Philos Stud (2021) 178:2759–2784
https://doi.org/10.1007/s11098-020-01579-9

Sets, lies, and analogy: a new methodological take

Giulia Terzian[1]

Accepted: 14 October 2020 / Published online: 3 November 2020

Abstract The starting point of this paper is a claim defended most famously by Graham Priest: that given certain observed similarities between the set-theoretic and the semantic paradoxes, we should be looking for a 'uniform solution' to the members of both families. Despite its indisputable surface attractiveness, I argue that this claim hinges on a problematic reasoning move. This is seen most clearly, I suggest, when the claim and its underlying assumptions are examined by the lights of a novel, quite general and, I contend, promising take on inter-theoretic analogy. The ensuing discussion is intended to serve as both a possible case study and a first step towards the broader aim of the paper: namely, to initiate a wider conversation on the methodology of paradox-solving on the one hand, and the use of inter-theoretic analogies on the other.

Keywords Inter-theoretic analogy · Moderate naturalism · Truth-theoretic desiderata · Solutions to the paradoxes · Philosophical methodology

1 Overview

A little over two decades ago, Graham Priest argued for the claim that, on the basis of certain observed structural similarities between the set-theoretic and the semantic paradoxes, we should be looking for a 'uniform solution' to the members of both families.[1] Given that the set-theoretic paradoxes have been solved and the semantic

[1] See Priest (1994) and Priest (1995). Some follow-up discussion can be found in Priest (2000) and Smith (2000).

✉ Giulia Terzian
giulia.terzian@gmail.com

[1] Honorary Research Associate, Department of Philosophy, University of Bristol, Bristol, UK

paradoxes have not, the potential significance of this claim can be neither overlooked nor underestimated.[2] And yet there has been relatively little discussion of whether the claim in question, and the argumentative move it appears to sanction, are plausible ones. I will in fact argue that Priest's analogy-based move is methodologically problematic, and that as a result, his unification effort does not succeed—as it stands. Part of the problem, I will suggest, is that the claim is not accompanied by—and more generally, the literature lacks—a sustained and collective discussion, either of what counts as a solution to a paradox; or of how far we should be willing to go in order to secure one.

This partially negative result will then be used as a springboard for the more constructive part of the paper, where I will offer an alternative perspective on the methodology of paradox-solving. While most of my observations will revolve around the specific case of the semantic vs. set-theoretic paradoxes, the discussion will also speak to the wider question of what it takes to 'solve' a paradox, in any domain. Regarding the paper's main case study, I will maintain that so far, the focus has been on the wrong question. That is, rather than concentrating on finding a putative uniform solution to the paradoxes, we should be asking: 'Do we want the same things from our theories of sets and truth?'

After canvassing a small section of the relevant literature I draw the preliminary conclusion that as far as current wisdom goes, the answer to this question is negative. While it remains to be seen whether these circumstances will change, I will however maintain that some good already comes out of all this: namely, the seeds of a methodologically more robust approach to inter-theoretic analogies in general.

The discussion will unfold as follows. Section 2 provides some basic and well-rehearsed background on the set-theoretic and the truth-theoretic paradoxes. Section 3 turns to Priest's analysis of the paradoxes in Priest (1994) and his proposed 'principle of uniform solution'. There I present and defend the first main claim of the paper: that the principle, while very attractive on the surface, is in fact much less straightforward than Priest (and others proposing similar lines of argument) takes it to be. To do so, I unpack the principle into its several components and put forward one possible reformulation of the same.[3] Section 4 sketches an alternative approach to the question of whether a 'uniform solution' to the paradoxes can and should be sought. More specifically I will argue that focusing on the *concepts* of set and of truth, and their respective *theories*, affords a more illuminating and better motivated perspective on any potentially interesting similarities between the two domains. I then lay out one way to assess the success of an inter-theoretic analogy, the gist of which is: look at the desiderata underlying

[2] This claim is coarse-grained and imprecise, of course; it is nonetheless accurate insofar as there exist theories of sets, and there exist theories of truth, that are not susceptible to paradox. Only the former are also regarded as satisfactory and successful *over and above* their ability to steer clear of inconsistency, however. I will return to this important observation later on.

[3] To reiterate: I make no presumption that my analysis is the best nor certainly the only one possible. On the contrary, I hope to spark a discussion on the question of what it takes—and how far we should be willing to go—to solve a paradox, in this and any other domain.

the respective theories. Sections 4.1 and 4.2 provide an illustration of this proposal at work in a different domain. I then revisit Priest's 'uniform solution' principle by the lights of the proposed approach, in Sect. 4.3. Finally, Sect. 5 offers a summary of the discussion and some forward-looking remarks.

2 Naïve theories

By way of preamble, the first thing to note is that the discussion will be largely non-technical. Thus, the formal background will be left in its rightful place—namely, the background. This option is available to us because the technical framework that we will be presupposing is entirely standard and well-rehearsed in the literature; and it is not what is at issue here. Accordingly, let $\mathcal{L}$ be a first-order language containing the usual connectives ($\neg$, $\vee$), quantifiers ($\exists$), variable and constant terms (x, y, ...; a, b...) and predicate letters (P, Q, ...), and abiding the usual rules of syntactic composition. I now consider two familiar extensions of $\mathcal{L}$ and briefly recall the details of two very natural theories for the resulting languages.[4]

First, we extend the base language by adding the monadic predicate 'is true'—henceforth *tr*. The resulting language $\mathcal{L}_{tr}(:= \mathcal{L} \cup \{tr\})$ now has the syntactic resources to allow for the construction of self-referential expressions, where the latter may also contain the truth predicate. Suppose then that $\mathcal{L}_{tr}$ is our language of truth. In a series of hugely influential papers, Alfred Tarski showed that the simplest, most natural theory for any such language is inconsistent.[5] The theory in question—known as the *naïve theory of truth*—can be captured schematically as follows:

(NT1) The language $\mathcal{L}_{tr}$ contains the names of all of its expressions.
(NT2) The theory proves the unrestricted T-scheme. That is, for all sentences $\phi \in \mathcal{L}_{tr}$:
(T) $tr(\ulcorner \phi \urcorner) \leftrightarrow \phi$[6]
(NT3) Classical logic holds.

Tarski argued that satisfaction of (NT1–NT3) is the *sine qua non* of a truth theory's descriptive accuracy: a predicate failing to satisfy all three would just not *be* the truth predicate. He also proved that liar-like paradoxes can be derived in NT, and thus that (NT1–NT3) are "the primary source of all semantical antinomies [...]. These antinomies seem to provide a proof that every language which is universal in the above sense, and for which the normal laws of logic hold, must be inconsistent" (Tarski 1936, 164).

Now let us extend $\mathcal{L}$ with a two-place predicate $\in$ for set-membership—i.e., the predicate 'is a member of'. The resulting language $\mathcal{L}_{\in}$ has the same syntactic

[4] To reiterate: nothing new is being done here. See for instance Leitgeb (2004) and Feferman (1984).

[5] See e.g. Tarski (1936) and Tarski (1944).

[6] Following standard practice, I assume that a suitable coding function is available in the background, mapping the language of truth to $\mathcal{L}_{PA}$. Hereafter, gödel corners will be dropped.

resources as $\mathcal{L}_{tr}$; in particular, it contains self-referential expressions. And here once again a very simple and natural theory of set-membership is available—namely, the *naïve theory of sets*:[7]

(NS1) The language $\mathcal{L}_{\in}$ contains the names of all of its expressions.

(NS2) The theory proves the unrestricted comprehension scheme. That is, for all $\phi \in \mathcal{L}_{\in}$:

$$\text{(CA) } \exists x \forall y (y \in x \leftrightarrow \phi(y))$$

(NS3) Classical logic holds.

Together, (NS1–NS3) are sufficient to allow the derivation of familiar set-theoretic paradoxes such as Russell's paradox; like its truth-theoretic counterpart, in other words, NS is inconsistent. In both cases, the discovery of inconsistency has prompted mathematicians, logicians and philosophers alike to revisit the naïve conceptions of truth and set-membership, and to take a more cautious approach in the quest for theories of these notions that are natural, simple, successful (in a broadly naturalistic sense of the term) and, of course, consistent.

3 How to solve a paradox

Thanks to Tarski, we know that 'naïve' theories of other semantic concepts such as denotation, definability and satisfaction can be produced along much the same lines as NT, from which counterparts of the liar paradox can be derived.[8] It is hardly surprising, then, that some authors should have endeavoured to reach even further beyond the (ultimately rather scarce) facts at our disposal, in search of deeper and more general lessons about the paradoxes. One such attempt was made by Frank P. Ramsey, who claimed that the paradoxes involving semantic concepts should be separated from those involving "only logical or mathematical terms such as class and number" (Ramsey 1925, 353), i.e. pertaining to the domain of set theory. Accordingly, Ramsey maintained, the respective disciplines should be kept strictly apart.

