On Forms of Justification in Set Theory

Neil Barton[*], Claudio Ternullo[†], and Giorgio Venturi[‡]

18 March 2019

Abstract

In the contemporary philosophy of set theory, discussion of new axioms that purport to resolve independence necessitates an explanation of how they come to be *justified*. Ordinarily, justification is divided into two broad kinds: *intrinsic* justification relates to how 'intuitively plausible' an axiom is, whereas *extrinsic* justification supports an axiom by identifying certain 'desirable' consequences. This paper puts pressure on how this distinction is formulated and construed. In particular, we argue that the distinction as often presented is neither *well-demarcated* nor sufficiently *precise*. Instead, we suggest that the process of justification in set theory should not be thought of as neatly divisible in this way, but should rather be understood as a conceptually indivisible notion linked to the goal of *explanation*.

Introduction

In what sense are mathematical claims *justified* and what, if any, processes constitute relevant and legitimate *justificatory* processes?

These are crucial issues for the philosophy of mathematics (and for philosophy, in general), as the answers to the questions above clearly bear on the acceptance or rejection of fundamental pieces of mathematical knowledge. Notable historical examples relating to the problem of justification in mathematics include the axioms of geometry, such as Euclid's Fifth Postulate and, more recently, set-theoretic propositions such as the Axiom of Choice and the Continuum Hypothesis. As is clear, a decision in favour or against the acceptance of each of these axioms or statements has strong consequences for the practice of several mathematical disciplines.

Justification has become even more pressing within contemporary set theory as a consequence of the independence phenomenon. As is known, there are important set-theoretic statements that cannot be decided by **ZFC** (our current most widely accepted theory of sets). A central question in the philosophy of set theory has thus concerned how we might settle these statements (and, indeed, if we should).

The project to settle independence through selecting and adopting new axioms was famously championed by Gödel in an influential paper on the Continuum Hypothesis and has since been known as Gödel's Programme.[1] Since its formulation,

[*]KGRC. E-mail: neil.barton@univie.ac.at. I would additionally like to thank the FWF (Austrian Science Fund) for their support through project P 28420.

[†]KGRC. E-mail: claudio.ternullo@univie.ac.at.

[‡]University of Campinas. Email: gio.venturi@gmail.com.

[1]See [Gödel, 1947] and the re-write in [Gödel, 1964].

it has become increasingly clear that the fulfilment of Gödel's Programme also requires examining (and making full sense of) the notion of *mathematical justification*. In this paper, we shall concern ourselves with examining the two most widely discussed forms of justification, both of which appear in Gödel's writings, and which have gradually come to be viewed as 'standard' in set theory, namely *intrinsic* and *extrinsic* justification. In particular, we shall focus on the way the distinction has been characterised and its relevance construed in recent work by Penelope Maddy, in particular her *Defending the Axioms* ([Maddy, 2011]).

Our goal is twofold:

(1.) To show that the distinction between intrinsic and extrinsic justification, and the notion that there might be a preferable kind, is fraught with problems.

(2.) To propose arguments in favour of a conception of justification as *multi-faceted* but fundamentally *indivisible* and linked to the notion of *explanation*. 'Intrinsic' and 'extrinsic' justifications should (on our view) be understood as manifestations of explanatory considerations.

The structure of the paper is as follows. First (§1), we provide an account of intrinsic and extrinsic justification as it appears in Maddy's aforementioned work. Next (§2), through analysing various case studies, we develop two problems for the account, one concerning the *tractability* of the distinction, and the other concerning the *demarcation* between the two kinds of justification. With these problems in view, we then (§3) argue that an understanding of justification as a process of determining which principles are explanatory yields a more satisfactory account of justification in set theory. Moreover, we shall also argue that our account is able to successfully respond to the issues of tractability and demarcation, by partly dissolving and partly re-considering the relevance of both issues. We then (§4) consider some objections to the account proposed in §4. Finally (§5) we conclude with some philosophical upshots and directions for future research.

1 Intrinsic and Extrinsic Justification

The distinction between intrinsic and extrinsic justification goes back plausibly as far as [Russell, 1907], but is most famously introduced and discussed in [Gödel, 1947] (with subsequent revisions in [Gödel, 1964]). Gödel's ideas have been widely debated in the literature. In particular, they have been extensively scrutinised by Penelope Maddy in a series of influential papers and monographs,[2] and more recently additional exploration has been provided by Peter Koellner[3]. For the purposes of this article, we will be primarily focussed on the origins of the distinction in Gödel, and the subsequent developments in Maddy, although we will relate back to the broader literature where possible.

We start with a couple of quotations from Gödel's 1947 paper. Concerning intrinsic justification, Gödel writes:

> "For first of all the axioms of set theory by no means form a system closed in itself, but, quite on the contrary, the very *concept of set* on which they

[2]See [Maddy, 1988a], [Maddy, 1988b], [Maddy, 1990], [Maddy, 1997], [Maddy, 2007], and [Maddy, 2011].

[3]See [Koellner, 2006] and [Koellner, 2009].

