily: if *x* is one of *yy*s, then *x* is necessarily one of *yy*s. It follows, then, that Stephen is morally responsible for a necessary truth—Stephen is morally responsible for the fact that *m* is one of *tt*s.

Though these are all examples where someone is responsible for a necessary truth, it is important to emphasize that this is once again an inessential feature of our diet of cases. Just like the examples invoking the actuality operator, the cases above all admit of parallel cases centrally involving contingent rather than necessary truths. For example, Franklin is responsible for the contingent fact that he is the inventor of bifocals, but he is not responsible for the *a priori* conditional that he is the inventor of bifocals *if* he is the actual inventor of bifocals. But given that he is the actual inventor of bifocals, the fact that he is the actual inventor of bifocals only if he is the inventor of bifocals entails that he is the inventor of bifocals. Taken all together, we have another counterexample to (B).

At this point in the discussion it is worth making a point about the role of counterexamples in responding to our puzzle.[21] In some good sense, we take it that rejecting (B) is a foregone conclusion, given the diverse counterexamples that can be marshaled against it. And since there are such counterexamples to (B), one might for this reason worry that the question about how to address a *puzzle* involving the principle (B) turns out to be somewhat trivial: just drop (B)! But this is a mistake.

[20] This and the next case modify arguments found in Merlussi and Lampert (2022). This naming-style argument is inspired by Tharp (1989).

[21] Many thanks to an anonymous reviewer whose feedback motivated these remarks.

Consider one type of (putative) counterexample to (B) discussed in the literature—cases involving causal overdetermination. The following representative case appears in Stump and Fischer (2000).[22] Betty freely sets off an avalanche sufficient to destroy an enemy camp in the valley below. Suppose, moreover, that the laws of nature and the intrinsic state of the world in the remote past determine that a separate avalanche sufficient to destroy the camp takes place at the same time. Betty is at least partly morally responsible for the destruction of the camp, even though the destruction of the camp is necessitated by factors for which Betty is not even partly morally responsible—*viz.*, the laws of nature and the state of the world in the remote past—contrary to (B).

Setting aside any question about the merits of this sort of (putative) counterexample, what matters for our purposes is just the following point: rejecting (B) on the basis of such a counterexample does not by itself constitute a promising response to the puzzle we are considering in this paper. For puzzles like ours arise for principles only slightly different from (B), including those devised as *ad hoc* responses to counterexamples to (B). For example, consider the following *ad hoc* modification of (B):

(B*) $\mathrm{N}\alpha, \Box(\alpha \rightarrow \beta) \vdash \mathrm{N}\beta$, provided the fact that β is not causally overdetermined

The Stump/Fischer case is not a counterexample to (B*), even if it is a counterexample to (B). But puzzles arise for (B*) on the assumption that someone is morally responsible for something *that is not causally overdetermined*. So even if one rejects (B) on the basis of the Stump/Fischer counterexample, such a maneuver has limited usefulness as a response to our puzzle: similar puzzles immediately arise to which the response does not apply.[23]

By our lights, then, counterexamples to (B) are important, but they are less than decisive in guiding our response to puzzles like the one discussed in this essay. If we were just concerned with whether (B) holds, one counterexample would settle the issue. But since we are concerned with whether dropping (B) is the right response to our puzzle—a puzzle which turns out to be somewhat flexible and robust—what matters is that the counterexamples on offer are not only multiform, but they are also widespread and systematic. This places a substantial defensive burden on defenders of (B), a burden that we as theorists are unwilling to shoulder. But it also suggests that modal closure principles like (B), including gerrymandered principles designed to avoid extant counterexamples, are *in general* not very promising. Counterexamples tell against (B), and broad theoretical considerations tell against principles like (B) more generally.

[22] These counterexamples build on early discussions by Ravizza (1994), and Warfield (1996), among others.

[23] A similar point applies to a modification of (B) where β is causally determined by previous factors outside of everyone's control, even though full determinism may not hold—that is, where only "pockets of local determination" exist. This case is discussed by Stump and Fischer (2000: 50).

