DOI: 10.1111/phpr.13130

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Dialetheism and the countermodel problem

Andreas Fjellstad | **Ben Martin**

FISPPA, University of Padova

Correspondence
Ben Martin, FISPPA, University of Padova.
Email: ben.lj.martin@gmail.com

Andreas Fjellstad and Ben Martin contributed equally to this article.

Funding information
Research for this paper was supported by a PNRR grant, under the European Union's NextGenerationEU research and innovation programme.

Abstract

According to some dialetheists, we ought to reject the distinction between object and meta-languages. Given that dialetheists advocate truth-value gluts within their object-language, whether in order to solve the liar paradox or for some other reason, this rejection of the object-/meta-language distinction comes with the commitment to use a glutty metatheory. While it has been pointed out that a glutty metatheory brings with it *expressive* deficiencies, we highlight here another complication arising from the use of a glutty metatheory, this time *evidential* in nature. According to this *countermodel problem*, while the thoroughgoing dialetheist who embraces a glutty metatheory can justify their *acceptance* of a rule of inference's *invalidity* using countermodels, to justify their renunciation of an unwanted rule they actually require the means to warrant their *rejection* of the rule's *validity*—which cannot be supplied by countermodels based on a standard dialetheic semantics. We end by sketching out a possible solution for the thoroughgoing dialetheist using a bilaterialist semantics.

KEYWORDS

bilateralism, dialetheism, explosion, invalidity, metatheory, refutation, rejection, self-referential paradoxes

1 | INTRODUCTION

According to advocates of dialetheism, some contradictions are true, as well as being false (Priest, 2006a; Weber, 2022). That is, there are truth-value gluts. Within the contemporary literature dialetheism has been motivated by a whole host of reasons, including inconsistent obligations (Priest, 2006b, Ch. 13), metaphysical considerations (Priest, 2006b, Ch. 11–12), and theological puzzles (Beall, 2021). However, probably those arguments for the position that have gained the greatest purchase in the literature are based upon the logico-semantic and mathematical paradoxes, including the liar, Russell, and sorites (Priest, 2006b; Weber, 2022).

As is well known, if the dialetheist wishes to avoid a trivial theory they require a paraconsistent logic, in which *ex falso quodlibet* (otherwise known as *explosion*) is invalid. While paraconsistent logics have the advantage of allowing the dialetheist to avoid triviality, they also come at the cost of invalidating other useful rules of inference, such as the disjunctive syllogism. Given that this latter inference rule is often thought to play an important role within mathematical reasoning (Burgess, 1983), its invalidation has led to the criticism that paraconsistent logics are unable to support mathematical results (Tennant, 2004).

Related to this concern, and sometimes motivated by it, dialetheists disagree over the extent to which they must always use (or, reason according to) a paraconsistent logic. Some, such as Beall (2013b), propose that it's possible to isolate and quarantine gluts, and thereby permit classically valid reasoning in those cases we have assurance that the relevant domain behaves classically. Consequently, if metalogic is such a domain, we can go ahead and use classical logic there with the assurance that we won't meet any gluts that trivialise the theory. One potential positive upshot for the dialetheist, if they could provide us with this reassurance, is that we could reason classically to prove the desired results about our target paraconsistent logic. This is *prima facie* advantageous given that classical logic is deductively stronger than paraconsistent logics, and so we are more likely to be able to successfully complete the desired proofs about our object theory than if we used a glutty paraconsistent metatheory. In what follows, we'll refer to variants of dialetheism which allow for a non-glutty metatheory, and thus a distinction between an object and meta-language, as *moderate* dialetheism.

In contrast, other dialetheists reject the possibility of isolating these gluts and thus drawing a distinction between those domains that behave classically and those which don't (most notably, Weber (2022)). One result of this view is that no viable distinction can be drawn between the logic of our metatheory and that of our object theory. If we need to endorse a glutty paraconsistent logic in virtue of the existence of gluts, then we need to accept one all the way up (so to speak). In what follows, we'll refer to versions of dialetheism which reject the distinction between an object language and metatheory as *thoroughgoing* dialetheism.[1]

[1] In addition to Weber's own work, thoroughgoing dialetheism is also defended, or at least explored, in Badia et al. (2022) and Tanaka and Girard (2023). More historically, Routley (1977a, 1977b, 1979) was probably also an advocate of thoroughgoing dialetheism, with his quest for a universal "ultra" logic that allowed us to keep the "simplicity" of naïve set theory and a semantically closed language. While Routley (1980) did at times suggest that a classical metatheory could be used, this seems more for pragmatic reasons to preach to the unconverted; classical shackles that could be subsequently thrown off. Priest (1990, p. 208) also probably counts as a thoroughgoing dialetheist, given his claim that the *telos* for a dialetheic solution to the semantic paradoxes is a rejection of the object-/meta-language distinction (though, see the bootstrapping argument for why it's acceptable for a dialetheist to use a classical model theory in Priest, 2006a, p. 257; cf. Meadows, 2015). Our impression is that the view is slowly maturing into a research programme that deserves proper consideration and discussion.

Our goal in this paper is not to assess the general motivations for being a dialetheist *simpliciter*, nor to adjudicate over the choice between a *thoroughgoing* and *moderated* form of dialetheism. Rather, it is to present a new interesting challenge for the thoroughgoing dialetheist, which we call the *countermodel problem*, that has no respective impact on moderate dialetheism.

One of the properties we desire from the semantics of our chosen logic is that it is able to license our acceptance or rejection of some given set of inference rules by demonstrating their validity or invalidity, respectively (Martin & Hjortland, 2021). This is just as true for the dialetheist as anyone else. In particular, they wish to renunciate certain rules of inference such as the disjunctive syllogism, in order to avoid the charge of trivialism. With a *non*-glutty metatheory the method for how one goes about showing a logic licenses the rejection of a rule of inference is straightforward—one simply demonstrates that it is invalid in the logic by providing a countermodel (relative to the logic's consequence relation).

For the thoroughgoing dialetheist who uses a glutty metatheory, however, the situation is not so simple. After all, as we shall see, once one embraces a glutty metatheory based on a standard semantics, demonstrating the invalidity of a rule of inference by producing a countermodel fails to license its rejection, as the inference may still also be satisfied by the very same model. An inference can have a countermodel *and* be satisfied by every model. Consequently, the thoroughgoing dialetheist currently lacks the means to warrant their rejection of the validity of these unwanted rules, and so a new dialetheic-friendly semantics that possesses the resources to provide this warrant is needed.[2]

The rest of the paper runs as follows. Section 2 provides the background to the countermodel problem, including the thoroughgoing dialetheist's favoured semantics and their recognised upshots. Section 3 then presents the problem itself and distinguishes it from existent problems in the literature, such as the "just true"-problem. Lastly, Section 4 sketches out a possible bilateralist solution to the problem, highlighting that there are available semantics which can license the thoroughgoing dialetheist's rejection of the unwanted rules.

2 | TOWARDS A GLUTTY SEMANTICS

2.1 | Motivating thoroughgoing dialetheism

Several motivations have been given for rejecting the possibility of isolating gluts and drawing the distinction between a glutty object-language and non-glutty metatheory.[3]

[2] *A brief note on terminology*: Throughout this paper we stick with the standard nomenclature of treating dialetheism (*simpliciter*) as the claim that some contradictions are true, or some truth-bearers are both true and false, with the latter being equivalent to the claim that some truth-value gluts actually exist (Priest 2006a, p. 1; Weber 2022, pp. x & 3). Jc Beall (2022) has suggested in a recent review of Weber (2022) that we ought to move away from this nomenclature, and instead refer to positions committed to true contradictions as simply "glut theories", and keep the title "dialetheism" for those positions that take on the further commitment of rejecting the meta-/object-language distinction (what we call in this paper "thoroughgoing" dialetheism). While sympathetic to Beall's terminological concerns, we don't follow his recommendation here so as not to confuse the general reader who will be more used to the standard terminology. However, of course, if Beall's new nomenclature catches on, the same points we make here can be re-expressed in his terms. Many thanks to an anonymous referee for asking us to clarify this point.

[3] What follow are just a few of the available reasons, not an exhaustive list. For a more detailed description of these and other motivations, see Weber (2022, Chs. 1–3).

First, one may believe that gluts are particularly abundant, permeating all of our theories and thus simply impossible to isolate. For instance, if one believes that gluts arise from instances of vagueness then, given the abundance of the latter, gluts themselves will be found everywhere. There is no domain which is immune, including in mathematics; "after all, almost every predicate is vague" (Weber, 2022, p. 95).

Second, putting considerations of vagueness to one side, if one believes that the domain of mathematics itself has been permeated by gluts, for instance due to the set-theoretic paradoxes, then the dialetheist might consider it ill-advised to expect a stable classically well-behaved metatheory using objects from this domain—after all, who knows what new paradoxes might crop up? As Weber (2022, p. 89) notes, "[a] non-classicist *might* assume that some finite abelian group is consistent, and reason accordingly, without embarrassment; but when it comes to theories that focus on the notions of truth and proof, I don't see how the reliability of classical metatheory can be taken for granted," given that this is *just where* we would expect the problems to arise!

Third, following philosophical tradition, if one presumes that logical laws provide us with the most general principles of reasoning which hold come what may regardless of subject matter, then one might well think this commits one to using the same logic across one's object and metatheory. After all, if one uses some logic $\mathcal{L}_1$ in one's object language, but another $\mathcal{L}_2$ in one's metatheory, then this is tantamount to admitting that logic is not wholly general. Thus, if the dialetheist is committed to a glutty paraconsistent logic in their object language and are "motivated by the allure of a closed, complete theory" (Weber, 2022, p. 93), they may well feel the further pull to use the same logic for their metatheory.

Finally, if one's dialetheism is motivated by the goal of providing a comprehensive solution to the liar paradox, then one might also be wary of allowing for a non-glutty metatheory. After all, according to dialetheists, all non-dialetheic solutions to the liar fail either because they: (i) are susceptible to revenge versions of the paradox, and so incomplete; (ii) unnecessarily restrict the expressibility of our natural languages, and are thus contrary to our non-logical commitments (for example, regarding semantic closure); or (iii) are ad hoc, by lacking independent motivation (Priest, 2006b, Ch. 1). In contrast, the dialetheist assures us that their theory suffers none of these weaknesses. Unlike non-dialetheic solutions, the dialetheist suggests they can provide a principled and comprehensive solution to the liar paradoxes that respects the expressibility of our natural languages (by accepting semantic closure). Yet, these motivations for rejecting non-glutty solutions to the liar also provide motivation for the dialetheist to reject the object-/meta-language distinction.