Ramsey's classification has since been called into question, mainly on the charge that it neglects certain rather striking commonalities between the two families of paradoxes. Standing out most prominently among Ramsey's critics are Bertrand Russell and, almost a century later, Graham Priest. Following in Russell's footsteps, Priest argues in several places that "all paradoxes of self-reference [...] fit the [same

[7] This is not the 'usual' presentation of the so-called naïve theory of sets, which is more often identified as the theory consisting of the unrestricted comprehension scheme, i.e. what I here label (NS2). In fact, one also often finds the naïve theory of truth characterised as consisting of just the unrestricted T-scheme, i.e. (NT2); see e.g. Incurvati and Murzi (2017). In both cases, however, a minimal syntactic condition amounting to (NT1) and (NS1), respectively, as well as closure under classical implication, are implicitly assumed. The formulation given in the text merely makes this explicit, thereby also highlighting the similarities in the background conditions of the two families of paradoxes.

[8] See e.g. Tarski (1936). An immediate and useful upshot of this is that we can take the liar paradox (in any of its versions) as representative of all semantic antinomies without much loss of generality.

schema].” Thus, “Ramsey was wrong”: all such paradoxes are of the same kind (Priest 1994, 25).[9] On this basis, Priest champions a ‘principle of uniform solution’ for all paradoxes of self-reference alike—henceforth labelled *Priest’s Principle*, to be abbreviated by the more felicitous acronym (PP). In Priest’s own words (1994, 32):

same kind of paradox, same kind of solution.

Priest’s Principle offers a single, simple, unifying—hence, very attractive—prognosis for a variety of paradoxes. On the face of it, moreover, this unifying effort does not appear forced, nor *ad hoc*. When the paradoxes and their sources are reconstructed in parallel—we have seen one way of doing so in Sect. 2—the resemblance is indeed striking. But there is an even more important reason for being interested in (PP), to which we now turn.[10]

As we have just seen, (PP) moves from an *observed* fact about the paradoxes (they are structurally similar) to infer—or perhaps more accurately, to predict—a *new*, as yet unconfirmed claim about their solutions (they too are, or should be, structurally similar). My first contention is that this move is more problematic than Priest, and others following similar lines of argument, seem to think. To see why, we need to take a closer look at the elements of these arguments; although I will henceforth focus on Priest’s, as much for ease of presentation as because it is the more explicitly articulated of its kind.[11]

To this end, two key matters need to be addressed. First, what does Priest mean by two paradoxes—and their solutions—being structurally similar? And just what is it for something to be a ‘solution’ to a paradox? I address these questions in Sects. 3.1 and 3.2, respectively. Building on those analyses, in Sect. 3.3 I pinpoint what I consider to be the main weaknesses in the argumentative move encapsulated by (PP).

3.1 When are two paradoxes structurally similar?

What is it for two paradoxes to be ‘of the same kind’? Here Priest takes his cue directly from Russell, and specifically “an idea dating from some years before he lighted upon the Vicious Circle Principle” (Priest 1994, 27). The idea in question is the so-called *Russell Schema* (1994, 27):

[Given] a property ϕ, and function δ, consider the following conditions:

[9] For earlier examples of similar lines of argument see also Russell (1906), Herzberger (1970) and Feferman (1984). The discussion is also reproduced—with one small amendment—in Priest (1995, chapter 9).

[10] In the interest of maintaining focus, and without any loss of generality (cf. fn. 8), from here on I will take the liar paradox and Russell’s paradox as representatives of the two families of antinomies, respectively.

[11] Put differently: my aim here is not to ‘attack’ Priest specifically; none of what follows is intended to be in any way antagonistic. It merely strikes me that the apparently intuitive conclusion that (PP) invites us to draw is ultimately under-supported, and at the very least worth examining more closely than has been done so far.

1. $w = \{x : \phi(x)\}$ exists
2. if x is a subset of w: a) $\delta(x) \notin x$ and b) $\delta(x) \in w$[...] Given (1) and (2) we have a contradiction. For when [(2a)] and [(2b)] are applied to w, an irresistible force meets an immovable object. The result: $\delta(w) \in w$ and $\delta(w) \in w$.

The Russell schema is essentially a recipe for diagonalization; as Priest also points out, it is actually a special case of what he calls the *Inclosure Schema* (Priest 1995, 147):

> We now require two properties, ϕ and ψ, and a function δ, satisfying the following conditions:
>
> 1. $\Omega = \{y : \phi(y)\}$ exists, and $\psi(\Omega)$;
> 2. if x is a subset of Ω such that $\psi(x)$: a) $\delta(x) \notin x$ and b) $\delta(x) \in \Omega$Given that these conditions are satisfied we still have a contradiction.

The Inclosure and the Russell Schema each identify two conditions which, taken together, produce a contradiction. In this sense they are effectively templates for generating a paradox. Priest then shows that "all the paradoxes of self-reference [...] fit the [Inclosure] Schema. The structure that this describes is, therefore, the structure that generates all the paradoxes" (Priest 1994, 32). This then is what Priest means by two paradoxes being 'of the same kind', or 'structurally similar': that they can be reconstructed in such a way as to fall out of one and the same template. In particular, both the liar and Russell's paradox may be recovered in this way.

We now have a better grip on the first half of (PP). What about the second half? Oddly, Priest remains somewhat vague on the matter:

> If two paradoxes are of different kinds, it is reasonable to expect them to have different kinds of solution; on the other hand, if two paradoxes are of the same kind, then it is reasonable to expect them to have the same kind of solution. (Priest 1995, 183)

Let's assume that the interpretation of the phrase 'are of the same kind' (respectively: 'are of different kinds') remains invariant throughout, i.e. that it translates to 'drop out of the same template' (respectively: 'drop out of different templates') in each of its occurrences.[12] This being Priest, the template in question—the one it would be 'reasonable to expect' the solutions to drop out of—is dialetheism: the view according to which there are true contradictions.[13]

[12] Alternative interpretations and/or formulations are of course possible. Since Priest himself doesn't settle on any one formulation but resorts to different interpretations in different places—we'll see some examples later, here I opt for one that sits at the more neutral end of the spectrum. Other candidates, going from least to most loaded, might include: 'follow the same pattern of behaviour'; 'can be reconstructed in parallel'; 'enjoy a common/the same explanation'; 'stem from the same cause'; and so on.

[13] How does dialetheism help 'solve' the paradoxes? Take the liar paradox. In a classical setting, the liar is problematic because no sentence can be both true and false. Dialetheism offers a more relaxed view: there are true contradictions, and the conclusion of the liar argument (λ is true and λ is not true) is among them.

What is odd is that, having gone to great lengths to argue for the structural similarity of the paradoxes (and even greater lengths to argue for dialetheism), Priest seems to think it unnecessary to argue *that in fact it follows*, from the similarity of any two paradoxes, that their solutions too must be similar.[14]

Note that it is not enough to offer up dialetheism as evidence here. For dialetheism on its own does not suffice to establish (PP). At best, it tells us that there is *a* way of putting all paradoxes to rest—namely, by accepting their conclusions as true. But we are left none the wiser about what it means for two solutions to be of the same kind, nor *why* we should expect them to be so. In sum, it has yet to be shown *what supports the move* from the first to the second half of (PP). Put differently, it is unclear why we should accept the normative claim that all these paradoxes *should* enjoy the same kind of solution. We will come back to this in Sect. 3.3, for now is a good time to turn to our second question: what is it for something to be a solution to a paradox?

3.2 What is a solution to a paradox?

The answer, it seems, is not far to seek. Consider Russell's paradox. We know that it can be derived in NS; we also know *why* this is the case, in the sense that we can point to exactly which features of NS allow for the existence (or construction) of the set of all non-self-membered sets. But as long as we subscribe to NS—say, because we prize its simplicity and intuitive appeal, Russell's paradox will remain with us. By contrast, we know that there are other set theories, meeting different meta-theoretical desiderata, in which neither Russell's paradox nor any of its kin arise. I want to say, as a first approximation, that each of these paradox-free theories is a *prima facie* candidate for a solution to the set-theoretic paradoxes. The qualification (candidate solution vs. solution) is an important one: as will become clear, I don't think all paradox-free theories are on a par in this respect. Among the available candidates, I will focus on one in particular: Zermelo-Fraenkel set theory (plus the Axiom of Choice).[15] Thus, again as a first approximation, I want to say that *ZFC is a solution to the set-theoretic paradoxes*.[16]

One reason for restricting my attention to ZFC has to do with this paper's overarching goal. The hope is to contribute a fresh perspective on specific instances of philosophical methodology; not to 'solve' the paradoxes once and for all. Accordingly, I stick to one chief case study in order to prioritize focus over breadth; and I choose ZFC for ease of discussion, given its relative familiarity. But I think

[14] As we'll see, this is because Priest takes (PP) to be 'little more than a truism' (Priest 2002, 287); I disagree. We'll come back to this in Sect. 3.3.