6 Hyperintensionality

A final theoretical cost associated with (B) is that it imposes the demanding requirement that the N operator be *non-hyperintensional*: as it were, the N operator cannot witness distinctions finer grained than mere modal distinctions. For suppose p and q are necessarily equivalent, and suppose no one is morally responsible for the fact that p. By assumption, p entails q. Given (B), it follows that no one is morally responsible for the fact that q. Thus necessary equivalence is also N-theoretic equivalence.

But there are reasons to suspect that the N operator is hyperintensional. The most concrete cases that bring this out are related to ones we have already considered, in suggesting that there are counterexamples to (B). If someone can be morally responsible for a necessary truth, but some necessary truths are such that no one can be morally responsible for them, then N is hyperintensional, since all necessary truths are necessarily equivalent. Cases favoring the hyperintensionality of N involving only contingent truths are also available.[24]

But perhaps more importantly, it would simply be surprising if the N operator turned out to be non-hyperintensional, at least given contemporary accounts of moral responsibility. To lack moral responsibility, it suffices to lack some necessary condition on moral responsibility, and many candidate necessary conditions on moral responsibility involve hyperintensional notions. (Given, that is, the common assumption that those notions *really are* hyperintensional; we omit this caveat in what follows.) For example, many accounts of moral responsibility set epistemic requirements on moral responsibility,.[25] and epistemic notions are frequently thought to admit of hyperintensionality; it would be somewhat surprising, then, if moral responsibility, or the lack thereof, were not likewise hyperintensional. Even the most widely discussed candidate requirement on morally responsible action—the ability to do otherwise, or the *Principle of Alternative Possibilities*—may motivate the hyperintensionality of N, if ascriptions of abilities are themselves hyperintensional.[26] In contrast to (B), however, (A) is perfectly acceptable even if N is hyperintensional. Once again, if there is a competition between (A) and (B), (A) wins hands down.

That said, principles like (B) are intuitively appealing to some people. And though (B) ultimately turns out to be problematic, identifying simple, natural, *unproblematic* principles in the neighborhood might suggest an explanation for why some people find principles like (B) plausible in the first place. Here is a conjecture:

[24] *Example*: as before, assume that someone is responsible for some contingent fact p and that no one is responsible for the logical truth $@p \rightarrow p$. Just as the actual truth of p gives us the entailment $\Box((@p \rightarrow p) \rightarrow p)$, so too we get the corresponding necessitated biconditional: $\Box((@p \rightarrow p) \leftrightarrow p)$. Thus we not only get a counterexample to (B), but we also get a counterexample to the non-hyperintensionality of N.

[25] See Rudy-Hiller (2018).

[26] See Spencer (2017), Lampert and Merlussi (2021a) and Merlussi and Lampert (2022) for cases motivating the hyperintensionality of ability ascriptions.

considerations about the hyperintensionality of moral responsibility might point the way forward.[27]

Until some decades ago, non-hyperintensional notions did most of the heavy lifting in philosophy, but some have argued that the metaphysician's toolkit ought to include hyperintensional resources as well. An example from Fine (1995) brings this out: the fact that Socrates exists is necessarily equivalent to the fact that the singleton set containing Socrates exists; nevertheless, one might think that one is *ontologically prior* to the other. If this is true, ontological priority is hyperintensional. The same has been claimed for other bits of metaphysical ideology, such as *ground*, whatever is expressed by metaphysically serious uses of 'because' or 'in virtue of', and more besides.[28] Our suggestion is this: perhaps a non-modal closure rule formulated in *hyperintensional* terms can perform better than (B), while also capturing some of (B)'s intuitive purchase.