Firstly, if the dialetheist were to use a non-glutty metalanguage in conjunction with their glutty paraconsistent object language, then revenge versions of the paradox evading a dialetheic solution would be bound to crop up. Just take whatever the "exclusionary" operator is from the non-glutty metalanguage and use it to construct a revenge liar which evades a dialetheic solution, on pain of the non-glutty metalanguage collapsing into a glutty language (Berto, 2014; Martin, 2015). Better, then, to stick with a glutty metalanguage. Secondly, in arguing for a dialetheic solution to the liar paradoxes, Priest (2006b) explicitly appeals to the ad hocness of discriminating between an object and meta-language as a reason to reject the Tarski-inspired language-hierarchy solution to the paradoxes. Consequently, if the dialetheist were required to distinguish between an object and meta-language, their solution too would be at risk of being ad hoc according to (at least) some dialetheists' own lights.

Thus, in order to live up to the promise of providing a comprehensive solution to the liar which doesn't fall foul of their own criticisms of other solutions, there's some motivation for the dialetheist to endorse a glutty metatheory too. Indeed, Priest (1990, p. 208) goes so far as to say that "the

whole *point* of the dialetheic solution to the semantic paradoxes is to get rid of the distinction between object language and meta-language" (cf. Weber (2022, p. 94)).

2.2 | Semantics for and with paraconsistent logic

In order to provide a dialetheic solution to the logico-semantic and mathematical paradoxes without committing themselves to trivialism, the dialetheist requires a logic which is both (i) *glutty*, that is allows propositions to be both true and false in an interpretation, and (ii) *paraconsistent*, whereby $\{\varphi, \neg\varphi\} \nvDash \psi$, for some φ and ψ (Martin, 2021a). Moreover, dialetheists often wish to take on the further commitment that the extensional connectives should behave like Boolean connectives to the extent that, for instance, adjunction is valid and the conjunctions of contradictory propositions are false (Priest & Routley, 1989, pp. 158–9 & 164–5).

Not all paraconsistent logics fulfil these two further criteria. The preservationist logic of Jennings and Schotch (1984), for instance, is paraconsistent without being glutty, and da Costa's (1982) $C_i(1 \leq i \leq \omega)$ logics are often shunned by dialetheists for not containing a negation which ensures that a proposition φ's negation is true if and only if φ is false (Priest & Routley, 1989, pp. 164-5). Some logics do fulfil all three criteria, however, with the most popular being the logic presented by Priest (1979) as the *Logic of Paradox* (LP).[4]

As has now become standard, we'll use a relational rather than a truth-functional presentation of LP's semantics.[5] Without going into details about the meta-theoretic machinery we use to define the interpretations, we assume that the formulas for which we provide a semantics are represented by constants, and that we have defined for each interpretation I (which is also represented by a closed term in the metalanguage) a two-place predicate r_I that satisfies the following principles where t and f are further constants intuitively representing truth and falsity (with the expression 'iff' being the biconditional of the language):[6]

- $r_I(\varphi, t)$ or $r_I(\varphi, f)$
- $r_I(\neg\varphi, t)$ iff $r_I(\varphi, f)$
- $r_I(\neg\varphi, f)$ iff $r_I(\varphi, t)$
- $r_I(\varphi \vee \psi, t)$ iff $r_I(\varphi, t)$ or $r_I(\psi, t)$
- $r_I(\varphi \vee \psi, f)$ iff $r_I(\varphi, f)$ and $r_I(\psi, f)$
- $r_I(\varphi \wedge \psi, t)$ iff $r_I(\varphi, t)$ and $r_I(\psi, t)$
- $r_I(\varphi \wedge \psi, f)$ iff $r_I(\varphi, f)$ or $r_I(\psi, f)$

[4] Originally formulated by Asenjo (1966) as the *Calculus of Antinomies*. Other options include the propositional fragment of Baten and De Clerq's (2004) CLuNs. Nothing rests on our particular choice of logic here, however—the same results apply to any logic that adheres to the three conditions above.

[5] For the rationale behind using a Dunn (1976)-style relational semantics for LP, see Weber et al. (2016).

[6] Given that modus ponens is invalid in LP for the material conditional, defined as $\neg A \vee B$, the Boolean connectives are usually supplemented with a further conditional validating the rule, and the language's biconditional 'iff' defined in terms of that conditional and conjunction in the usual way. Dialetheists disagree over the best account of this new conditional; see, for instance, Beall (2009) and Priest (2006b), but also more recent proposals such as Weber (2010), Badia and Weber (2019), and Badia et al. (2022). We won't take a stand here on which of these candidates are best for the dialetheist's purposes. Exactly which properties such a conditional should possess, for instance whether it should contrapose or not with the paraconsistent negation, are certainly substantive matters for any dialetheic project. However, we are not of the impression that the exact properties of the conditional beyond modus ponens (as a rule of proof) are relevant to either the formulation of the countermodel problem or our solution to it. Of course, if we are wrong on this score, this could provide additional support for one of these specific candidate conditionals within the literature.

We can furthermore define a two-place predicate representing a notion of validity as follows: $\varphi_0, \ldots, \varphi_n \vDash \psi$ iff for every I if $r_I(\varphi_i, t)$ for each φ_i then $r_I(\psi, t)$.[7]

2.3 | Expressive limitations?

As was recognised relatively quickly, without the aid of further communicative resources the dialetheist's semantics would possess expressive limitations (Parsons 1990). After all, as it stands LP's semantics do not facilitate the dialetheist's ability to *express disagreement* over the truth of a proposition. If a dialetheist's interlocutor were to assert φ, it isn't enough for the dialetheist to disagree by responding that φ *is false*, given that φ's falsity fails to preclude φ's simultaneous truth. Nor can the dialetheist successfully disagree by asserting the negation of φ for, again, it is perfectly possible that both φ and $\neg\varphi$ are true according to the dialetheist. Consequently, the dialetheist requires another means to express their disagreement with the interlocutor's assertion that φ, by somehow communicating their *rejection* of φ; where rejection is the mental state of refusing to believe φ (Priest, 2006b, p. 98).

The dialetheist's now common line of response to this challenge is to accept that their object-language semantics cannot itself successfully express rejection, and propose that the problem is dealt with at the level of pragmatics with the introduction of two independent and exclusionary speech-acts, assertion and denial (Priest, 2006b). While the speech-act of *assertion* serves to express the dialetheist's acceptance of a given proposition (as for the non-dialetheist), *denial* of a proposition φ expresses the rejection of φ by precluding its simultaneous assertion. Thus, to express their rejection of φ, and subsequent disagreement with their interlocutor over φ's truth, the dialetheist can do so by *denying* φ.[8]

Whether the introduction of this new *sui generis* speech-act communicating rejection solves all of the expressive deficiencies of the dialetheist's semantics is a moot point. After all, as Shapiro (2004) has pointed out, even the dialetheist may wish to express that a proposition is *true only* or *false only* within contexts that the speech-act of denial is inappropriate. For instance, they may wish to *assume* that a given proposition fails to be true, and then infer what follows. Or, more generally, they may want to include the presumption that a proposition is false *only* within a truth-functional setting, such as ⌜ If φ isn't true, then consequences $\psi_1, \psi_2, \ldots \psi_n$ follow ⌝.[9]

The problem we are interested in here, however, arises independently of whether denial can be used by the dialetheist within all important contexts to express their rejection of a proposition's truth (or, falsity). Or, indeed, other concerns over whether the dialetheist is always capable of expressing important semantic properties and their own commitments accurately. While associated with these *expressive* concerns, the countermodel problem is instead a concern about the *justificatory* limitations of the thoroughgoing dialetheist's semantics.[10]

[7] Of course, because one can have both $r_I(\delta, t)$ and $r_I(\delta, f)$, for some interpretation I and formula δ, one can also have $r_I(\psi, f)$ without the argument's validity being impacted.

[8] This may suggest that bilateralism, which treats both acceptance and rejection as primitive (as opposed to defining rejection as simply the acceptance of a negation), is particularly well-suited for the dialetheist's purposes, as has been suggested by Restall (2005). We come back to this point with our own solution to the countermodel problem in Section 4.

[9] For more on the debate over the putative expressive limitations of glutty semantics, and the consequences for the dialetheist's overall proposal, see (Littmann & Simmons, 2004; Jenny, 2017; Martin, 2015, 2021b; Omori & Weber, 2019; Young, 2015).

[10] We leave a more detailed discussion of how exactly the proposed countermodel problem differs from these established *expressive* limitations until later, in Section 3.3 below.

2.4 | A parallel asymmetry for justification?

Just as the means through which the dialetheist is able to *express* their rejection of a proposition is different to a non-dialetheist, so is the means through which they may *justify* their rejection of a proposition. After all, for the non-dialetheist, possessing justification for φ's falsity (and thus $\neg\varphi$'s truth) is equivalent to possessing justification for rejecting φ. In contrast, this inferential path is closed for the dialetheist, given that neither a proposition's falsity nor the truth of its negation preclude its simultaneous truth. In fact, they must reject the following two principles which we shall henceforth refer to as (FR) and (NR), respectively:

Falsity-Rejection: Justification for asserting the falsity of φ justifies the rejection of φ.

Negation-Rejection: Justification for asserting the negation of φ justifies the rejection of φ.

This ensures that for the dialetheist, providing justification for the rejection of a proposition can be quite different than for the non-dialetheist. Indeed, the dialetheist using LP, or any other glutty paraconsistent logic respecting the normal semantics for negation, in principle divorces the process of justifying one's rejection of a proposition φ from the processes of justifying acceptance of φ's falsity or φ's negation.

In order to re-establish a close connection between the justified rejection of φ and evidence for $\neg\varphi$, the dialetheist requires assurances that the sphere of enquiry in which φ is contained behaves consistently. Only then can the dialetheist rely upon the same considerations that the non-dialetheist does to justify the rejection of φ.[11] Otherwise, as Priest (2006b, p. 103) recognises, "the arguments for the negation of something are not, without some other considerations pertaining to the consistency of the situation, a complete case against the claim being negated. Hence, arguments *pro* and *contra* are *sui generis*." To emphasise the point—just as the speech-acts of assertion and denial are *sui generis*, and only the latter communicates the rejection of a proposition φ, so the requirements for *justifying* the acceptance of a proposition, and justifying its rejection are *sui generis*.

How the dialetheist successfully goes about providing justification for their rejection of a given proposition will, of course, depend upon the circumstances and forms of evidence relevant to the proposition's evaluation. One option is to show that the relevant proposition φ entails another ψ which we have independent reasons to reject. This is simply a dialetheic friendly version of *reductio*.[12] Another is to provide independent evidence for the target state of affairs failing to hold. This could come in the form of direct empirical evidence. For instance, if observing the colour of a wall, we can rationally reject the claim that the wall is blue on the basis that an absence of blue shades is observed on the wall. Or, in the case of mathematics, it could come in the form of a direct counterexample; in order to justify one's rejection of "All linear functions in one variable are perpendicular to one another", it suffices (even for the dialetheist) to provide a linear function that fails to possess this property.