[15] To reiterate: I am claiming that Russell's paradox is blocked in ZFC (whereas it can be derived in NS). I am *not* claiming that we have a direct proof of the consistency of ZFC. I will however follow the standard, widespread practice of talking *as if* this were the case, given the copious indirect evidence to this effect.

[16] Notice the indeterminate article! To repeat, far be it from me to claim that ZFC is the only paradox-free set theory. At the same time, for the reasons outlined in the main text I believe that ZFC still—for now—holds privileged status over its rivals.

analyses in a similar spirit can and should be undertaken with different set theories taking centre stage. A second and more important reason will become apparent shortly.

The 'relative familiarity' of ZFC is no mystery, of course: it is well known that since the axiomatic turn at the end of the 19th century, the cumulative hierarchy account has represented the set-theoretic orthodoxy. To this day, there is little question that ZFC is widely regarded as the 'default' set theory, even while it is flanked by a variety of competitors vying for its title as mathematical foundation. Happily, I need not concern myself with the whether, when or why of a potential passing of the torch: I do not need to take a stand on the 'best' extant set theory, whereas it is enough that there be an account that has withstood the test of progress in the way that ZFC has.

Putting this last part differently, it is enough for my purposes that ZFC passes the test of what Øystein Linnebo and Richard Pettigrew label *moderate naturalism*:

> Moderate naturalists about a particular scientific discipline hold that the opinions of scientists working in that discipline can suffice to establish that there exists a justification for some philosophically significant claim. [...] For instance, a moderate naturalist about mathematics might take the opinions of mathematicians to establish that there is a justification for the existential claims of traditional membership-based set theory. [...] And, in the iterative conception of set, she may take herself to have found it. By contrast, extreme naturalists claim that the very existence of the opinions of working scientists by itself provides the required justification for the claim. No further justification is needed beyond the fact that competent scientists with the relevant expertise assent to the claim in question. (Linnebo and Pettigrew 2011, 251)

Linnebo and Pettigrew draw this useful distinction between moderate and extreme naturalism in the context of a discussion of the prospects of category theory—in particular, the categorical theory of sets ETCS—as an autonomous foundation for mathematics. Ultimately, their conclusion is that the latter's "foundational credentials" will depend on where we stand on the naturalist spectrum. In particular, they argue that if we adopt a moderate stance, i.e. "if one requires that justifications be more substantive than those provided by extreme naturalism, then it seems doubtful that ETCS will have justificatory autonomy" (Linnebo and Pettigrew 2011, 252). Thus, by the lights of moderate naturalism ZFC enjoys a more robust justificatory basis than its rivals.

Linnebo and Pettigrew's tightly defended verdict allows me to sharpen my earlier claim: I propose that a 'solution' to the set-theoretic paradoxes is a set-theoretic account that is *successful* in the sense of being *sanctioned by moderate (set-theoretic) naturalism*.[17] Since ZFC is sanctioned—indeed, firmly supported—by

[17] I take it to be implied that such an account would be immune to paradox; that is, that being paradox-free is a necessary, though by now means sufficient, criterion of 'success' of a set-theoretic account (ditto for an account of truth, knowledge, social preferences, etc.). Indeed, I take it this is a given by the lights of both extreme and moderate naturalism.

moderate naturalism, it follows immediately that, by the lights of the proposed criterion, ZFC is a solution to the set-theoretic paradoxes. Notice, moreover, that the above-mentioned distinction between 'merely' candidate solutions and solutions can now also be made a little more precise: in particular, we can say that being paradox-free is not by itself enough to warrant the status of solution to the paradoxes; the viability of a candidate solution will depend on where it sits along the moderate naturalism spectrum.

As a different way of arriving at the same conclusion, consider what a solution to a paradox *shouldn't* (just) be: (i) a 'fix' to the problem, no matter what the (theoretical, methodological) cost; (ii) a (mere) diagnosis of the problem. It should be clear that (i) is just unsatisfactory and misguided. Arguably, moreover, (ii) is not enough on its own. We need to also have a way of reasoning with the concept in question without incurring into paradox. So we need to do better than (ii) and simultaneously safeguard against (i). And it seems to me that focusing on theories—where these are of course understood to be more than just a list of axioms—that have been subjected to, and have at least partially met, certain standards of philosophical and methodological robustness (among other things), is a plausible way of going about this.[18]

Generalizing, then, I propose that a solution to a paradox of X is an account of X that is successful (also) in the sense of being sanctioned by moderate naturalism (about X). Thus, when we look for a solution to a paradox—any paradox—we should not settle for a quick fix; rather, we should search for a well-motivated account of the concept in question. In particular, we should expect no less from a solution to the liar paradox or to Russell's paradox.

The methodological point I have just described is neither particularly deep nor particularly controversial; it deserves reminding of, more than it requires defending. Here are two (among many) passages bearing witness to the support it finds in the truth-theoretic literature:

> [...] we would not, in general, regard it as a satisfactory defense of an entrenched but inconsistent theory to weaken our logic so that we hide the contradiction. [...] To understand a concept properly, do not focus all your attention on how the concept behaves when it is on philosophical holiday. Look at the routine, unproblematic, non-philosophical work the concept does. (McGee 1989, 533–534)

> It is remarkable that despite the enormous effort that has been devoted to solving or dissolving the Liar paradox, the paradox and the mystery that surrounds it are still with us. It seems to us that a reason for this failure may be

[18] An anonymous reviewer suggested that diagnoses, rather than theories, might be the more appropriate terms of comparison in the sort of cases Priest and I are interested in. I am not convinced this would be enough, for the reasons given in the main text; but of course I could be wrong! By contrast, it seems to me that comparing what Incurvati (2020) calls *conceptions* of sets and of truth may well be an alternative way of framing the discussion. Indeed, given that my construal of theories (of sets/truth) treats these as more philosophically substantial than a list of axioms, I think that a conception-based comparison would not depart dramatically from the present proposal.

> that insufficient attention has been paid to the ordinary, unproblematic uses of the concept of truth. [...] Often we come to understand the extraordinary only when we see it in terms of the ordinary. [...] In order to gain a better understanding of the Liar, we need to give *less* attention to the paradoxes than we have given them. (Gupta and Belnap 1993, 17)

It should be noted at this point that ZFC has itself not been immune to philosophical criticism despite its undeniable and unparalleled success. A number of worries have been raised concerning the theory's rationale, among other things.[19] Underpinning these worries are three key contentions: that it was not until the 1930s that the axioms of ZF (without Choice) were brought together into a unified theory; that ZFC significantly departs from the original naïve notion of set; and that such a departure was motivated by the discovery of the set-theoretic paradoxes. The first claim is of course unimpeachable as a matter of historical fact. The second and third are also true; their implications, on the other hand, are much less cut and dry.

There is indeed little doubt that the paradoxes led mathematicians to look beyond the original 'naïve' account for a more satisfactory theory of sets. But there is a significant difference between recognizing this to be true and regarding the ZFC axioms as merely a clever way of patching things up so as to create a paradox-free haven. ZFC was not created overnight, and it is historically plausible that the paradoxes played a non-trivial role along the way. But from a naturalistic standpoint these considerations are easily outweighed by the fact that ZFC is still the most successful—in the sense described earlier—set theory around, and moreover one that preserves—albeit partially—the intuition that came before all others, namely unrestricted comprehension. Finally, note that similar worries could be raised against virtually any other 'solution' out there: it is hard to find a theory that has not been influenced, at *some* point along the way, by the paradoxes. Importantly, this is also true of dialetheism.

On the other hand, a look at the historical development of modern set theory suggests that the ZFC axioms were the end-product of what Potter (2004) calls a constructive process: that is, they originated from principles formulated as a means to explicate the pre-theoretic notion of set on the one hand, *and* reconcile the same with the desideratum of (classical) consistency, on the other.[20] The resulting theory has been currency ever since: not because it is better supported by philosophical argument, but because it is more compellingly sanctioned by mathematical practice, and so in turn by naturalistic standards. Incidentally, the same cannot be said about dialetheism, either in its truth-theoretic application or, certainly, in set-theoretic guise. In light of the foregoing, ZFC will henceforth be taken as our paradigm solution to the set-theoretic paradoxes.

[19] Priest himself gives voice to this particular worry in Priest (2006), for instance.