Here's a toy example to work through this suggestion. Consider the fact that dinosaurs are extinct. No one is morally responsible for this fact. But neither is anyone morally responsible for the fact that dinosaurs do not exist. Being extinct and failing to exist are not the same thing. Unicorns do not exist, but they are not extinct. Yet, the fact that dinosaurs are extinct entails that they do not exist. This is an instance of (B). But it is also an instance of the following principle:

(Bec) $\mathrm{N}\alpha, \beta$ because $\alpha \vdash \mathrm{N}\beta$

According to (Bec), if no one is morally responsible for the fact that α, and the fact that β obtains because of α, then no one is morally responsible for the fact that β. Dinosaurs do not exist because they are extinct, they are extinct because of a massive asteroid impact, etc.

At first blush (Bec) performs as well as (B) does; plausibly better, since (Bec) handles overdetermination cases in stride. If one finds (Bec) plausible, or tends to reason in accordance with (Bec), that can help in explaining the initial plausibility of ultimately untenable principles like (B). Our first point, then, is that accepting a principle like (Bec) can help to explain what is initially plausible about (B), and gives correct verdicts about some cases where (B) arguably fails. But also, on standard views, the terms in which (Bec) is formulated are hyperintensional.[29] So here is our second point: what we said earlier about the hyperintensionality of moral responsibility gives us reason to think that a principle like (Bec) is the *right sort* of closure principle to appeal to in explaining the initial plausibility of (B).

For our purposes, the choice of (Bec) rather than some other hyperintensional closure principle matters little. We are concerned with the more general point. The failure of (B) is, by our lights, clear. But the intuitive purchase of (B) might

[27] Thanks to an anonymous reviewer for encouraging us to make this point.

[28] See, for example, Rosen (2010) and Nolan (2014). See also Todd (2013) for similar claims about *soft facthood*. We are not here endorsing the claim that all or any of these notions really are hyperintensional, nor are we endorsing any particular verdicts about the importance of these notions for metaphysics.

[29] Here we have in mind views according to which what is expressed by (metaphysically serious) uses of 'because' is hyperintensional; see Nolan (2014) as well as Schnieder (2011) and De Rizzo (2022).

be recovered. Finding simple and comparably intuitive principles that cover much of the same ground is a promising route towards this recovery. Noting the hyperintensionality of moral responsibility recommends starting with non-modal principles like (Bec).

7 A corollary: the failure of the direct argument

In previous sections, we have recommended abandoning (B) as a solution to the puzzle with which this paper began. An important corollary of this solution is that Peter van Inwagen's Direct Argument against the compatibility of determinism and moral responsibility is invalid.[30]

For van Inwagen's argument, let P_0 describe the complete state of the universe at some point in the remote (pre-human) past, and let L describe the laws of nature. On van Inwagen's conception of determinism, roughly, P_0 and L together entail every truth. Van Inwagen then recruits the operator N and assumes that it obeys the principles NECESSITY and DISTRIBUTION discussed in §3 above:

NECESSITY: $\Box\alpha \vdash \mathrm{N}\alpha$
DISTRIBUTION: $\mathrm{N}\alpha, \mathrm{N}(\alpha \to \beta) \vdash \mathrm{N}\beta$

It is also assumed that no one is even partly morally responsible for either P_0 or L. Taken together, then, these assumptions deliver the conclusion that determinism is true only if no one is morally responsible for anything. Where p is any arbitrary truth:

(5) $\Box((P_0 \,\&\, L) \to p)$
(6) $\Box(P_0 \to (L \to p))$
(7) $\mathrm{N}(P_0 \to (L \to p))$
(8) $\mathrm{N}P_0$
(9) $\mathrm{N}(L \to p)$
(10) $\mathrm{N}L$
(11) $\mathrm{N}p$

(5) follows from the assumption of determinism; (6) follows from (5) in standard modal logic; (7) comes from (6) by NECESSITY; (8) codifies the assumption that no one is morally responsible for the fact that P_0; (9) comes from (7) and (8) by DISTRIBUTION; (10) codifies the assumption that no one is morally responsible for the fact that L; and (11) comes from (9) and (10) by DISTRIBUTION.

What makes this argument interesting is that it does not assume that having moral responsibility requires the ability to do otherwise (nor even more modest requirements on the causal history of the agent in question).[31] For van Inwagen's argument,

[30] See van Inwagen (1983: 182–188).