Establishing how exactly the dialetheist can go about justifying their rejection of a proposition in all relevant cases is beyond the scope of this paper but unnecessary for our purposes.

[11] How such an assurance could itself be justified is unclear. Priest has previously attempted to provide criteria for judging when gluts are likely to occur and not (Priest, 1995). However, given the concerns since raised over these criteria (Beall, 2001, 2014), we consider it an open question whether (and how) the dialetheist can provide us with the required assurances.

[12] One needs to be careful when using *reductios* within the context of dialetheism, as the traditional formal meta-rule of *reductio* is invalid for dialetheists (given that contradictories can be simulatenously true). However, this does not stop dialetheists from freely using *reductio ad absurdum* arguments which show a proposition to be rationally rejectable because it entails an absurdity (which for the dialetheist need not include all contradictions). On this, see Priest (2006a) and Martin (2021b).

Rather, we need only recognise here that the dialetheist's requirements for rationally rejecting a proposition φ are independent of the requirements for rationally accepting $\neg\varphi$, and thus φ's falsity (*unless*, that is, we have assurances the relevant scenario behaves consistently). In the next section, we consider what impact the dialetheist's rejection of (FR) and (NR) has on their ability to reject an unwanted rule of inference once they embrace a glutty paraconsistent metatheory.

3 | THE COUNTERMODEL PROBLEM

3.1 | The need to *reject* rules of inference

There are various classical inference schemas that the dialetheist desires to avoid because they would commit them to trivialism. However, it follows from our discussion in section 2 that, in order to do so, the dialetheist must supply reasons to *deny their validity* rather than simply reasons to *assert their invalidity*. Indeed, what the dialetheist really wishes to do is preclude the validity of these rules of inference, for only then will they be assured to have blocked the implication to triviality. It isn't enough that triviality doesn't follow, if it *also does follow* from one's commitments. In other words, what they require is justification for the *rejection* of the rules in question. Recognition of this requirement is sometimes, at least, reflected in discussions of dialetheism and paraconsistency:

> [In order to permit a paraconsistent set theory, one must] allow for the set theory to entail contradictions, but reject the principle ex contradictione quodlibet. (Priest, 2006b, p. 247)

> Dialetheic paraconsistent logicians, by contrast, precisely reject explosion because they think that at least some contradictions are (or may be) true. (Allo, 2010, p. 28)

Despite this, it still seems common in the literature to (implicitly) assume that although the dialetheist rejects the principle (NR), in the particular case of logical consequence, demonstrating the invalidity of a schema is *sufficient* for precluding (and so rejecting) its validity. After all, when one presumes a consistent metatheory, this is true. To demonstrate the invalidity of some argument schema $\Gamma \vDash \psi$, it suffices to show that for some interpretation r', $\forall\varphi \in \Gamma, r'(\varphi, t)$ and *it is not the case that* $r'(\psi, t)$. Given that within a consistent metatheory there is no interpretation r' in which both $r'(\varphi, t)$ and *it is not the case that* $r'(\varphi, t)$, for any formula φ, this countermodel precludes $\Gamma \vDash \psi$ and thus suffices to justify the dialetheist's *rejection* of the schema's validity.

The problem, of course, is that the dialetheist only has the right to accept (NR) in those cases in which they are assured that the situation behaves consistently. While this is true in the case of (in)validity when we have a consistent metatheory, in the case of the thoroughgoing dialetheist who desires a glutty metatheory the situation is far trickier, for they have no means to ensure that interpretations behave consistently. Indeed, one of their goals in embracing a glutty metatheory is to ensure that they need not behave consistently. As we shall now see, this spells trouble for their semantics.

3.2 | No consistent metatheory, no useful countermodels

So far we've highlighted three points. Firstly, there are a sub-group of glut-theorists, the *thoroughgoing* dialetheists, who reject the object-/meta-language distinction. Thus, in virtue of endorsing a glutty paraconsistent object-language, they also accept a glutty paraconsistent metatheory for their logic. Secondly, dialetheists (*simpliciter*) reject (NR) and (FR), except for those cases in which they are assured the situation behaves consistently. Thus, justification for the acceptance of $\neg\varphi$ does not suffice as justification for the rejection of φ. Finally, to be assured their logic blocks absurd consequences (namely, triviality), the dialetheist must be justified in believing their logic sanctions the *rejection* of the validity of certain argument schema.

These three commitments combined spell trouble for the *thoroughgoing* dialetheist. First, by using a glutty metatheory, they are committed to accepting that there are interpretations r' such that for any formula φ and truth-value t both $r'(\varphi, t)$ and *it is not the case that* $r'(\varphi, t)$. Interpretations can be inconsistent. In fact, this is the *whole point* of having a glutty metatheory. Thus, possessing justification for the invalidity of some argument schema Δ will not suffice for being justified in rejecting the validity of these schema Δ; they require some independent means with which to justify their *rejection* of the validity of these schema. Unfortunately, their current semantics provide them with no such means.

Using these semantics, in order to justify their rejection of the validity of some given argument schema $\Gamma \vDash \varphi$, they would need to somehow preclude the possibility that, for all interpretations r, if $\forall\psi \in \Gamma, r(\psi, t)$ then $r(\varphi, t)$. However, the only means the thoroughgoing dialetheist has in their semantics to attempt to preclude this possibility is by providing a countermodel in which $\forall\psi \in \Gamma, r'(\psi, t)$ and $\neg r'(\varphi, t)$. But $\neg r'(\varphi, t)$ fails to preclude $r'(\varphi, t)$ in a glutty metatheory. So, their current semantics provide no means to justify the desired rejection of the validity of these unwanted schema. All the thoroughgoing dialetheist has at their disposal are countermodels that within a glutty metatheory demonstrate invalidity; they do not license the rejection of validity.

The *thoroughgoing* dialetheist who accepts a glutty metatheory, therefore, is in a very different situation to the non-glutty logician.[13] Given that for the non-glutty logician the acceptance of a statement φ automatically sanctions the rejection of φ's negation (and vice versa), in virtue of having the right to accept the *in*validity of some inference they concurrently gain the right to reject the validity of the inference (and vice versa). In comparison, by rejecting (NR) and insisting upon a glutty metatheory, the thoroughgoing dialetheist's demonstration of the *in*validity of an inference licences *solely* the acceptance of its invalidity rather than the required rejection of its validity.

Where, then, does the thoroughgoing dialetheist go from here? They seem to be in a bind. After all, they need their semantics to sanction the rejection of unwanted rules of inference, given that they require their semantics to demonstrate how they are able to endorse contradictions without triviality following. Yet, their current semantics with a glutty metatheory fails to deliver the goods. Further, removing the requirement for a glutty metatheory isn't a viable option, as this would amount to simply retreating to a more *moderate* dialetheism.

The challenge then is to provide a way for the thoroughgoing dialetheist to preclude the validity of an inference within their semantics, and so license the rejection of the unwanted rules of inference, *without* jettisoning the glutty metatheory and losing the putative benefits of their position. We see two possible routes to achieve this.

First, the thoroughgoing dialetheist could provide an extra-logical means to license the rejection of the validity of the unwanted rules of inference. For instance, they could hope to show through

[13] And, indeed, the *moderate* dialetheist, who allows for a non-glutty metatheory.

empirical observations that, for some instances of the rule, the premises *actually are true* while the conclusion *fails* to be true.[14] In this case, the dialetheist could hope to provide real-life "countermodels" by specifying instances of premise sets that they ought to (jointly) accept, and conclusions that they ought to reject.[15] This proposal is not without its challenges, however. Firstly, one would need to establish that there are available empirical observations which would indisputably license the rejection of the rules of inference, by showing that there are instances of the argument schema in which the premises ought to be accepted and conclusion rejected (this is a topic considered in the past by Priest (2006a), without conclusive results). Secondly, endorsing this option would be tantamount to admitting that, unlike their non-dialetheic colleagues, the thoroughgoing dialetheist fails to possess a semantics which shows whether they are licensed to accept or reject some argument schema. Given that possessing a semantics that is able to deliver on a position's goals is an important criterion for a theory of logic (Martin & Hjortland, 2021), this lack would itself count against the plausibility of thoroughgoing dialetheism in comparison to its non-dialetheic rivals.

This leads us onto the second route, to develop a semantics that rectifies the shortcomings of the present approach. This requires providing a semantics that: (i) respects the need for a glutty metatheory, whilst (ii) including within its semantics markers that justify the *rejection* of a proposition, and thus successfully preclude its simultaneous acceptance, facilitating the production of countermodels that license the rejection of an inference. The challenge for this latter approach is to provide such a semantics without revenge paradoxes resurfacing. It is the possibility of just such a semantics we explore in Section 4.

Before we move onto discuss these semantics however, it will be instructive to briefly outline how the *countermodel problem* is distinct from those existent *expressive* concerns raised over the dialetheist's semantics, noted in Section 2.

3.3 | Four distinct problems

There are three notable concerns that have already been raised against the dialetheist's glutty semantics: (i) the *just-true* problem (Beall, 2013a; Littmann & Simmons, 2004; Young, 2015); (ii) the *recapture*, or *exclusion*, problem (Beall, 2013b; Berto, 2014; Jenny, 2017); and (iii) the *invalidity revenge* problem (Young, 2019). Each, though connected to the *countermodel* problem, are distinct from it.

According to the *just-true* problem, the dialetheist lacks the ability to express within their semantics that a given proposition is true *only* (or, false *only*), thereby precluding its simultaneous falsity (or, truth). Showing that this is so with a glutty metatheory is straightforward. For any interpretation r and proposition φ, the most that the dialetheist can do within their semantics to express that φ is true (or false) only is to state that *it is not the case that* $r(\varphi, f)$ and *it is not the case that* $r(\varphi, t)$, respectively. Yet, given that the metatheory is glutty, both effectively fail to preclude that φ is *also* false or true, as we can have both $r(\varphi, f)$ and $r(\varphi, t)$, respectively, as well. Thus, the

[14] Whether the dialetheist is able to *express* that the conclusion *fails to be true*, given a glutty metatheory, is of course another point. However, we are interested here in the *evidential* constraints placed upon dialetheists by a glutty metatheory, not *expressive* constraints.