[20] This case is made convincingly by Moore (1982), so I will not be arguing for it here. See also Hallett (1986).

3.3 Priest's principle

We are now in a better position to assess (PP). The latter, recall, is the claim: "same kind of paradox, same kind of solution." Spelled out in full, it has the form of a conditional:

(PP) If two paradoxes are structurally similar, then (we should expect that) their solutions are structurally similar.[21]

We've seen that Priest (1994) argues for the antecedent of (PP) in the case of self-referential set-theoretic and truth-theoretic paradoxes. By straightforward reasoning we then have:

(PP) If two paradoxes are structurally similar, then (we should expect that) their solutions are structurally similar.

(PP_a) The set-theoretic and semantic paradoxes are structurally similar.

Therefore,

(PP_c) (We should expect that) the solutions to the set-theoretic and semantic paradoxes are structurally similar.

If the above argument could be vindicated, it would indeed be an extremely neat result, and certainly a game-changer in the philosophy of truth. Should we accept the argument and so, in particular, (PP_c)? It is not exactly supported by the evidence on existing solution attempts. The only other reason we have to endorse the conclusion is (PP). And so we come back to the question raised in Sect. 3.1: Should we, in fact, accept (PP)? Should we follow Priest's apparently seamless move from "same paradox" to "same solution"?

As far as I can tell, Priest himself offers two sorts of defenses of this principle. One is in the name of something like common sense, according to which (PP) is "little more than a truism." The other appeals to allegedly analogous examples from other domains of discourse. The two are sometimes interwoven, as witnessed by the first of the following passages:

> [Consider] an analogous principle: same kind of illness, same kind of cure—if two people have the same illness, they are to be cured in the same way. In a superficial sense, this is obviously false. For example, the same illness can be treated with two different drugs. But in a more profound, and more important, sense, the principle is clearly correct: if we have one illness in the two people, this must be due to the same cause. So the two people must be cured in the same way, namely, by attacking that cause. (Priest 2002, 287–288)

> [...] it is not necessary to be able to define something in order to be able to recognise it. Mathematicians usually recognise an explanation when they see

[21] Importantly, I am not hereby attributing to Priest the trivially false descriptive claim that all structurally similar paradoxes *do in fact* enjoy the same kind of solution. As already noted, Priest is making the more interesting claim that *we should expect* this to be the case. My reformulation of (PP) in slightly more articulate terms is intended to preserve this latter reading. In what follows I will sometimes drop the bracketed clause in the interest of readability.

> one; and I see one in the shape of the Inclosure Schema: an inclosure is what explains why the paradoxes arise. This is why such paradoxes are of a kind. (Priest 2002, 288)

At a glance, these somewhat anecdotal analogies seem quite effective at shoring up the intuitive plausibility of (PP). My worry is that as *arguments* for the same, they fall short of delivering the promised goods.

Suppose I don't share the intuition that (PP) is essentially a "truism" (as indeed I don't). Given that (PP) is also not (even close to being) vindicated by current truth or set theory, what's left to convince me otherwise? The medical analogy, it seems, except that too is of little help—for it is much too vague and subject to idiosyncratic interpretation to be effective, I think. Consider, for instance, that for a given illness there are typically various factors—hereditary, environmental, spurious, etc.—that qualify as its causes. Suppose A, B and C each have a heart attack: A's results from a vasospasm due to cocaine abuse, in absence of any other risk factors; B's is due to a spontaneous coronary artery dissection, also in absence of risk factors; C's is due to years of heavy smoking and poor eating habits, combined with genetic factors. One might object that these are only distal causes of A, B and C's respective conditions, whereas the principle of 'uniform cure' is supposed to apply to the respective proximal causes, and these are in fact one and the same. But this would-be objection seems weak. For one thing, in the rather banal scenario conjured up here, it is perfectly conceivable that the causal paths in question should only converge at the terminal stage, i.e. the illness itself: myocardial infarction. For another, information about the case-specific, distal causal factors can be crucial for the purposes of devising a 'cure', i.e. a treatment plan. Finally, even if we glossed over the differences in A, B and C's causal histories, that still wouldn't warrant an expectation of 'one illness, one cure.' For instance, A's heart might be fine after being shocked once; B's might require heart surgery and a pacemaker and subsequent pharmacological therapy; C might have to undergo a heart transplant.

It could be objected that I am misconstruing the paradox-as-illness metaphor. But this is in fact precisely my point: the metaphor is vague enough that one could easily construe it in more than one way and arrive in each case at a different conclusion. Equally importantly, just because a vague metaphor *appears* to be natural (to some), this doesn't automatically qualify it as robust—much less as a satisfactory substitute for a non-metaphorical argument.

My point here could be simply put as: we shouldn't be too quick in calling something a truism. More soberly, it seems that once we look past the enticing, intuitive shine of (PP), we see that it relies on a rather thin rationale.

Let's now piece together some of the foregoing. In Sect. 3.1, we saw that two paradoxes are structurally similar, on Priest's account, when they can be recovered from the same template. This is the bar set by the antecedent of (PP). I then turned my attention to the latter's consequent, in Sect. 3.2. I argued (or rather rehearsed existing arguments), that a solution to a paradox cannot just be a quick fix to an inconsistency, but should rather drop out of a well-motivated theory of the concept in question. This was recognised as an eminently reasonable methodological standard, which is also met—to general satisfaction—in the set-theoretic case.

I then took this point a short step further by recasting talk of a 'solution' to a paradox of X to that of a 'successful theory' of X: that is, a theory of X that is (paradox-free and) sanctioned by moderate naturalism about X. Thus, a 'solution' to Russell's paradox (note once again the indeterminate article) is supplied by ZFC; other 'solutions' may be supplied by other accounts of sets, as long as they pass muster by the lights of moderate naturalism.

We can now reformulate the earlier argument to reflect both these points:

(PP*) If two paradoxes are structurally similar, then (we should expect that) the theories of the respective concepts are structurally similar.[22]

(PP$_a$) The set-theoretic and semantic paradoxes are structurally similar.

Therefore,

(PP$^*_c$) (We should expect that) The theories of set-membership and of truth are structurally similar.

In its new formulation, the argument seems to lose much of its force. More specifically, and crucially for Priest's (and my) purposes, it casts (PP) in an unmistakably controversial light. For, taken on its own and at face value, (PP) is committing a methodological blunder: it is moving from an observed fact about 'irregular' (the term 'degenerate' is also sometimes used in the literature) applications of the concepts of set-membership and truth, to an otherwise groundless conclusion about the general, 'regular', unproblematic applications of the same, captured by and embodied in their respective underlying theories. And without any further supporting argument, such a move is unwarranted.

Unsurprisingly, an immediate upshot of this is that (PP$_c$) is now also cast in the same uncertain light. For, either we reject the idea that a solution to the truth-theoretic (respectively, set-theoretic) paradoxes *just is* a successful theory of the concept in question, and thereby enter methodologically dubious territory; or we accept this is so, and then it follows easily that the truth of (PP$_c$) depends on the truth of (PP$^*_c$).[23] And (PP$^*_c$), according at least to current wisdom, is unfounded. Note

[22] I take it to be implicitly understood that I am talking about theories that are at least moderately successful (in the above sense) here, since it is these that make for potentially interesting cases of analogy between 'solutions' to corresponding paradoxes. I readily acknowledge that it won't be possible to make a clearcut or definitive ruling on the success of every theory under the sun, since on the present characterization, a theory's success piggy-backs on its being sanctioned by moderate naturalism, and the latter's verdicts are often unstable in the long run. That said, I think that in most cases we have a pretty good—and shared—sense of where a particular theory may lie on the spectrum. I take this to be enough for present purposes, though admittedly not entirely satisfactory.

[23] As an anonymous referee noted, there is of course a third option: reject moderate naturalism across the board, and minimally as a standard of acceptability in both the set-theoretic and the truth-theoretic camps. The same referee suggested that this option would plausibly be favoured by Priest himself. Be that as it may, as things currently stand I am not aware of a fully satisfactory defense either of the claim that moderate naturalism is an inadequate methodological standard, or of the claim that a paraconsistent set theory of the sort advocated by Priest meets the standard of moderate naturalism. And so I think this is more of a worry for those who would favour this more radical option than for myself. See also Incurvati (2020, Chapter 4) for a particularly clear discussion of precisely this point. In addition to the foregoing, it seems to me that it would be (more) interesting if (PP) could be legitimated even with moderate

that this does not detract from its attractiveness: if either claim could somehow be rehabilitated—with or without (PP)—it could well be a genuine game-changer in the philosophy of truth. In the next section, I examine the prospects for upholding (PP^*_c) in a little more detail.