[31] Views of the latter kind originated with Frankfurt (1969) and have been offered in various forms, for example, by Frankfurt (1971), Frankfurt (1987)), Fischer and Ravizza (1998), and Fischer (2006).

one need only assume that when it comes to moral responsibility—or, really, the lack thereof—NECESSITY and DISTRIBUTION hold.

But, if we are right, even this is too much to assume. For, as mentioned above, NECESSITY and DISTRIBUTION together guarantee (B),[32] and (B), we have argued, does not hold. If it did, together with (A), we would get the unacceptable result that no one is morally responsible for anything, regardless of whether determinism is true. So we reject (B) and, with it, we take the Direct Argument to be invalid.[33]

This is by no means the first challenge to the validity of the Direct Argument in the literature. As noted in the previous section, closure principles like (B) are hotly disputed. Much has been written suggesting that different closure principles required for running a version of the Direct Argument are invalid. But this literature has mainly proceeded by trading theorists' intuitions and debating the merits of tempting counterexamples. What this essay brings out, we think, is a much stronger basis for rejecting the sort of closure principles at issue. Our case against (B) is systematic, proceeding by citing clear, simple, and plausible principles that motivate rejecting (B). What previous discussions show is, at best, that there are a handful of (candidate) counterexamples to principles like (B); what our discussion shows is that, given other plausible principles like (A), *every* case where someone is morally responsible for something is a counterexample to (B). To our minds, this is a fresh and forceful case against the validity of the Direct Argument.

If what we have said so far is correct, this essay not only tells against the validity of the Direct Argument in a new way, but it also tells us something new about the Direct Argument more generally. So, in closing, let us suggest a new diagnosis of what is wrong with the Direct Argument. The Direct Argument purports to show that no one is morally responsible for anything if determinism is true, given the principles NECESSITY and DISTRIBUTION. But the Direct Argument proves too much: given the plausible claim that no one is morally responsible for logical truths, NECESSITY and DISTRIBUTION themselves suffice to establish that no one is morally responsible for anything. Given (A), in other words, the question of determinism is a red herring.

Rejecting determinism is not a promising response to the Direct Argument. The puzzles discussed in this paper show that the threat to moral responsibility highlighted by the Direct Argument is robust, regardless of the status of determinism. If we simply deny determinism without also questioning the logical assumptions at work in the Direct Argument, the threat remains. Happily, questioning these logical assumptions is independently worthwhile.[34]

Funding Open Access funding enabled and organized by Projekt DEAL.

[32] See footnote 9 above.

[33] This response applies to other versions of the Direct Argument as well, such as those offered by Warfield (1996).

[34] Many thanks to two anonymous reviewers for helpful comments; thanks as well to the editorial staff at *Philosophical Studies* for their labors. Special thanks to Brian Cutter and Pedro Merlussi for helpful conversations.