[15] We consider a quasi-experimental approach to formal systems, where one rejects inference rules because the result of adding them to some preferred system trivialises the system, as a variation on this first route. Thanks to an anonymous referee for pushing us on this point.

semantics fail to be able to express that a given proposition is true (or, false) *only*, leading to an expressive limitation in the thoroughgoing dialetheist's semantics.

How much this expressive limitation matters depends on the extent to which one thinks the dialetheist ought to be able to distinguish between propositions which are glutty and those which behave consistently. After all, the (non-trivial) dialetheist doesn't think that all propositions are glutty: most will think that "Boris Johnson is a fried egg" is only false, and most will believe that "2 + 2 = 4" is only true. On this point, given that the dialetheist criticises the non-dialetheist's semantics for being expressively deficient, it seems that the inability to express their own commitments counts equally against the dialetheist's semantics (Littmann & Simmons, 2004).[16]

The *recapture*, or *exclusion*, problem by contrast is the criticism that the dialetheist is unable to "recapture" classical validity in cases where we have assurances the situation behaves consistently. That we should desire this from the dialetheist's semantics is due to the fact that: (i) as dialetheic semantics only invalidate certain important rules of inference such as the disjunctive syllogism and modus ponens because of inconsistent scenarios, we ought to be able to use these important rules of inference when we have assurances that the situation is consistent, such as in mathematics. Not licensing these rules of inference in such cases will lead to important inferential deficiencies, including when it comes to proving mathematical results (Williamson, 2017); further, (ii) rival semantics, such as the gappy K3, are able to recapture classical rules of inference without problem in non-gappy situations (Jenny, 2017). What we have then is not only potentially a theoretical weakness on the part of the dialetheist's semantics, but a *comparative* weakness in light of rivals' ability to meet the challenge.

While the *recapture* problem is connected to the *just-true* problem, they are nonetheless distinct (Young, 2015). A solution to the *just-true* problem should also bring a solution to the *recapture* problem. By being able to specify that a set of propositions are either *true only* or *false only*, and thus precluding the troublesome glutty interpretations, the dialetheist can ensure the interpretations behave classically and suitably recapture classical validity.[17] However, solving the *just-true* problem is not the only possible route to solving the *recapture* problem. Instead, one could introduce into one's semantics non-logical constants which somehow communicate that the relevant situation behaves consistently, without one's semantics thereby having the capacity to express that a given proposition is *true only* or *false only*. Such a solution has been explored by both Beall (2013a) with his "Shrieking" operator, and Berto (2014) with his "exclusionary" operator on predicates, which is intended to communicate that two properties are somehow primitively metaphysically incompatible.

While obviously associated with these existent problems, the *countermodel* problem is nonetheless distinct from both. Unlike these previous concerns, it is not primary a problem of *expressive* limitations. Rather, it highlights an inability on the thoroughgoing dialetheist's part to demonstrate with their semantics that they are justified in rejecting the validity of unwanted argument schemas. It is one matter not being able to *express* a commitment within one's object language, and another for one's semantics to fail to *justify* the results one demands. As Weber (2022, 104) notes, while the thoroughgoing dialetheist's putative *expressive* limitations may lead to social or interpersonal problems in virtue of not being able to effectively communicate with non-dialetheists, these limitations do not themselves undermine the truth of the dialetheist's

[16] See Weber (2022, pp. 103–104) for the dissenting view, that the force of the *just-true* problem has been overstated.

[17] On the assumption, of course, that the connectives are given their usual meaning. However, given that most dialetheists explicitly desire this property of their logics (as noted in Section 2.1), this is a fair assumption to make.

thesis or their justification for holding the position. In contrast, the *countermodel* problem gets to the heart of the dialetheist's justification for their position, calling into question the thoroughgoing dialetheist's ability to justify (including to themselves) their rejection of unwanted rules of inferences, which is itself required to demonstrate their non-commitment to triviality.

In this regard, the *recapture* problem is probably closest in kind to the *countermodel* problem, for the former emphasises the inability of the dialetheist's semantics to show that we have the right to rely upon classically valid, but dialetheically invalid, rules of inference within consistent scenarios. However, it is one matter for a semantics to be at a comparative disadvantage because it is unable to license the recapture of classical validity in desirable circumstances (which is what the *recapture* problem shows), and another entirely to be unable to license the rejection of those rules of inference which call into question the viability of the proposal. The whole feasibility of the dialetheist's position relies upon the capacity of their semantics to demonstrate that they are not committed to these unwanted rules of inference, just as *any* research programme in logic is required to provide a semantics which delivers on their desired inferential commitments. This is the challenge the *countermodel* problem poses the thoroughgoing dialetheist.

Further, there's reason to think that the countermodel problem will continue to be problematic for the thoroughgoing dialetheist even if they are willing to deny the importance of the recapture problem. For instance, Weber (2022, pp. 96–102) is understandably unmoved by the latter problem, on the basis it presupposes that the thoroughoing dialetheist *should* be attempting to achieve the results already accomplished within classical mathematics. By Weber's lights, this is just another consistentist prejudice. An inconsistent mathematics should not be hamstrung by the requirement to (re-)establish everything that classical mathematics can.[18] In comparison, the *countermodel* problem makes no such presumption in favour of consistency or classical mathematics. It only relies upon the assumption that each research programme in logic should be able to demonstrate it can deliver on its own promises and desired results. In the case of the thoroughgoing dialetheist, this means showing through their semantics that they are justified in *rejecting* the unwanted inference rules that would otherwise commit them to triviality. No part of the countermodel problems conceals a presumption in favour of classicality, and thus has force against thoroughgoing dialetheism even when the recapture problem does not.

Lastly, we have the *invalidity revenge* problem presented by Young (2019), according to which the use of a glutty metatheory commits the thoroughgoing dialetheist to (at least) the invalidity of any argument schema where the conclusion contains only extensional connectives.[19] The invalidity revenge problem concerns an overgeneration of invalidities, and Weber (2024) sees this issue as solvable through the addition of a constant $\bot$ to the language defined in such a way that $r(\bot, t)$ *iff* $\bot$. This effectively blocks the construction of countermodels for a significant subset of the inferences.

The countermodel problem, on the other hand, does not concern an overgeneration of invalidities. Instead, it highlights that a proof that some inference is invalid through the construction of a countermodel fails to provide a warrant to reject the validity of the inference for a thoroughgoing

[18] Weber's view here seems markedly different from those of the other potential thoroughgoing dialetheists, such as Priest and Routley, both of whom at one time or another note the importance of not losing a "great chunk" of classical mathematics (cf. Weber, 2022, pp. 96–102). For Weber, in comparison, how much of classical mathematics should be retained is an open question.

[19] Depending upon the dialetheist's treatment of the conditional, and certain other additional semantic principles they endorse, this can be enough to ensure that they are committed to *all* argument schemas being invalid (see Young, 2019, Sect. 3). The particular details here are not important for our purposes, however.

dialetheist. In other words, the countermodel problem calls into question the epistemic value of the thoroughgoing dialetheist's countermodels, since those theorems about the existence of some model involves a paraconsistent negation expressing that the conclusion is untrue. Yet, being untrue is compatible with being true when the metatheory is dialetheic; in which case, the inference remains satisfied by the same model.

Ultimately, all four of the (putative) problems have their source in the fact that the inconsistency of the thoroughgoing dialetheist's metatheory impacts the behaviour of the logic's consequence relation. But, that is hardly a surprise. Despite this, the *countermodel* problem is notably distinct from those expressive problems already recognised in the literature. Firstly, it is fundamentally an *evidential* shortcoming on the part of the thoroughgoing dialetheist's semantics, not an expressive problem, and secondly it continues to hold weight against the thoroughgoing dialetheist's position even if the other concerns can be rejected as based upon consistentist prejudices.

Now that we have outlined this novel problem for the thoroughgoing dialetheist's current semantics, and shown how it differs from those already in the literature, it's time to sketch out a possible solution. In doing so, we'll highlight the case for the thoroughgoing dialetheist building their semantics upon the primitive acts of acceptance and rejection; a solution which should be natural for the dialetheist given that for them these acts are exclusionary, unlike the truth-values of truth and falsity.[20]

4 | BILATERALIST SEMANTICS FOR THE DIALETHEIST

4.1 | Desiderata for the solution

In order to provide a successful solution to the countermodel problem for the thoroughgoing dialetheist, it isn't enough to ensure that they are able to preclude the validity of unwanted rules of inference in their semantics. After all, as we noted at the beginning of this paper, the problem arises *because of* other commitments they wish to take on. Thus, our solution should also

[20] In this paper, we have taken on the assumption shared by most dialetheists, even those entertaining a glutty metatheory, that both the speech-acts of assertion and denial, and the related cognitive acts of acceptance and rejection, are mutually exclusive (Berto, 2014; Priest, 2006a, 2006b). Perhaps this is ultimately wrong-headed, and the thoroughgoing dialetheist ought to consider the supposed incompatibility between accepting and rejecting the same proposition as yet another residual byproduct of years of inconsistency intolerance in logic. In which case, the dialetheist could propose that a warrant to believe the falsity of a proposition sufficed to reject the proposition, and thus any warrant to believe the simultaneous truth and falsity of a proposition sufficed for its simultaneous acceptance and rejection. Let us just briefly note why we haven't considered this option for the thoroughgoing dialetheist. Firstly, without an advocate for the proposal, we are unclear how it would work exactly. For instance, how rejection in this case could be said to constitute a refusal to believe something if it is compatible with the simultaneous acceptance of the proposition. Secondly, the proposal would also need to independently address those concerns over expressive limitations which led to the introduction of these exclusionary mental acts in the first place. Again, it isn't clear to us how the proposal would achieve this. Lastly, it isn't clear that allowing for the simultaneous acceptance and rejection of a proposition would solve the *countermodel problem*. In particular, if rejection is compatible with acceptance then demonstrating that one is justified in rejecting the validity of an unwanted rule of inference would fail to preclude one's simultaneous acceptance of its validity. Yet, now we are in the same position as before—the thoroughgoing dialetheist still requires (for their own sake) assurances that they are not committed to the validity of these unwanted rules of inference. Thus, while we may ultimately be mistaken in taking on the assumption that acceptance and rejection are mutually exclusive, without detailed proposals to the contrary, it seems best for us to work within the confines of this assumption that (at least most) dialetheists currently adhere to. Many thanks to an anonymous referee for pushing us on this point.

respect these wider desiderata for the thoroughgoing dialetheist's semantics. These are: (i) that the resulting logic should be *glutty* and *paraconsistent*, (ii) that the logic should respect the normal semantics for the Boolean connectives, (iii) that the logic should require no distinction to be made between object and meta-language, and (iv) that the semantics should not produce new versions of a liar sentence which don't admit of a dialetheic solution. We call these the *background desiderata* for our solution to the problem.