Before I do so, let me briefly sum up the main reasoning carried out so far. I've argued that the plausibility of (PP) depends on that of (PP^*), insofar as the latter, contrary to the former, offers a plausible account of what a solution to the paradoxes should look like. In turn, for something like (PP^*) to be upheld, a minimal desideratum is that we find at least a putative instance in which the above argument is sound, and so in which (PP^*_c) is true. And for the latter to be the case, we need to establish, at a minimum, the existence of at least one pair of truth- and set-theoretic accounts (a) that are sanctioned by moderate naturalism about truth and set-membership, respectively; *and* (b) that bear some degree of structural similarity with one another. One way to spell out the latter will be put forward in the next section.

4 Set theory and truth theory compared

How might we understand the claim that a theory of X and a theory of Y are structurally similar, for any two subject matters X, Y? There are no doubt many ways of doing so, although there hasn't been much discussion of this question to date; much less of how such putative structural similarities might support fruitful inter-theoretic analogies. The proposal I sketch here takes the key to such analogies to lie in the desiderata, or *norms*, underpinning our theories of X and Y.

The idea, in brief, is as follows. A norm of X may be thought of as an answer to the question: What should a theory of X be like? In the case of truth, examples of such answers might be: A theory of truth should preserve classical logic. Or: A theory of truth ought to be type-free. Or: A theory of truth ought to have a compositional semantics. And so on.

In general, answers to the question of what a theory of truth should be like will each be predicated on a single intuition or insight or thesis about truth itself; combined appropriately, they will form the basis for a theory of this concept. In other words, truth-theoretic norms are "the *prima facie* goals that we ought to reach for when we set up a theory of truth" (Leitgeb 2007, 276). Different answers—more precisely, different maximal consistent combinations of norms—will then point to, distinguish between, (sometimes uniquely) characterise, different philosophical views of truth. Indeed thinking about how to evaluate, rank, choose among theories from this perspective turns out to be particularly useful in the truth-theoretic arena; it affords us valuable insights on what has come to be a rather muddied debate, and thus an opportunity for making genuine progress in the area.

Footnote 23 continued
naturalism in the background. Put differently: if the only way for (PP) to go through is dependent on the rejection of moderate naturalism, then it would strike me as a much less interesting claim than I take it to be.

I do not propose to give a fully articulated defense of these claims here, for it would divert attention from our main focus; a more detailed account of this notion of norm has been offered elsewhere, in the context of a discussion of how we might adjudicate between competing truth theories.[24] But for another brief and straightforward illustration of the notion I have in mind, consider once again NT. In Sect. 2, this theory was seen to arise from the conjunction of the principles (NT1–NT3). I wish to say that the latter are in fact all and only the norms of NT. For, each is a partial answer to the question of what a truth theory should be like: it should have certain syntactic resources, it should prove the unrestricted T-scheme, it should preserve classical logic. Together, they 'make up', 'constitute' NT: they are the bare bones of one among many views of what truth should be like. Take away or tweak any of those principles, moreover, and the resulting theory will no longer correspond to the original, 'naïve' view of truth.

Importantly, and hopefully rather obviously, the conception of norm I have just sketched is not specifically tied to the truth-theoretic context, and can be transposed to other domains—for instance, set theory. There too we can ask after the "*prima facie* goals that we ought to reach for when we set up a theory of [set-membership]". And there, too, each set-theoretic account can be stripped down to its normative skeleton, i.e. the (possibly unique) combination of answers to that question. Crucially, focusing on a theory's underlying norms allows us to take a step back from the theory's specific (and potentially distracting) details; in particular, the task of juxtaposing and comparing any two theories can be safely simplified, at least initially, by taking their respective collections of norms to act as the relevant terms of comparison.

In sum, focusing on collections of norms as a proxy for full-blown theories can be a handy strategy in the context of different meta-theoretical tasks. In particular, we are now in a methodologically better position to approach the question that originally prompted this paper: namely, whether a genuine, robust analogy can be drawn between expected solutions to the set-theoretic and truth-theoretic paradoxes.

4.1 Norms elsewhere

First, though, let us look at one more illustration of the conception of norm at work, since it will be underpinning much of what follows. And this time, in order to get a sense of the potential versatility of this notion, we'll pick an entirely different domain of discourse: social choice theory.

Suppose we ask: What should a theory of social choice be like? From the meta-theoretical perspective adopted here, different answers to this question will correspond to different norms of social choice; and different combinations of these norms will in turn give rise to different accounts. For instance, so-called orthodox utilitarianism has it that individual and collective rational action (should) both tend towards the same goal: maximizing utility. More precisely, "social alternatives should be evaluated by the total utility they bring to all individuals in society [...].

[24] In Terzian (2012), which in turn builds on Leitgeb (2007); see also Terzian (2015).

So in effect, utilitarians claim that social choice should aim to maximise 'social utility', where this quantity is defined as the total [...] of individual utilities" (Okasha 2009, 572).

Thus, orthodox utilitarianism might give something like the following answer to our question: A theory of social choice should require that social preferences are ranked according to the utilitarian rule. From this, a putative norm of social choice emerges:

(U) Social preferences ought to be ranked according to the utilitarian rule.[25]

Why putative? Because it turns out that there is more than one mathematical route to the utilitarian rule; different routes are marked by different sets of theoretical assumptions which one may or may not be willing to accept. So in order to identify "the *prima facie* goals that we ought to reach for when we set up a theory of [social choice]" we need to backtrack to a stage that is conceptually and methodologically prior to (U). Ditto in the truth-theoretic case: one might at first be tempted to answer the original question by demanding that a truth theory ought to Tarskian, or that it behave according to Belnap-style revision semantics, or the like. But while these are certainly formulated as normative claims, they are also more like complex, 'post-theory' desiderata than basic, *prima facie* goals. Simply to demand that a truth theory be Tarskian, for instance, still tells us nothing about what this entails nor about the rationale for such a choice.

So what sort of assumptions do the trick in the social choice case? The literature identifies a number of these: for instance, the principles known as Pareto Indifference (P), Weak Pareto (WP), Anonymity (A), and so on.[26] Conjoined, these principles give rise to the utilitarian rule.[27] But each of these also corresponds to a *norm* of social choice: each expresses a desideratum on what a theory of social choice should be like.

This parallels the truth-theoretic picture we sketched earlier. We may naturally enough think of (P), (WP), (A) etc. as the bare bones of (a particular brand of) utilitarianism. Take away or tweak any of these principles, and the resulting theory will no longer correspond to the account of social choice we started from. Other, perfectly legitimate views may of course emerge as a result; the point is simply that the identity of that particular theory is inextricably tied to its norms.

[25] For more on the utilitarian rule see Okasha (2009), or any introduction to decision theory.

[26] In the interest of coherence of terminology and notation, I will be drawing directly and solely from Okasha (2009, 569–570) here. The Pareto axiom "says that if all individuals receive the same amount of utility in alternative x as they do in alternative y, then x and y are socially indifferent." Weak Pareto "says that [...] if all individuals are unanimous in preferring x to y, the social preference order should respect this fact." Anonymity "says that [the social preference order] is impartial, i.e. it pays no attention to the identity of individuals."

[27] More precisely, recall that utilitarianism can be brought about by more than one combination of such principles. Thus there is more than one way to complete the list so as to obtain a utilitarian choice function. Once again, the details do not matter to us here; see Okasha (2009) for a more comprehensive overview.

4.2 Norms and inter-theoretic analogy

The conception of norm presented above will be informing the rest of the discussion. In particular, it will prove valuable when it comes to assessing (PP*). I will now spell out in more detail why this is so, thereby also outlining the reasoning strategy I plan to follow in the rest of the section.

A theory's norms, we have seen, constitute a bona fide proxy for the former's distinctive features—its identity, as it were. By looking at its norms, we can tell what a theory is like: what it says about its subject matter, what its philosophical rationale amounts to, and so on. We can also *compare* two theories with respect to their norms. I've already alluded to one case in which this strategy was implemented—to produce a partial ranking of truth theories with respect to different parameters—that proved illuminating in several respects. We could describe an exercise of this sort as a comparison relative to normative *content*.

Alternatively, we could compare two theories relative to normative *structure*, in the following sense. Let X, Y be concepts and let N^X, N^Y be norms of X and Y, respectively. Thus N^X will be of the form 'A theory of X should $\Phi(X)$', and N^Y of the form 'A theory of Y should $\Psi(Y)$'. First, we define a criterion of structural analogy for norms (SAN):

SAN Two norms N^X, N^Y are structurally analogous iff the result of replacing X with Y within N^X (modulo minimal contextual tweaks) is N^Y, and viceversa.