References

Crossley, J. N., & Humberstone, L. (1977). The logic of "actually''. *Reports on Mathematical Logic, 8*, 11–29.
Davies, M., & Humberstone, I. L. (1980). Two notions of necessity. *Philosophical Studies, 38*(1), 1–30.
De Rizzo, J. (2022). No choice for incompatibilism. *Thought, 11*(1), 6–13.
Evans, G. (1979). Reference and contingency. *The Monist, 62*(2), 161–189.
Fine, K. (1995). Ontological dependence. *Proceedings of the Aristotelian Society, 95*, 269–290.
Fischer, J. M. (2006). A Framework for moral responsibility. In J. M. Fischer (Ed.), *My way: essays on moral responsibility* (pp. 1–37). New York: Oxford University Press.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control*. Cambridge: Cambridge University Press.
Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*, 829–39.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.
Frankfurt, H. (1987). Identification and wholeheartedness. In Ferdinand Schoeman (Ed.), *Responsibility, character, and the emotions* (pp. 27–45). Cambridge: Cambridge University Press.
Goodsell, Z., & Yli-Vakkuri J. (xxxx). (forthcoming). *Logical Foundations*.
Grim, P. (1984). There is no set of all truths. *Analysis, 44*(4), 206–208.
Hanson, W. (2006). Actuality, necessity, and logical truth. *Philosophical Studies, 130*(3), 437–459.
Hanson, W. (2014). Logical truth in modal languages: reply to Nelson and Zalta. *Philosophical Studies, 167*, 327–339.
Hazen, A. (1978). The eliminability of the actuality operator in propositional modal logic. *Notre Dame Journal of Formal Logic, 19*(4), 617–622.
Kaplan, D. (1989). Demonstratives. In Joseph Almog, John Perry, & Howard Wettstein (Eds.), *Themes from Kaplan* (pp. 481–563). Oxford: Oxford University Press.
Köpping, J., & Zimmermann, T. E. (2018). Looking backwards in type logic. *Inquiry, 64*(5–6), 646–672.
Kripke, S. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic, 24*, 1–14.
Kripke, S. (1963). Semantical analysis of modal logic. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik, 9*(5–6), 67–96.
Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
Kearns, S. (2011). Responsibility for necessities. *Philosophical Studies, 155*, 307–24.
Lampert, F., & Merlussi, P. (2021a). Counterfactuals, counteractuals, and free choice. *Philosophical Studies, 178*, 445–469.
Lampert, F., & Merlussi, P. (2021b). How (not) to construct worlds with responsibility. *Synthese, 99*(3–4), 10389–10413.
Merlussi, P., & Lampert, F. (2022). Naming and free will. *Grazer Philosophische Studien, 99*(4), 475–484.
Nelson, M., & Zalta, E. (2012). A defense of contingent logical truths. *Philosophical Studies, 157*(1), 153–162.
Nolan, D. (2014). Hyperintensional metaphysics. *Philosophical Studies, 171*(1), 149–160.
Ravizza, M. (1994). Semi-compatibilism and the transfer of nonresponsibility. *Philosophical Studies, 75*, 61–93.
Rosen, G. (2010). Metaphysical dependence: grounding and reduction. In Bob Hale & Aviv Hoffmann (Eds.), *Modality: Metaphysics, Logic, and Epistemology* (pp. 109–135). Oxford University Press.
Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/.

Salmon, N. (2019). Impossible odds. *Philosophy and Phenomenological Research, 99*(3), 644–662.
Schnieder, B. (2011). A logic of 'because'. *The Review of Symbolic Logic, 4*(3), 445–465.
Spencer, J. (2017). Able to do the impossible. *Mind, 126*(502), 466–497.
Stump, E., & Fischer, J. M. (2000). Transfer principles and moral responsibility. *Philosophical Perspectives, 14*, 47–55.
Tharp, L. (1989). Three theorems of metaphysics. *Synthese, 81*(2), 207–214.
Todd, P. (2013). Soft facts and ontological dependence. *Philosophical Studies, 164*(3), 829–844.
Turner, P. R., & Capes, J. (2018). Rule A. *Pacific Philosophical Quarterly, 99*, 580–595.
Uzquiano, G. (2015). A Neglected Resolution of Russell's Paradox of Propositions. *The Review of Symbolic Logic, 8*(2), 328–344.
Van Inwagen, P. (1983). *An essay on free will*. Oxford: Oxford University Press.
Warfield, T. (1996). Determinism and moral responsibility are incompatible. *Philosophical Topics, 24*, 215–226.
Weatherson, B. (2001). Indicative and subjunctives. *Philosophical Quarterly, 51*, 200–216.
Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.
Yli-Vakkuri, J., & Hawthorne, J. (2021). Intensionalism and propositional attitudes. In Uriah Kriegel (ed.), *Oxford Studies in Philosophy of Mind Volume 2*. Oxford (pp. 114–174).
Zalta, E. (1988). Logical and analytic truths that are not necessary. *Journal of Philosophy, 85*, 57–74.