4.2 | Bilateralism with provability and refutability

As our discussion from Section 3 made clear, if the dialetheist is to be able to provide countermodels within their logic, they will need to preclude or reject certain possibilities. Now, given that for the dialetheist truth-values fail to possess this preclusionary character but the speech-acts of assertion and denial, as well as the mental acts of acceptance and rejection do, it makes sense to attempt to answer the countermodel problem by couching our semantics in terms of such acts. In other words, by using a bilateralist semantics.

Bilateralism in logic is typically associated with either Smiley (1996) and Rumfitt's (2000) approach to natural deduction systems for classical logic based upon signed formulas (with one sign for acceptance and another for rejection), or Restall's (2005) interpretation of two-sided sequents, where a sequent represents a position according to which one has asserted everything in the antecedent and denied everything in the succedent. A sequent is then valid just in case asserting everything in the antecedent and denying everything in the succedent is "out of bounds"; that is, it is in some sense inappropriate.

However, rather than couching our bilateralism directly in terms of assertion and denial, or acceptance and rejection, we will favour here a semantics in terms of provability and refutability. The case for this adaptation is due to the nature of models. Models are formal abstract objects of some sort, whether constructed or real (depending upon your metaphysical leanings), and we evidence our claims about them by proving or refuting these claims. Thus, provability and refutability belong more naturally to the domain of semantics than to the domain of pragmatics or mental states, as is the case with assertion and denial, and acceptance and rejection, respectively. This makes them a more natural addition to our metatheoretic vocabulary than acceptance and rejection. Nonetheless, this choice should not obscure the fact that we take provability and refutability to be useful features of the semantics *because* they serve as suitable analogues for acceptance and rejection. In other words, we take provability to suffice for a warrant to accept, and a refutation to suffice for a warrant to reject.

The notion of refutation in philosophical logic is typically associated with refutation calculi, the aim of which are to recursively capture the set of formulas that are refutable according to some logic (Goranko et al., 2020). Our use of refutation will differ from this however, as we do not seek to present separate calculi for what is provable and what is refutable according to some standard, such as classical or paraconsistent logic. Instead, our approach is closer to research in provability logic, where it is common to treat *it is provable that* as an embeddable modal operator.[21]

[21] For an introduction to provability logic, see Verbrugge (2017). We are not aware of any corresponding research on refutability logic. However, one could perhaps understand the approach of Rosenblatt (2021) along these lines, where a sequent calculus with both sequents and "anti-sequents" is developed in order to define validity predicates for some non-classical logics using a non-classical meta-theory, as long as we understand the sequent arrow and the anti-sequent arrow as object language operators, rather than expressions of some metalangauge. Indeed, there could be an interest-

Thus, while it is common to sign formulas, or present two calculi (one for what is provable and another for what is refutable), we shall instead take the status of being refutable or provable to be representable with modal operators in the language. By treating both provable and refutable as embeddable operators, we can subsequently use them to define a notion of validity from which a notion of countermodel may be extracted. From a thoroughgoing dialetheic perspective, simply signing formulas would seem to enforce a distinction between object and meta-language which they actively resist, in line with desideratum (iii) above. By using operators on the other hand, we are in a position that we may include the metatheoretic vocabulary within the object language, and thus respect *background desideratum* (iii) by making no distinction between an object and meta-language.

4.3 | The sketch of a formal framework

To illustrate the approach, it will be useful to be slightly more formal in our presentation. We'll now sketch one way to develop a formal framework with embeddable modal operators for provability and refutability, with which countermodels can be provided for the thoroughgoing dialetheist.

As noted above, it is common to include within a glutty paraconsistent semantics an additional conditional that satisfies at least modus ponens as a rule of proof; that is, as a rule which may be applied on theorems but not assumptions.[22] We do not want to commit ourselves here to one particular such conditional, but will assume for the sake of simplicity that it satisfies the properties of BCK.[23]

We now expand the language with two unary operators $\oplus$ and $\ominus$ representing provable and refutable, respectively.[24] They should be defined in such a way that the following rules of proof are admissible:

$$\frac{A_0 \wedge \ldots \wedge A_n \rightarrow B_0 \vee \ldots \vee B_m}{\oplus A_0 \wedge \ldots \wedge \ominus B_m \rightarrow \oplus B_0 \vee \ldots \vee \ominus A_n}\,\mathrm{D}$$

$$\frac{\exists x(\oplus A_0 \wedge \ldots \wedge \oplus A_n \wedge \ominus B)}{\ominus \forall x((\oplus A_0 \wedge \ldots \wedge \oplus A_n) \rightarrow \oplus B)}\,\exists\forall \qquad \frac{A}{\oplus A}\,\mathrm{N}$$

$$\frac{}{(\oplus A \wedge \ominus A) \rightarrow B}\,\mathrm{E}$$

Being rules of proof, it is important to note that any piece of reasoning that applies these rules on assumptions, or formulas obtained from (undischarged) assumptions, may be rejected.

These rules are intended as necessary conditions for $\oplus$ and $\ominus$, not a complete characterisation of the operators. A complete characterisation would depend on the rest of the language; especially

ing connection between the approach we develop here and the system obtained by applying the approach of Rosenblatt (2021) to a paraconsistent logic, as opposed to the substructural logics focused on in that paper. However, exploring such connections goes beyond the scope of this paper.

[22] For an introduction to the notion of a rule of proof, see Humberstone (2010).

[23] This is the basic non-contractive conditional in multiplicative affine logic, also found in Cantini (2003). This assumption is in line with recent work on paraconsistent metatheory and inconsistent mathematics by Badia and Weber (2019), Weber (2022), and Badia et al. (2022).

[24] We're aware that $\oplus$ is sometimes used as a binary operator for additive disjunction in linear logic, but we do not think this will be an issue for the typical reader. Instead, we think that it would have been worse to use simply $+$ and $-$ due to their use for signed formulas by Smiley (1996) and Rumfitt (2000).

the choice of conditional, but also the exact rules for conjunction, disjunction, negation, and the quantifiers. For example, one could opt for a logic based on LP with either the BCK conditional or the conditional proposed by Beall (2009), or one could opt for a logic along the lines of that proposed by Badia et al. (2022). Each would deliver us with a bilateralist semantics that fulfils *background desideratum* (ii). At this point, we see no reason to take a stand on which of these choices is optimal (though, that isn't to say it won't turn out that one of them is).

With regard to representing provability and refutability, rule N ensures that provability tracks theoremhood, while E expresses the inappropriateness of refuting and proving the same formula. Thus, E shows that $\ominus A$ is a contrary of $\oplus A$, rather than (the truth of) A; there is no inconsistency in refuting something that happens to be true. This, we take it, suffices to ensure $\ominus A$ is not equivalent to the negation of A, and so is not a classical negation in disguise.[25]

The purpose of rule D, on the other hand, is to distribute provability and refutability operators across a conjunctive antecedent and a disjunctive consequent of a valid conditional, in a fashion familiar from a Tarskian theory of compositional truth. It can also be understood as a generalisation of contraposition for refutability. In the basic case, D permits the derivations of both $\ominus B \rightarrow \ominus A$ and $\oplus A \rightarrow \oplus B$ as theorems from $A \rightarrow B$ as a theorem.

The distribution over disjunctions in the consequent is required for the operators to match the semantic clauses adequately. After all, we would like to prove, for instance, that if it is provable that $r_I(\varphi \vee \psi, t)$ and refutable that $r_I(\varphi, t)$ then it is provable that $r_I(\psi, t)$, or alternatively that if it is provable that $r_I(\varphi \vee \psi, t)$ then it is either provable that $r_I(\varphi, t)$ or provable that $r_I(\psi, t)$. A derivation of this could look like the following which starts with one direction of the positive clause for $\vee$ in the semantics presented above in section 2.2:

$$\frac{r_I(\varphi \vee \psi, t) \rightarrow r_I(\varphi, t) \ \vee \ r_I(\psi, t)}{\oplus r_I(\varphi \vee \psi, t) \wedge \ominus r_I(\varphi, t) \rightarrow \ \oplus r_I(\psi, t)} \ \mathrm{D}$$

However, we do not want to be committed to something akin to bivalence in general for $\oplus$ and $\ominus$, so that $\oplus A \vee \ominus A$ is a theorem for every formula A. After all, we would not expect every proposition to be either accepted or rejected by an agent; one may rather suspend judgement. To avoid D implying this result for every formula A, we prefer the variant of D with a suitable restriction ensuring that the disjuncts in the conclusion-sequent must be formulas expressing decidable semantic facts. After all, analogously, we do think that a bivalence-like principle holds for acceptance and rejection under *complete information*, and thus it's equally plausible for provability and refutability over decidable fragments.

Finally, the $\exists\forall$-rule we expect to be admissible using D, E and the suitable rules for $\wedge$, $\vee$, $\forall$ and $\exists$, but we mention it explicitly here as it is crucial for our solution. It ensures we can translate the existence of a countermodel into a refutation of a claim about what is valid. We thus think of the relational atoms that assign t or f to a formula at an interpretation as a ternary relation, and our intended applications of the $\exists\forall$-rule will involve quantification into the "interpretation"-position of such relational atoms. The rule as stated is thus stronger than we require, since it allows us *in general* to transform the premise-formula into the conclusion-formula in such a way that whatever the existential quantifier bound is now bound by the universal quantifier.

This proposed semantics delivers a definition of validity according to which an inference from Γ to A is valid just in case, for every interpretation, if for every formula $B \in \Gamma$ it is provable that B is true, then it is provable that A is true. A countermodel will, therefore, be a model in which

[25] Thanks to an anonymous referee for pushing us on this point.

for every $B \in \Gamma$ it is provable that B is true, but A is refutable. In other words, switching back momentarily to talk in terms of acceptance and rejection, a model in which every member of Γ is accepted, but A is rejected. Having a countermodel for an inference then suffices with the $\exists\forall$ rule to reject the inference's validity.

However, to obtain countermodels at all, we require that there is at least one formula whose truth is refutable at some interpretation. There are at least three ways to achieve this. First, we could follow Badia et al. (2022) and stipulate that the language contains a constant $\bot$ defined in such a way that, for every interpretation, it is refutable that it is true. Second, along the lines of Carnap (1947), we could assume that for every propositional variable there is an interpretation at which it is refutable that it is true. In the current context, this corresponds to the position that trivialism is merely a possibility. Third, we could provide a formula with the desired property using $\oplus$ and $\ominus$. After all, our semantics should ultimately interpret formulae involving $\oplus$ and $\ominus$ in such a way that ensures $\oplus A \wedge \ominus A \vDash B$, given the exclusionary character of provability and refutability. Exactly how this is accomplished is beyond this paper. However, it's reasonable to expect that we would need at least one interpretation where it is refutable that $\oplus A \wedge \ominus A$ for every formula A to produce the desired result.