Equivalently: N^X, N^Y are structurally analogous iff $\Phi(X), \Psi(Y)$ can be substituted for one another (modulo minimal contextual tweaks) without loss of content.

Then we can formulate a criterion of structural analogy of theories (SAT) along the following lines.

Strong SAT Given a theory T^X and a theory T^Y, and the respective collections of norms N_1^X–N_n^X, N_1^Y–N_n^Y: T^X is *strongly structurally analogous* to T^Y iff the collections N_1^X–N_n^X, N_1^Y–N_n^Y are pairwise structurally analogous.

Weak SAT Given a theory T^X and a theory T^Y, and the respective collections of norms N_1^X–N_n^X, N_1^Y–N_n^Y: T^X is *weakly structurally analogous* to T^Y iff $\exists m \in \mathbb{N}, 1 < m \leq n$ such that N_1^X–N_m^X, N_1^Y–N_m^Y are pairwise structurally analogous.

In other words: if two theories are structurally analogous to one another, *this should be mirrored by their respective norms*. Call this an inter-theoretic structural analogy litmus test:

SALT Two theories T^X, T^Y are structurally analogous if they satisfy at least Weak SAT.

SALT gives us one way (though by no means the only possible such) to more precisely articulate condition (b) from the conclusion of Sect. 3. Mirroring the formulation adopted there, I propose that for something like (PP*) to be upheld, it is

not enough to identify a common pattern underlying the set-theoretic and truth-theoretic paradoxes; we also need to establish, at a minimum, the existence of at least one pair of truth- and set-theoretic accounts (a) that are sanctioned by moderate naturalism about truth and set-membership, respectively; *and* (b) which jointly satisfy SALT.

In Sects. 4.2 and 4.3 I will run a partial litmus test on (PP*) with ZFC as the default representative for the set-theoretic camp. First, though, let us return once more to our social choice case study. A number of authors have noted that there are certain analogies between utilitarianism in decision theory and an otherwise unrelated domain of discourse, namely Darwinism in biology. Specifically, the key similarity between these areas seems to revolve around the notion of *maximisation of welfare*: respectively, maximization of utility by rational agents, and of fitness by organisms. A particularly clear and succinct articulation of this fact is given once again by Okasha:

> Just as rational choice theorists argue that much human behaviour can be explained by the hypothesis of utility-maximisation, so evolutionary theorists argue that much animal behaviour (and other aspects of phenotype) can be explained by the hypothesis of fitness-maximisation.
> The utility/fitness analogy does not mean that an agent who seeks to maximise their utility will thereby maximise their biological fitness; this is certainly not true, given the actual preferences of many people. Rather, *the idea is that utility and fitness play structurally similar roles in the respective theories in which they feature*. (Okasha 2009, 574; emphasis mine)

Okasha then suggests that the analogy between utilitarianism and Darwinism might actually extend beyond a mere surface appearance, by examining what happens if one replaces the notion of utility, within a decision-theoretic framework, with that of fitness. What is relevant to us is not so much the result of this analysis as the methodology being applied: Okasha shows, in effect, that *the observed surface analogy is backed up by an analogy at the level of the norms underpinning the two theories*.

In sum, this exemplifies—in a hopefully illuminating way—precisely the kind of methodological approach I have been advocating thus far. With this in mind, we can return to our main case study.

Now recall from Sect. 3 that if (PP) is to regain any plausibility, it had better be the case that we can find independent grounds for upholding (PP*), and in particular (PP$^*_c$)—the claim that we should expect our theories of truth and set-membership to be structurally similar. The good news is that we are now in a methodologically more favourable position to make some headway on this front: namely, run SALT on pairs of theories of truth and set-membership, to see whether we find a match.

Like many things, this is a simple enough task on paper; much less so in practice, given—among other things—the sheer number of both set and truth theories out there. Completion of this endeavour is thus deferred to a future paper; better yet, to a future collective effort on the part of interested readers. As previously announced,

the scope of the present investigation will be limited to a small subset of possible pairings, featuring ZFC and just a few truth-theoretic candidates.[28]

Since the asymmetry between the two sides weighs in favour of set theory—those paradoxes have been 'solved', after all—I start by identifying and briefly outlining the norms of ZF, in Sect. 4.3. In Sect. 4.4 I turn to the question of whether a counterpart to at least one of those norms might plausibly be thought to underlie a satisfactory theory of truth; in other words, whether a truth theory can be found that, together with ZF, satisfies SALT.

4.3 Norms of set-membership

What are the norms of ZF set theory? The literature points rather compellingly to the following principles, known as the *iterative conception* and the *limitation of size conception*:

(ZF1) A collection of objects is a set iff it is produced in a *linear*, *well-founded* process of the following sort: at stage 0 we have the empty set $\emptyset$; at stage 1, $\emptyset$ and its singleton set $\{\emptyset\}$; at stage 2, $\emptyset$ and $\{\emptyset\}$ and $\{\emptyset, \{\emptyset\}\}$; and so on, into the transfinite.

(ZF2) A collection of objects is a set iff it is *small enough*: that is, iff the objects of the collection are not in one-to-one correspondence with the universe of sets.

In addition—though it is almost invariably left implicit—we have, of course:

(ZF3) Classical logic holds.

Together, these principles paint a picture of the set-theoretic universe.[29] Each captures certain fundamental aspects of what it is for something to be a set; they jointly underwrite the account known as ZF set theory. Simply by surveying the principles—even in prose formulation, as above—we may readily recover many of the familiar characteristics of sets (according to ZF): they are (primarily) extensional objects, they are built 'from the ground up', they stand in a membership relation that is anti-symmetric, combinations of sets produce new sets, and so on. Moreover, as shown most famously by Boolos (1989), (ZF1)–(ZF2) are amenable to formalisation in such a way that the ZF axioms can be fully recovered from these two principles alone (always assuming that classical logic is in the background).[30]

Let's quickly take stock. We have seen three examples of how the proposed conception of norm can be a useful tool for bringing a theory's distinctive features into relief. To the question 'What should a theory of social choice be like?' a utilitarian replies: it should abide by (P), (WP), (A), etc. In the case of truth, a naïve

[28] Hereafter I will be somewhat sloppy and often omit mention of Choice.

[29] This is the 'picture' often described as the cumulative hierarchy of sets.

[30] See once again Incurvati (2020) for an admirably clear and philosophically engaged discussion of the iterative and limitation of size conceptions; in particular, of their differences with respect to the strength of their respective justificatory bases, and the diagnoses of the paradoxes they embody.

truth-theorist replies that it should satisfy (NT1)–(NT3). And an 'orthodox' ZF set-theorist answers: a theory of set-membership should abide by (ZF1)–(ZF3).

We have also seen that the plausibility of (PP) depends on that of (PP^*_c): the claim that the theories of set-membership and of truth are structurally similar. Finally, we identified a methodologically promising strategy for testing a claim of inter-theoretic similarity—namely, by verifying whether and to what extent this similarity extends to their respective underlying norms (SALT). We now put this strategy to work in order to reach a verdict on (PP^*_c).

There are two ways we could go about this: trawl the truth-theoretic landscape in search of independently motivated norms that fit the bill; or forge truth-theoretic norms that do, and thus produce a new theory of truth that is structurally analogous to ZF by construction. The former is the more conservative and therefore arguably preferable option; for this reason, as well as because it would take up far too much space, the latter approach is deferred to future work. Our next step is then to search for norms of truth that could stand as plausible counterparts to (ZF1)–(ZF3). If we succeed, we could seriously countenance the possibility that these norms might underpin an analogous 'solution' to the semantic paradoxes.

In fact, the task at hand can be simplified further, at least for the time being. First of all, since we're seeking for ways to satisfy Weak SAT, we can restrict our search to partial matches for (ZF1)–(ZF3). Secondly, thanks once again to Boolos, we know that (ZF1) and (ZF3) are sufficient to derive the subtheory Z^-, obtained from ZF by removing the axioms of replacement and extensionality.[31] Of course, Z^- is an impoverished version of our original target theory. On the other hand, it comprises all but two of its axioms, one of which—Extensionality—is widely thought to come as close to a self-evident principle as a non-logical axiom could hope to do. But for our purposes, this is enough: for, "[the] iterative conception is the only natural and [...] consistent conception of set we have, and it implies Z^-" (Boolos 1989, 90). And Z^-—together with the near-self-evident principle of extensionality—comprises the bulk of our target theory.

In sum: in order to reach at least a preliminary verdict on (PP^*_c), it is enough—for now—to run the litmus test on prospective truth-theoretic counterparts to (ZF1) + (ZF3).