We won't take a stand here on which of these three options is best, but simply assume there is some formula κ with the property that its truth is refutable at an interpretation. This assumption is in no way controversial. Even the thoroughgoing dialetheist wishes to admit some sentences that ought to only be rejected, such as the aforementioned "Boris Johnson is a fried egg".

With that assumption in place, we can now show how this machinery suffices to obtain a countermodel to modus ponens for the material conditional defined as $\neg A \vee B$ (i.e. disjunctive syllogism), as the dialetheist desires. With minor modifications, this procedure can be used to provide any desired countermodel, demonstrating how the approach solves the current problem.

To simplify the notation, we'll write $T(x, y)$ for the claim that $r_x(y, t)$ and $F(x, y)$ for the claim that $r_x(y, f)$. We assume that the metalanguage has a closed term for every formula of the object language and, taking inspiration from Gödel-codes, we assume that $\ulcorner A \urcorner$ is a closed term for the formula A. The following formula expresses that there is an interpretation that refutes the truth of κ:

$$\exists x \ominus T(x, \ulcorner \kappa \urcorner)$$

For simplicity, we don't quantify explicitly over numbers naming propositional variables. Now, with λ as an instance of the standard liar, it follows that:

$$\forall x[\oplus T(x, \lambda) \wedge \oplus F(x, \lambda)]$$

From which it follows that:

$$\exists x[\oplus T(x, \lambda) \wedge \oplus T(x, \ulcorner \neg\lambda \vee \kappa \urcorner) \wedge \ominus T(x, \ulcorner \kappa \urcorner)]$$

Which subsequently implies that:

$$\ominus\forall x[((\oplus T(x, \lambda) \wedge \oplus T(x, \ulcorner \neg\lambda \vee \kappa \urcorner)) \rightarrow \oplus T(x, \ulcorner \kappa \urcorner)]$$

In other words, there is a countermodel to modus ponens and this countermodel implies that the inference is refutable. Clearly, the same approach can be used to provide countermodels for other (meta-)inference schemas that the dialetheist wishes to reject, such as explosion and

reductio. For instance, to obtain a countermodel for explosion, we simply employ the claim that $\forall x[\oplus T(x,\lambda) \wedge \oplus T(x,\neg\lambda)]$.

This suffices to show that there is an available solution to the countermodel problem that meets the thoroughgoing dialetheist's desiderata of producing a logic that (i) is glutty and paraconsistent, (ii) respects the normal semantics for the Boolean connectives, and (iii) makes no distinction between an object and meta-language. This leaves us with the requirement to show that the semantics fail to produce any new version of the liar which doesn't admit a dialetheic solution, and thus threatens the comprehensiveness of the dialetheist's proposal.

As is well known, introducing some form of "exclusionary" operator into the dialetheist's semantics buys expressive power at the cost of producing a revenge version of the liar paradox using this new operator (Berto, 2014; Martin, 2015). Yet, in the case of our bilateralist semantics there's good reason to think this won't occur, as $\ominus$ does not behave like a paracomplete negation defined as $A \rightarrow \bot$, where $\rightarrow$ is the suitable non-contractive conditional. As mentioned above, $\ominus A$ is not the contrary of A, but rather of $\oplus A$. Thus, it is not $A \wedge \ominus A$ that implies anything, but rather $\oplus A \wedge \ominus A$. For a bilateralist, this is precisely how things should be: the contrary of rejecting A is not the truth of A but rather the acceptance of A.[26] From a perspective based on provability and refutability, this also seems correct: while A could be true even if it is refutable, A being provable and refutable should make your logic explode, metaphorically speaking. Further, this property of the semantics captures the dialetheist's thought that one cannot both accept and reject a proposition (simultaneously), which is what provides these acts with the exclusionary character that truth-values fail to possess.

Finally, and importantly, the standard mechanism by which one produces paradoxical cases is blocked precisely because it is the combination of a proposition's provability and refutability which explode, and not the combination of being true and refutable. In the latter case, one would proceed by turning a negated formula into a truth with which it is equivalent, and then contract the two copies of the same truth into one. In contrast, with the semantics presented here, using this mechanism merely produces the truth of a refutation and the acceptance of what is being refuted, and so there is nothing to contract. This suffices to handle any potential revenge "refutation"-liar such as:

(R) it is refutable that *R* is true

However, what about *being unprovable*? Considering the underlying bilateralism, it is natural to define unprovable as simply the negation of being provable, and not in terms of the *sui generis* property of being refutable. Will unprovable then become a dialetheia, as in the approach to paraconsistent metatheory sketched by Weber (2024)? While this certainly depends on the finer details of the proposal, our preliminary answer is that this shouldn't be the case. Let U be an "unprovable"-liar: "it is not provable that U is true", and assume that "it is not provable that U is true or it is provable that U is true" is derivable. With transparent truth and the proposed rules for being provable, we can either say that the entire sentence is provable, or distribute the provability operator across each disjunct to obtain "it is provable that U is true or it is provable that it is provable that U is true". Yet the latter fails to entail "it is provable that U is true or it is provable that U is true", from which the contradiction would follow.

[26] Of course, if $A \wedge \ominus A$ is a theorem then $\oplus A \wedge \ominus A$ is also a theorem, but this does not imply explosion involving $\ominus$, i.e. $A \wedge \ominus A \rightarrow B$. One shouldn't mix reasoning from theorems with reasoning from assumptions. Correspondingly, we are not committed to the factivity of provability, that is, $\oplus A$ implies A; this should, of course, only be the case if $\oplus A$ is a theorem.

In the same vein, one could speculate upon whether $\oplus A \vee \ominus A$ as a theorem would be problematic and lead to contradiction in the same way as $A \vee \neg A$. The simple answer to that is "no". Let's take the above sentence (R) as an example. Then the relevant instance would be $\oplus \mathrm{Tr}(\ulcorner R \urcorner) \vee \ominus \mathrm{Tr}(\ulcorner R \urcorner)$. Applying transparent truth gives us now $\oplus \mathrm{Tr}(\ulcorner R \urcorner) \vee \mathrm{Tr}(\ulcorner R \urcorner)$. However, from this point there is not much more we can do beyond distributing further layers of $\oplus$ and $\ominus$ across both formulas, but there will not be an even number of $\oplus$ or $\ominus$ to justify a contraction.

Another example, equally instructive, is the explosive reasoning starting out from a reductio rule according to which if $A \rightarrow \ominus A$ is a theorem then so is $\ominus A$. This would allow us to derive anything with a liar-like sentence involving $\ominus$. However, the admissibility of this rule does not follow from our given rules D, $\forall\exists$, N and E. Instead, applying D on $A \rightarrow \ominus A$ delivers (for instance) $\oplus A \rightarrow \oplus\ominus A$ or $\ominus\ominus A \rightarrow \ominus A$. With an unrestricted D, we could have obtained $\ominus A \vee \oplus\ominus A$ from $A \rightarrow \ominus A$. But, that is still a far cry from the troublesome $\ominus A \vee \ominus A$ which we could contract into $\ominus A$, leading us to paradox.[27]

Consequently, there's good reason to think our solution also meets the thoroughgoing dialetheist's desideratum (iv).[28]

4.4 | Alternatives to bilateralism

As we see it, the thoroughgoing dialetheist has two live options with which to provide a semantics that delivers countermodels to those inferences they desire to reject, and so solve the countermodel problem. Either they can (i) go down a bilateralist route, such as that sketched here, or (ii) use an incompatibility operator route along the lines of Berto (2014), where the resulting operator is a contrary-forming device. We end with a few words on why we prefer the former route over the latter.

Whichever tools one employs to define and justify the incompatibility route, the key will be obtaining a statement representing a countermodel for an inference that suffices for a warrant to reject the inference. A contrary-forming operator like $A \rightarrow \bot$ could certainly both imply the desired warrant and be used for such a purpose. However, a countermodel cannot then be a formula of the form $A \wedge (B \rightarrow \bot)$, since assuming this is a theorem does not suffice for deriving $(A \rightarrow B) \rightarrow \bot$. Instead, we would need an intensional (i.e. multiplicative) conjunction for that purpose. The "substructural" paraconsistent logic recently presented by Badia et al. (2022) would seem to be a better fit, given that it includes such a connective and we can even show that the truth of $A\&(B \rightarrow \bot)$ implies the truth of $(A \rightarrow B) \rightarrow \bot$.

[27] We thank an anonymous referee for suggesting we consider these instructive cases.

[28] Interestingly, one can show with a classical metatheory the non-triviality of the theory obtained by expanding a paraconsistent theory of truth based on LP and a basic non-contractive conditional with the operators $\oplus$ and $\ominus$ satisfying the rules above, assuming the non-triviality of the original paraconsistent theory. We sketch a partial cut-elimination proof in the appendix. The proof establishes the conservativeness of the rules for $\oplus$ and $\ominus$ with regard to the underlying theory. This tells us that, looking at the resulting theory from a classical perspective, such a representation of provable and refutable does *not* introduce any revenge paradox. It follows that, as far as the classical perspective goes, our approach is revenge-free since assuming the derivability of a revenge paradox is inconsistent with the partial cut-elimination proof. While this result will please a reader that still prefers their classical metatheory, a dialetheically minded reader is, of course, free to find the result insufficient given its reliance on a classical metatheory. In any case, we think it's safe to deny the possibility of a rejection liar involving the currently formalised concepts in the system. Of course, one can always think of further principles that one might like regarding provability or refutability for expressive reasons, which when added to the current package reintroduces triviality. We can't rule this out. However, at present, we have good reason to think the proposal can handle all liars it can express.

However, we are hesitant to go down this path just yet, given that the multiplicative fragment of the logic presented by Badia et al. (2022) can be shown to be ω-inconsistent (Fjellstad, 2024). Now, perhaps ω-inconsistency is a result that the thoroughgoing dialetheist ought to embrace, because it is necessary or at least useful for certain model-theoretic purposes (as illustrated by Badia et al., 2022), but we don't know that yet. Thus, at present, it's best to recommend caution even if it would certainly fit the label of "thoroughgoing" to embrace ω-inconsistency.

In order for an approach like Badia et al. (2022) to avoid ω-inconsistency, we would need a working theory of multiplicative quantification that doesn't introduce sufficient contraction to introduce ω-inconsistency. Unfortunately, the current approach to multiplicative quantification by Zardini (2011) is clearly unsuitable for such purposes.[29] Thus, on the one hand, we take the proposed multiplicative strengthening of the additive quantifier by Badia et al. (2022) as indicating that any serious metatheoretical reasoning involving a multiplicative conjunction requires also a quantifier that is, to some extent, more multiplicative than the additive/extensional quantifier. On the other hand, we take the problems that arise from that strengthening to show that any solution to the countermodel problem along these lines requires formal tools that are still under development.

Of course, a proponent of an incompatibility operator can also choose to work with the logic presented by Weber (2024), in which the quantifiers haven't been strengthened in the same "multiplicative" way. Assessing whether that will be satisfactory is beyond the scope of this brief discussion, but we note that $\bot$ is not assigned such a role in the semantics presented by Weber (2024).

Both the bilateralist and incompatibility operator routes are open possibilities that should be further developed. However, as we note above, in ultimately choosing among these proposals it is not enough that they solve the problem at hand, but that they are in keeping with the thoroughgoing dialetheist's additional background commitments. In that regard, the bilateralist proposal is motivated by the long-standing commitment among (some) glut theorists that whereas truth-bearers can simultaneously be true and false, it is a physical impossibility to jointly assert(/accept) and deny(/reject) a claim.

Of course, this long-standing commitment may ultimately be ill-advised for the thoroughgoing dialetheist, but again *prima facie* it seems beneficial to have a solution to the problem which is underpinned by a prevalent commitment among the group for whom the problem is pertinent. Ultimately, the goal of the solution to the countermodel problem sketched out in this section has been to instigate the debate over what would constitute a good solution to the problem for the thoroughgoing dialetheist. We don't intend it to be the final word on the matter, and we look forward to further developments.

5 | CONCLUSION

In this paper, we have presented a new justificatory problem for the thoroughgoing dialetheist, who wishes to embrace a glutty paraconsistent metatheory. According to the *countermodel problem*, the thoroughgoing dialetheist is unable using their existent semantics to provide countermodels which sanction the rejection of unwanted rules of inference. In this sense, their current semantics fail to deliver on the justificatory burden we require of our semantics. In response, we have proposed that one fruitful means to address this problem is by embracing a bilateralist

[29] See Fjellstad and Olsen (2021) for a discussion of the problems with the approach in Zardini (2011).

semantics in terms of provability and refutability, treating these as analogues for the mental acts of acceptance and rejection, respectively, which even the dialetheist agrees are exclusionary, unlike truth and falsity.

ACKNOWLEDGEMENTS

We are grateful to Jc Beall and an anonymous referee for their useful comments on a previous version of this paper. Research for this paper was supported by a PNRR grant, under the European Union's NextGenerationEU research and innovation programme.

Open access publishing facilitated by Universita degli Studi di Padova, as part of the Wiley - CRUI-CARE agreement.

ORCID

Andreas Fjellstad https://orcid.org/0000-0002-3239-4484
Ben Martin https://orcid.org/0000-0003-2280-0781

REFERENCES

Allo, P. (2010). A classical prejudice? *Knowledge, Technology & Policy*, *23*, 25–40. https://doi.org/10.1007/s12130-010-9098-4

Asenjo, F. G. (1966). A calculus of antinomies. *Notre Dame Journal of Formal Logic*, *7*, 103–105. https://doi.org/10.1305/ndjfl/1093958482

Badia, G., & Weber, Z. (2019). A substructural logic for inconsistent mathematics. In A. Rieger & G. Young (Eds.), *Dialetheism and its applications* (pp. 155–176). Springer.

Badia, G., Weber, Z., & Girard, P. (2022). Paraconsistent metatheory: New proofs with old tools. *Journal of Philosophical Logic*, *51*, 825–856. https://doi.org/10.1007/s10992-022-09651-x

Batens, D., & Clercq, K. D. (2004). A rich paraconsistent extension of full positive logic. *Logique et Analyse*, *47*, 227–57.

Beall, J. (2001). Dialetheism and the probability of contradictions. *Australasian Journal of Philosophy*, *79*, 114–8. https://doi.org/10.1080/713659182

Beall, J. (2009). *Spandrels of truth*. Oxford University Press.

Beall, J. (2013a). Shrieking against gluts: the solution to the 'just true' problem. *Analysis*, *73*, 438–45. https://doi.org/10.1093/analys/ant057

Beall, J. (2013b). A simple approach towards recapturing consistent theories in paraconsistent settings. *The Review of Symbolic Logic*, *6*, 755–64. https://doi.org/10.1017/s1755020313000208

Beall, J. (2014). End of inclosure. *Mind*, *123*, 829–49. https://doi.org/10.1093/mind/fzu075

Beall, J. (2021). *The contradictory christ*. Oxford University Press.

Beall, J. (2022). Review of Zach Weber, Paradoxes and Inconsistent Mathematics. *Notre Dame Philosophical Reviews*. https://ndpr.nd.edu/reviews/paradoxes-and-inconsistent-mathematics/

Berto, F. (2014). Absolute contradiction, dialetheism, and revenge. *Review of Symbolic Logic*, *7*, 193–207. https://doi.org/10.1017/s175502031400001x

Burgess, J. P. (1983). Common sense and "relevance". *Notre Dame Journal of Formal Logic*, *24*, 41–53. https://doi.org/10.1305/ndjfl/1093870219

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to id. *Journal of Symbolic Logic*, *55*, 244–259. https://doi.org/10.2307/2274965

Cantini, A. (2003). The undecidability of grisin's set theory. *Studia Logica*, *74*, 345–368. https://doi.org/10.1023/a:1025159016268

Carnap, R. (1947). *Meaning and necessity: A study in semantics and modal logic*. University of Chicago Press.

da Costa, N. C. A. (1982). On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic*, *15*, 497–510. https://doi.org/10.1305/ndjfl/1093891487

Dunn, J. M. (1976). Intuitive semantics for first-degree entailments and 'coupled trees'. *Philosophical Studies*, *29*, 149–168. https://doi.org/10.1007/BF00373152

Fjellstad, A. (2022). Expressing logical disagreement from within. *Synthese*, *200*, 1–33. https://doi.org/10.1007/s11229-022-03667-1

Fjellstad, A. (2024). *A multiplicative ingredient for omega-inconsistency*. (Unpublished manuscript)

Fjellstad, A., & Olsen, J.-F. (2021). Ikt$^{\omega}$ and łukasiewicz-models. *Notre Dame Journal of Formal Logic*, *62*, 247-256. https://doi.org/10.1215/00294527-2021-0012

Goranko, V., Pulcini, G., & Skura, T. (2020). Refutation systems: An overview and some applications to philosophical logics. In F. Liu, H. Ono, & J. Yu (Eds.), *Knowledge, proof and dynamics* (pp. 173–197). Springer.

Humberstone, L. (2010). Smiley's distinction between rules of inference and rules of proof. In T. J. Smiley, J. Lear, & A. Oliver (Eds.), *The force of argument: Essays in honor of timothy smiley* (pp. 107–126). Routledge.

Jennings, R. E., & Schotch, P. K. (1984). The preservation of coherence. *Studia Logica*, *43*, 89–106. https://doi.org/10.1007/BF00935743

Jenny, M. (2017). Classicality lost: K3 and lp after the fall. *Thought*, *6*, 43–53. https://doi.org/10.1002/tht3.231

Littmann, G., & Simmons, K. (2004). A critique of dialetheism. In G. Priest, J. Beall, & B. Armour-Garb (Eds.), *The law of non-contradiction: New philosophical essays* (pp. 314–35). Oxford University Press.

Martin, B. (2015). Dialetheism and the impossibility of the world. *Australasian Journal of Philosophy*, *93*, 61–75. https://doi.org/10.1080/00048402.2014.956768

Martin, B. (2021a). Identifying logical evidence. *Synthese*, *198*, 9069–95. https://doi.org/10.1007/s11229-020-02618-y

Martin, B. (2021b). Searching for deep disagreement in logic: The case of dialetheism. *Topoi*, *40*, 1127–38. https://doi.org/10.1007/s11245-019-09639-4

Martin, B., & Hjortland, O. T. (2021). Logical predictivism. *Journal of Philosophical Logic*, *50*, 285–318. https://doi.org/10.1007/s10992-020-09566-5

Meadows, T. (2015). Unpicking Priest's bootstraps. *Thought*, *4*, 181-188. https://doi.org/10.1002/tht3.172

Metcalfe, G., Olivetti, N., & Gabbay, D. M. (2008). *Proof theory for fuzzy logics*. Springer.

Negri, S. (2005). Proof analysis in modal logic. *Journal of Philosophical Logic*, *34*, 507–544. https://doi.org/10.1007/s10992-005-2267-3

Nicolai, C. (2021). Cut elimination for systems of transparent truth with restricted initial sequents. *Notre Dame Journal of Formal Logic*, *62*, 619–642. https://doi.org/10.1215/00294527-2021-0032

Omori, H., & Weber, Z. (2019). Just true? on the metatheory for paraconsistent truth. *Logique et Analyse*, *248*, 415–433. https://doi.org/10.2143/LEA.248.0.3287323

Parsons, T. (1990). True contradictions. *Canadian Journal of Philosophy*, *20*, 335–54.

Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic*, *8*, 219–241. https://doi.org/10.1007/BF00258428

Priest, G. (1990). Boolean negation and all that. *Journal of Philosophical Logic*, *19*, 201–215. https://doi.org/10.1007/bf00263541

Priest, G. (1995). *Beyond the limits of thought*. Cambridge University Press.

Priest, G. (2006a). *Doubt truth to be a liar*. Oxford University Press.

Priest, G. (2006b). *In contradiction: A study of the transconsistent (second editon)*. Oxford University Press.

Priest, G., & Routley, R. (1989). Systems of paraconsistent logic. In G. Priest, R. Routley, & J. Norman (Eds.), *Paraconsistent logic: Essays on the inconsistent* (pp. 142–155). Philosophia Verlag.

Restall, G. (1999). *An introduction to substructural logics*. Routledge.

Restall, G. (2005). Multiple conclusions. In P. Hájek, L. Valdés-Villanueva, & D. Westerståhl (Eds.), *Logic, methodology and philosophy of science*. College Publications.

Rosenblatt, L. (2021). Towards a non-classical meta-theory for substructural approaches to paradox. *Journal of Philosophical Logic*, *50*, 1007-1055. https://doi.org/10.1007/s10992-020-09589-y

Routley, R. (1977a). Ultralogic as universal? (part i). *Relevance Logic Newsletter*, *2*, 51–90.

Routley, R. (1977b). Ultralogic as universal? (part ii). *Relevance Logic Newsletter*, *2*, 138–175.