4.4 Norms of truth

The iterative conception describes the process by which sets come into existence: start from the empty set (stage 0); apply the power-set operation to form a new set with the empty set as its (only) member (stage 1); iterate. At each stage, all elements present at that stage are collected to form the next element of the progression; the result is a hierarchy of sets and of levels. This process of generation is *cumulative*, in the sense that a set can only enter the hierarchy, but never exit. It is also *well-founded*: the chain of set-membership starts (or ends, depending on the point of

[31] For details, see Boolos (1989). A precursor of the latter is found in Boolos (1971).

view) with the empty set. All of the sets in the hierarchy are therefore also well-founded, in the sense that each set *depends on* the empty set.

One of the norms of ZF set theory, then, is (also) a norm of well-foundedness. This is particularly interesting for present purposes because a notion of well-foundedness also appears in the truth-theoretic literature. In particular, it takes centre stage in a number of prominent accounts spanning the past half-century. Here I focus on two representative accounts, due to Kripke (1975) and Leitgeb (2004) respectively.[32]

4.4.1 Grounded truth I: Kripke

In his seminal paper, Kripke gave a now famous portrayal of the process by which a semantic agent learns the meaning of the concept of truth. The idea is that our semantic agent—call them Sandy—constructs the extension of the truth predicate in a stage-wise manner, as follows. To begin with, Sandy has at their disposal only a purely non-semantic language, i.e. one containing expressions such as 'apple', 'no', '$0 = 0$', 'Tigers have stripes', and so on—but no truth predicate (stage 0).[33] They then learn that "we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself" (Kripke 1975, 701; stage 1). By holding firm this crucial first step, and by helping themselves to relatively elementary reasoning, Sandy "will eventually be able to attribute truth to more and more statements involving the notion of truth itself" (1975, 701). By the same token, moreover, they will also learn how to attribute falsity to both non-semantic and semantic statements of the language (stage n). Eventually, Sandy will be able to continue iterating this procedure indefinitely.

Formally, the process in question is captured by a monotone inductive operator—the *jump* operator—which produces a sequence of extensions (and anti-extensions) of the truth predicate. This sequence reaches several fixed-points, one of which is least (and unique).[34] Importantly, both the sequence and the resulting least fixed-point are well-founded, in the sense described earlier: in the terminology that has since gained currency, they are *grounded* in the lower-most stage of the construction (containing only non-semantic facts).

Kripke's inductive hierarchy and the set-theoretic hierarchy are visibly similar in a number of respects. First, both are grounded constructions: all sets in the ZF hierarchy are grounded in the empty set, and all sentences in the least fixed-point are grounded in non-semantic facts.[35] Secondly, both are cumulative: grounded

[32] It was Kripke's paper, of course, that first turned the spotlight on well-foundedness as a truth-theoretic constraint. Leitgeb's 2004 account is one amidst a numerous progeny of Kripke (1975); it also provides a healthy contrast to its forebear that is exactly relevant for the present discussion. Hence my choice to focus on these particular accounts for the purposes of a preliminary evaluation.

[33] As should be clear, the attribute 'non-semantic' is to be understood as meaning that the language initially available to Sandy does not include concepts such as truth, falsity, denotation, etc.

[34] I am not interested in rehearsing the technical details of the construction here. Countless sources providing such details are available, starting of course with Kripke (1975).

[35] Again, see e.g. Leitgeb (2004, 168).

sentences can never be cast out of the least fixed-point (nor at any earlier stage, once they enter the extension or anti-extension of *tr*). Finally, from any level σ, σ-many iterations of the inverse power-set and jump operations (respectively) eventually lead back to stage 0 of the construction. Thus, for any set (respectively: sentence) that makes an appearance in the hierarchy, we can trace its dependence path back to its respective ground.

For all the similarities between the Kripkean and the set-theoretic hierarchy, there is nonetheless one prominent element of disanalogy. Recall that for an object to qualify as a set, and so to appear in the set-theoretic hierarchy, *and so* to stand in a dependence relation to another set, it must be grounded in the empty set; and vice versa. In the Kripkean construction, a sentence can be assigned a truth value if and only if it is grounded, but ungrounded sentences aren't thereby banished from existence, unlike their set-theoretic counterparts; rather, we simply won't find them in the least fixed point. The discrepancy is explained by the fact that Kripke-dependence is strictly compositional: the liar sentence (call it λ) is ungrounded, but it depends on the sentence $\neg tr(\lambda)$ because the two are materially equivalent, by definition of λ; and of course $\neg tr(\lambda)$ depends on λ, and so on.

The crucial point is that unlike the set-theoretic dependence relation, the compositional analysis underwriting Kripke-dependence fails to discriminate circular from well-founded dependence paths. In sum, Kripkean groundedness and ZF well-foundedness do not satisfy the criterion of structural analogy: our first litmus test returns a negative match.

4.4.2 Grounded truth II: Leitgeb

In brief, Leitgeb's aim is to identify a class Φ, with $\mathcal{L}_{PA} \subset \Phi \subset \mathcal{L}_{tr}$, consisting of *all and only* those sentences $\phi \in \mathcal{L}_{tr}$ that depend directly or indirectly on non-semantic states of affairs. Formally, this is accomplished by defining a dependence operator $D(\Phi)$ which collects into a set all the sentences of $\mathcal{L}_{tr}$ that stand in an appropriate dependence relation to one or more other sentences. The condition for a sentence ϕ to 'appropriately depend' on another, or more generally on a set of sentences $\Phi \subseteq \mathcal{L}_{tr}$, is the following:

$$\phi \text{ depends on } \Phi \quad \text{iff} \quad \forall \Psi \subseteq \mathcal{L}_{tr} : \ Val_{\Psi}(\phi) = Val_{\Psi \cap \Phi}(\phi).$$

A distinctive feature of Leitgeb's notion of dependence (henceforth: L-dependence) is that "[this] notion [...] is a kind of *supervenience*" (Leitgeb 2004, 160), in the sense that the truth value of ϕ cannot change without a change in the extension of the truth predicate. This means that for two sentences $\phi, \psi \in \mathcal{L}_{tr}$ to stand in a relation of L-dependence, it must be the case that the semantic facts about ϕ, ψ, *and* the extension of *tr*, are already settled.

Leitgeb's dependence operator is applied iteratively to sets of sentences of the language; because of the 'supervenience' character of L-dependence, at each stage there will be a collection of sentences that are fed into the evaluating mechanism together, rather than one at a time. (This is in contrast with the compositional aspect of Kripkean dependence, for instance.) The L-dependence operator is also monotone inductive, and therefore has a least fixed-point Φ_{lf}. The latter is

particularly interesting for two (related) reasons: the first is that it contains the unrestricted instances of all of its T-biconditionals; the second is that it contains all and only the grounded sentences of the language.

Leitgeb also observes that L-dependence is not specifically tied to the truth-theoretic context: "Since *tr* is left uninterpreted and is a fortiori not yet regarded as expressing truth, our theory of dependence is not specifically a part of a theory of truth but its scope is more general; other applications of the theory are conceivable" (Leitgeb 2004, 174). Not altogether surprisingly, one such application is none other than set theory. Indeed, the monotonicity of L-dependence ensures that Leitgeb's construction shares remarkably many structural features with the set-theoretic hierarchy. As Leitgeb himself puts it,

> [...] if the domain of this dependence relation is restricted to the set of sentences that depend on non-semantic states of affairs, a well-founded hierarchy of dependency up to the ordinal level of [the least fixed-point] is determined. This is [...] analogous to the situation in set theory, where [...] if the domain of the membership relation is restricted to the class of well-founded sets, membership turns out to be arranged according to a similar ordinal system of levels (though, of course, of "unbounded height"). (Leitgeb 2004, 174–175)

To sum up, we have seen two ways in which sentences can stand in a dependence relation: compositional, Kripkean dependence on the one hand; supervaluational L-dependence on the other. And we have seen that the former offers insufficient normative grip to settle questions about truth-evaluability; whereas L-dependence gives a way of settling such questions. Also importantly, the choice of a supervaluational scheme ensures that the logic of Leitgeb's account is thoroughly classical. As far as the prospective analogy with ZF set theory goes, the latter is a point in Leitgeb's favour; although the advantage is diminished precisely by the need to resort to a supervaluational logic. Is this enough for Leitgeb's account to satisfy at least Weak SAT? It doesn't fail completely, for sure, insofar as it guarantees a full match to (ZF3). It also doesn't perform significantly better than this, insofar as it achieves only a partial match to (ZF1). If we adopted a strongly charitable stance, we could chalk this up as a moderate success for Leitgeb's account, by the lights of SALT.

A similar conclusion seems warranted for Kripke's account, as long as similar standards of charity are in place. In this case, compositionality is on Kripke's side; not so his preferred non-classical valuation scheme. Once again, we have at best a partial match with our set-theoretic counterpart.