Routley, R. (1979). Dialectical logic, semantics and metamathematics. *Erkenntnis*, *14*, 301–331. https://doi.org/10.1007/bf00174897

Routley, R. (1980). The choice of logical foundations: Non-classical choices and the ultralogical choice. *Studia Logica*, *39*, 77–98. https://doi.org/10.1007/bf00373098

Rumfitt, I. (2000). Yes and no. *Mind*, *109*, 781–823. https://doi.org/10.1093/mind/109.436.781

Schröder-Heister, P. (2016). Restricting initial sequents: The trade-offs between identity, contraction and cut. In R. Kahle, T. Strahm, & T. Studer (Eds.), *Advances in proof theory, progress in computer science and applied logic 28*. Birkhauser.

Shapiro, S. (2004). Simple truth, contradiction, and consistency. In G. Priest, J. Beall, & B. Armour-Garb (Eds.), *The law of non-contradiction: New philosophical essays* (pp. 336–54). Oxford University Press.
Smiley, T. (1996). Rejection. *Analysis*, *56*, 1–9. https://doi.org/10.1111/j.0003-2638.1996.00001.x
Tanaka, K., & Girard, P. (2023). Against classical paraconsistent metatheory. *Analysis*, *83*, 285–294. https://doi.org/10.1093/analys/anac093
Tennant, N. (2004). An anti-realist critique of dialetheism. In G. Priest, J. Beall, & B. Armour-Garb (Eds.), *The law of non-contradiction: New philosophical essays* (pp. 355–384). Clarendon Press.
Verbrugge, R. (2017). Provability logic. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2017/entries/logic-provability/
Weber, Z. (2010). Transfinite numbers in paraconsistent set theory. *Review of Symbolic Logic*, *3*, 71–92. https://doi.org/10.1017/s1755020309990281
Weber, Z. (2022). *Paradoxes and inconsistent mathematics*. Cambridge University Press.
Weber, Z. (2024). True, untrue, valid, invalid, provable, unprovable. *Logic and Logical Philosophy*, Advanced online publication. https://doi.org/10.12775/llp.2024.008
Weber, Z., Badia, G., & Girard, P. (2016). What is an inconsistent truth table? *Australasian Journal of Philosophy*, *94*, 533–548. https://doi.org/10.1080/00048402.2015.1093010
Williamson, T. (2017). Semantic paradoxes and abductive methodology. In B. Armour-Garb (Ed.), *The relevance of the liar* (pp. 325–46). Oxford University Press.
Young, G. (2015). Shrieking, just false and exclusion. *Thought*, *4*, 269–276. https://doi.org/10.1002/tht3.187
Young, G. (2019). A revenge problem for the dialetheist. In A. Rieger & G. Young (Eds.), *Dialetheism and its applications* (pp. 21–46). Springer.
Zardini, E. (2011). Truth without contra(di)ction. *Review of Symbolic Logic*, *4*, 498–535. https://doi.org/10.1017/s1755020311000177

How to cite this article: Fjellstad, A., & Martin, B. (2024). Dialetheism and the countermodel problem. *Philosophy and Phenomenological Research*, 1–25. https://doi.org/10.1111/phpr.13130

APPENDIX

The aim of this appendix is to show that the approach sketched above in subsection 4.3 is non-trivial if defined within a classical metatheory. To that purpose we will assume that we are working with a sequent calculus where sequents are pairs of sets for LP expanded with a suitable conditional and a transparent truth predicate. The truth-predicate Tr where $\ulcorner A \urcorner$ is a closed term functioning as a name for the formula A satisfies the following rules:

$$\frac{A, \Gamma \Rightarrow \Delta}{Tr(\ulcorner A \urcorner), \Gamma \Rightarrow \Delta} \quad \frac{\Gamma \Rightarrow \Delta, A}{\Gamma \Rightarrow \Delta, Tr(\ulcorner A \urcorner)}$$

The conditional satisfies the following rules:

$$\frac{\Rightarrow A \quad B \Rightarrow \Delta \qquad C \Rightarrow A \quad B \Rightarrow D}{\Gamma, A \to B \Rightarrow \Delta \quad \Gamma, A \to B \Rightarrow C \to D, \Delta} \qquad \frac{A \Rightarrow B}{\Gamma \Rightarrow A \to B, \Delta}$$

We could have increased the complexity of the sequents in order to get a more interesting conditional, for example by using hypersequents as in Metcalfe et al. (2008), labelled formulas inspired by Negri (2005), or let the sequents be pairs of structures along the lines of Restall (1999). This would merely add noise with regard to the proof strategy presented here since the rules for $\oplus$ and $\ominus$ presented below can be generalised to such settings in ways that are somehow obvious.

For our purposes, it suffices to establish that the addition of suitable rules for $\oplus$ and $\ominus$ is a conservative extension. This can be achieved with a partial cut-elimination theorem. With the partial cut-elimination proof we show that, if a sequent is obtained with a derivation containing cuts on subderivations that involve applications of the rule for $\oplus$ and $\ominus$, then that sequent can also be obtained with a derivation in which the cuts are applied on subderivations that do not contain applications of the rule for $\oplus$ and $\ominus$. In effect, we show how to transform every derivation with cuts after applications of the rule for $\oplus$ and $\ominus$ to a derivation where each cut is before the applications of those rules. The proof sketch is based primarily on the cut-elimination proofs presented by Cantini (1990) and Nicolai (2021).

As regards the sequent calculus, we will assume that it is expanded with the following initial sequents and rules. If A is of the form $Tr(\ulcorner B \urcorner)$, $\oplus B$, or $\ominus B$, then the following are initial sequents:

$$A, \Gamma \Rightarrow \Delta, A$$

We also expand the calculus with the following rule:

$$\frac{A_0, \ldots, A_n \Rightarrow B_0, \ldots, B_m}{\Gamma, \oplus A_0, \ldots, \ominus B_m \Rightarrow \oplus B_0, \ldots, \ominus A_n, \Delta} D'$$

where every formula in the premise-succedent may be "$\ominus$-ed" into the conclusion-antecedent, and every formula in the premise-antecedent may be "$\ominus$-ed" into the conclusion-succedent. The special case where there is only one formula in the premise-succedent is also permitted.

We now proceed to sketch how to prove that applications of the following rule can be partially eliminated from a derivation:

$$\frac{\Gamma \Rightarrow \Delta, A \qquad A, \Gamma' \Rightarrow \Delta'}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'} \text{ cut}$$

To that purpose, we define two measures on a derivation in this sequent calculus: *modal depth* and *truth depth*.

Based on Cantini (1990), the modal depth of a derivation d is defined inductively as follows: if d is an initial sequent, the modal depth of that sequent is 0. If d is obtained from derivations d_i, where $0 < i < j$ with any rule except D', then the modal depth of d is the supremum of the modal depths of d_i for each $i < j$. If d is obtained from d' with D', then the modal depth of d is that of d' plus 1.

Based on Schröder-Heister (2016) and Nicolai (2021), the truth depth of a formula A relative to the position (antecedent or succedent) in the endsequent of a derivation d is defined inductively as follows: if d ends with an initial sequent, then every formula in the sequent has the truth-depth 0. If d is obtained from d' with an application of a truth rule, then the truth-depth of the principal formula is 1 plus the maximum of the truth-depth of the active formula in d' and the possible copy of the principal formula already occurring in that position of the sequent in d'. The other formulas keep their truth-depth from d'. If d is obtained with some other rule, then the truth-depth of the principal formula is the supremum of the truth-depth of the active formulas and the possible copy of the principal formula already occurring in that position in d'. The other formulas keep their truth-depth from d' and contracting formulas get the supremum of each copy.

We now claim that the following holds:

> If there is a derivation of $\Gamma \Rightarrow \Delta$ ending with an application of cut with modal depth $m > 0$, then there is a derivation of $\Gamma \Rightarrow \Delta$ with modal depth $n < m$.

Proof is by induction on the modal depth of the derivation, with subinductions on the truth depth of a formula, the complexity of a formula, and the cut-height. The latter two measures are defined in the standard way.

Since most of the cases are standard, and details can be found in Negri (2005), Nicolai (2021) and Fjellstad (2022), we focus here on the single case that distinguishes this proof from the full cut-elimination proofs presented by Cantini (1990) and Nicolai (2021) (for significantly different consistent theories of truth). Consider the following (slightly simplified) cut where the left premise is an initial sequent and the modal depth of the right premise is greater than zero:

$$\frac{Tr(\ulcorner A\urcorner) \Rightarrow Tr(\ulcorner A\urcorner) \quad \frac{Tr(\ulcorner A\urcorner), A \Rightarrow \Delta}{Tr(\ulcorner A\urcorner) \Rightarrow \Delta}}{Tr(\ulcorner A\urcorner) \Rightarrow \Delta}$$

Given that the truth depth of $Tr(\ulcorner A\urcorner)$ as it occurs in the right premise is greater than the truth depth of $Tr(\ulcorner A\urcorner)$ as it occurs in the antecedent of the left premise, we cannot simply "eliminate" this cut by just keeping the left derivation. Had we rejected initial sequents of that form, then this case would not have been a problem and we could in fact have concluded consistency along the lines of Nicolai (2021). But then our theory of truth would have been non-reflexive.

Instead, we accept that we will only obtain a partial cut-elimination, and in particular we shall only push cuts until the modal depth is reduced to zero. Now, the calculus is such that $A \Rightarrow A$ is derivable with modal depth 0 and truth depth 0 for every formula A. Consider now the following transformation of the above case:

$$\dfrac{\dfrac{A \Rightarrow A \qquad A, Tr(\ulcorner A\urcorner) \Rightarrow \Delta}{A, Tr(\ulcorner A\urcorner) \Rightarrow \Delta}}{Tr(\ulcorner A\urcorner) \Rightarrow \Delta}$$

This transformation may increase the complexity of the cut formula and the cut-height, but does not increase the modal depth of the end sequent and decreases the truth depth of the cut formula. Consider now the following case involving not truth but $\oplus$:

$$\frac{\oplus A \Rightarrow \oplus A \quad \frac{A \Rightarrow \Delta}{\oplus A \Rightarrow \oplus \Delta}}{\oplus A \Rightarrow \oplus \Delta}$$

We can transform it into the following:

$$\dfrac{\dfrac{A \Rightarrow A \qquad A \Rightarrow \Delta}{A \Rightarrow \Delta}}{\oplus A \Rightarrow \oplus \Delta}$$

In this case, we have decreased the modal depth of the cut. With this observation, we conclude that we can push each application of cut up beyond applications of D'. D' is a conservative extension of the underlying sequent calculus. It follows that if a theory defined with a sequent calculus satisfying the conditions assumed in this appendix is trivial, then that triviality result is obtainable without the D'-rule.