Where does this leave us? I've only examined two truth theories among many, so no definitive conclusions are possible at this point. But for the sake of argument, let's suppose for a moment that these two are the closest conceivable matches to ZF in the truth-theoretic camp. Would this scenario speak in favour of the plausibility of (PP*), or against it? As we've just seen, neither account comes close to satisfying Strong SAT; we could at best say that they (weakly) satisfy Weak SAT. Still, in this putative scenario there are no better options on the table, so perhaps a case could be made for relaxing our standards for inter-theoretic analogy. While I would find this

an unsatisfactory compromise, it is conceivable that one might be willing to bite the bullet for the sake of seeing (PP*) go through. Even then, *it would still need to be argued* that either account is also successful as a theory of truth; in particular, that it is sanctioned by moderate truth-theoretic naturalism. And as things stand, neither Kripke's nor Leitgeb's account seems quite up to the challenge.

Finally, notice once again that dialetheism does no better than either Kripke's or Leitgeb's theories of truth. More accurately: it does no better... according to current wisdom. As I've been arguing, the burden of proof on potential candidates for inter-theoretic analogy is significant—as we should expect it to be. And in the case of dialetheism, as with any of its rivals, meeting these standards requires more than being able to put the paradoxes to rest. Crucially, it would require both of its truth- and set-theoretic counterparts to be sanctioned by moderate naturalism; for the time being at least, dialetheist 'solutions' fall short of the mark.

It could be legitimately objected that I have not, after all, discovered anything we didn't know already. I agree, although with some qualifications. First, I hope to have made it clear that it was never my aim (much less expectation) that I would somehow magically discover a solution to the paradoxes, or stumble across a pair of perfectly analogous theories of truth and set-membership. Nor, just as importantly, did I set out to 'prove' Priest wrong. It may even be that the latter's preferred dialetheic 'solution' to the paradoxes will eventually be sanctioned by moderate truth-theoretic *and* set-theoretic naturalism, even though that isn't currently the case. My concern throughout was rather to clarify what we are (or should be) talking about when we talk of 'solving' a paradox; and correspondingly to invite caution, and greater transparency, when claims such as (PP) are put forward. In a way, my only real disagreement with Priest concerns his claim that (PP) is a truism. As I hope to have shown, it is far from that.

5 Conclusion

This paper began by calling into question the claim, defended by Priest among others, that the presence of structural similarities between paradoxes of truth and set-membership supports the conclusion that the paradoxes should have a common solution. I argued, first, that in order for it to be adequately assessed, this claim ought at the very least to be accompanied by an account of *what counts as a solution* to a paradox. I then examined one possible contender. I proposed that, in the first instance, a reasonable proxy for a 'solution' to a paradox of X is a *theory* of X; specifically, one that is well-motivated and satisfactory by moderately naturalistic standards. I then refined this idea by arguing that the question, of whether the two families of paradoxes might after all share a solution, ought to be reformulated as the question of whether our theories of set-membership and truth might share (at least some of) the same *norms*, where these are subjected to the above-mentioned naturalistic standards.

Section 4 took concrete steps towards addressing this question. In the course of that discussion, I also took (PP) as a springboard to raise the broader question: What warrants claims of inter-theoretic structural analogy such as (PP^*_c)? I went on to

propose a pair of criteria for testing the plausibility of the latter, and applied them to a specific subclass of possible set and truth theory pairings: namely, ZF *vis-à-vis* theories of grounded truth. The ensuing discussion established that for this particular class, the analogy fails. Thus far, then, (PP^*_c)—and therefore a fortiori (PP_c) and (PP)—remain unwarranted.

Does the negative result reached here sanction the definitive failure for this and any other analogy between truth and set theories, and thus between the solutions to the respective paradoxes? Might Ramsey have been on to something after all? Not necessarily, I think. For one thing, there are plenty more possible pairings of set-theoretic and truth-theoretic norms that have not yet been examined. For another, there is at least one other possible approach, briefly mentioned earlier on: namely, building a theory of truth that *is* analogous to ZFC, ETCS, or whichever theory one takes to be the strongest candidate in the field. Arguably, the resulting truth theory would be artificial; but it is an open question whether it might nevertheless turn out to be an adequate account of the notion of truth.[36] Finally, it remains an open possibility that dialetheism might eventually come to be sanctioned by moderate naturalism from both the set-theoretic and the truth-theoretic camps, thus vindicating the uniform 'solution' originally championed by Priest.

In conclusion, the key upshots of this investigation amount to the following: first, having clarified the terms of the debate in which Priest and others have engaged and which has so far been vitiated by an important ambiguity; secondly, having set up a cleaner, more transparent framework within which this and similar debates can now play out. It is worth emphasising once more that although the few concrete results obtained thus far are not conclusively positive, they are not conclusively negative either. If nothing else, continuing along the path we have traced promises to deliver interesting, perhaps even novel insights into the normative fabric of our theories of set-membership and truth.

Acknowledgements This paper has been a (very) long time in the making. My greatest debt of gratitude is to Hannes Leitgeb: many years ago, he patiently listened to what were, at the time, my ill-formed and scattered thoughts on this issue, helped me unpack them and encouraged me to write them up. I am also grateful to audiences at the University of Bristol and at several conferences where I presented this material. Within these audiences, special thanks are due to Roy T Cook and the late Jaakko Hintikka for valuable feedback and much-appreciated encouragement. Finally, I wish to thank two anonymous reviewers from this journal for their helpful comments, as well as several other reviewers who read previous versions of this paper and offered feedback over the years.

[36] Or, at least, no less adequate than rival existing theories.

References

Boolos, G. (1971). The iterative conception of set. *The Journal of Philosophy*, *68*(8), 215–231.
Boolos, G. (1989). Iteration again. *Philosophical Topics*, *17*(2), 5–21.
Feferman, S. (1984). Toward useful type-free theories. I. *The Journal of Symbolic Logic*, *49*(1), 75–111.
Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge: MIT Press.
Hallett, M. (1986). *Cantorian set theory and limitation of size* (Vol. 10). Oxford: Oxford University Press.
Herzberger, H. G. (1970). Paradoxes of grounding in semantics. *The Journal of Philosophy*, *67*(6), 145–167.
Incurvati, L. (2020). *Conceptions of set and the foundations of mathematics*. Cambridge: Cambridge University Press.
Incurvati, L., & Murzi, J. (2017). Maximally consistent sets of instances of naive comprehension. *Mind*, *126*(502), 371–384.
Kripke, S. A. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*(19), 690–716.
Leitgeb, H. (2004). What truth depends on. *Journal of Philosophical Logic*, *34*(2), 155–192.
Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, *2*(2), 276–290.
Linnebo, Ø., & Pettigrew, R. (2011). Category theory as an autonomous foundation. *Philosophia Mathematica*, *19*(3), 227–254.
McGee, V. (1989). Applying Kripke's theory of truth. *Journal of Philosophy*, *86*(10), 530–539.
Moore, G. (1982). *Zermelo's axiom of choice: its origins, development and influence*. New York: Springer.
Okasha, S. (2009). Individuals, groups, fitness and utility: Multi-level selection meets social choice theory. *Biology and Philosophy*, *24*(5), 561–584.
Potter, M. (2004). *Set theory and its philosophy: a critical introduction*. Oxford: Oxford University Press.
Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind*, *103*(409), 25–34.
Priest, G. (1995). *Beyond the limits of thought*. Cambridge: Cambridge University Press.
Priest, G. (2000). On the principle of uniform solution: A reply to Smith. *Mind*, *109*(433), 123–126.
Priest, G. (2002). *Beyond the limits of thought*. Oxford: Oxford University Press.
Priest, G. (2006). *In contradiction: A study of the transconsistent*. Oxford: Oxford University Press.
Ramsey, F. P. (1925). *The Foundations of Mathematics*. Abingdon: Routledge & Kegan Paul.
Russell, B. (1906). On some difficulties in the theory of transfinite numbers and order types. *Proceedings of the London Mathematical Society*, *4*(14), 29–53.
Smith, N. J. (2000). The principle of uniform solution (of the paradoxes of self-reference). *Mind*, *109*(433), 117–122.
Tarski, A. (1936). The concept of truth in formalized languages. In A. Tarski (Ed.), *Logic, semantics, metamathematics* (pp. 152–278). Oxford: Oxford University Press.
Tarski, A. (1944). The semantic conception of truth: And the foundations of semantics. *Philosophy and Phenomenological Research*, *4*(3), 341–376.
Terzian, G. (2012). *Uncovering the norms of truth: A meta-theoretic inquiry*. PhD thesis, University of Bristol.
Terzian, G. (2015). Norms of truth and logical revision. *Topoi*, *34*(1), 15–23.