> are based suggests their extension by new axioms which assert the existence of still further iterations of the operation "set of". [...] Probably there exist others based on hitherto unknown principles; also there may exist, besides the ordinary axioms, the axioms of infinity and the axioms mentioned in footnote 17 [here Gödel means large cardinal axioms] other (hitherto unknown) axioms of set theory which a more profound understanding of the *concepts underlying logic and mathematics* would enable us to recognize as *implied* by these concepts." [italics are all ours] ([Gödel, 1947], p. 181)

Immediately thereafter, Gödel also explains that there might be further criteria for the acceptance of an axiom:

> "Furthermore, however, even disregarding the intrinsic necessity of some new axiom, and even in case it had no intrinsic necessity at all, a decision about its truth is possible also in another way, namely, *inductively by studying its "success"*, that is, its *fruitfulness* in consequences and in particular in "verifiable" consequences, i.e., consequences demonstrable without the new axiom, whose proofs by means of the new axiom, however, are considerably simpler and easier to discover, and make it possible to condense into one proof many different proofs." [italics are all ours] (*ibid.*, p. 182)

The quotations above provide us with the bones of an account of what one should take the two forms of justification to consist in: *Intrinsically justified* new axioms are those that follow from the concept of set, or, more generally, 'concepts' underlying logic and mathematics, whereas *extrinsically justified* new axioms are those which are justified through studying their success, fruitfulness and consequences.

In more recent times Gödel's distinction between intrinsic and extrinsic justifications has been taken up and further developed. For example, Maddy writes:

> "When a principle is defended in terms customarily classified as intrinsic, various descriptors typically appear: the principle is intuitive, self-evident, obvious; it's part of the meaning of the word 'set'; it's implicit in the very concept of set; and so on. Of course, each of these glosses raises its own suite of questions. These days, I think that the most common idea is the last-mentioned—implicit in the concept of set—and that the concept of set intended is the iterative conception." ([Maddy, 2011], p. 124)

Intrinsic justification thus includes a cluster of ways in which we might justify a particular principle. However, Maddy's focus is what we take to be 'implicit' in the relevant concept. Extrinsic justification, on the other hand, is not concerned with whether or not a principle results from a successful conceptual 'unfolding', but rather concerns the *consequences* that it has. For our purposes, this is the key aspect of the distinction. For example, again, Maddy re-phrases Gödel's description of the two kinds of justification as follows:

> "It has become customary to describe these two rough categories of justification as 'intrinsic'—self-evident, intuitive, part of the 'concept of set', and such like—and 'extrinsic'—effective, fruitful, productive." ([Maddy, 2011], p. 47)

Thus, extrinsic justification consists in justifying a principle through identifying its consequences. In particular, if a particular proposed principle (axiom) has 'effective', 'fruitful', or 'productive' consequences, then we can count it as receiving extrinsic justification.

Now, for our purposes, it is important to emphasise the following fact (that we shall question later): extrinsic and intrinsic justifications are often seen as *orthogonal* and *competing*. For instance, Maddy is circumspect about the usefulness of intrinsic justifications for anything stronger than the most basic set-theoretic axioms, and believes that extrinsic justifications are to be preferred:

> "Ultimately we aim for consistent theories, for effective ways of organizing and extending our mathematical thinking, for useful heuristics for generating productive new hypotheses, and so on; intrinsic considerations are valuable, but only insofar as they correlate with these extrinsic payoffs. This suggests that the importance of intrinsic considerations is merely instrumental, that the fundamental justificatory force is all extrinsic. This casts serious doubt on the common opinion that intrinsic justifications are the grand aristocracy and extrinsic justifications the poor cousins. The truth may well be the reverse!" ([Maddy, 2011], p. 136)

For Maddy then, the important facts are the *consequences* a principle has.[4] Her view is supported by the fact that mathematical practice is usually dictated by the relative fruits of a body of mathematics at any particular time. Many mathematicians care mainly about proving theorems, and will (for the most part) use whatever available tools seem most appropriate to them, no matter whether or not they have intuitive arguments.

A view opposite to Maddy's is that (set-theoretic) mathematics should be derived from intuitively supported principles. For instance, Mary Tiles counts as a supporter of such a view in the following quote:[5]

> "To claim this [i.e. foundational] status for set theory it is necessary to claim an independent and intrinsic justification for the assertion of set-theoretic axioms. It would be circular indeed to justify the logical foundations by appeal to their logical consequences, i.e., by appeal to the propositions for which they are going to provide the foundations." ([Tiles, 1989], p. 208)

Tiles' point is that set theory is precisely meant to be *providing* the foundations for mathematical reasoning. To have foundations for our mathematical reasoning, we need then to be based on intuitively evident principles. To appeal to the consequences of a principle is, for Tiles, to presuppose the very thing for which we are trying to provide foundations.

[4] At least the Maddy of [Maddy, 2011]. Her views change quite substantially from the account provided in [Maddy, 1990]. It is with this more recent version of Maddy's conception that we are concerned in this paper.

[5] William Tait is another example here, as in the following quotation:

> "To introduce a new axiom as "true" on this [i.e. iterative] conception because of its "success" would have no more justification than introducing in the study of Euclidean space points and lines at infinity because of their success. ... A "probable decision" about the truth of a proposition from the point of view of the iterative conception can only be a probable decision about its derivability from that conception. Otherwise, how can we know that a probable decision on the basis of success might not lead us to negate what we otherwise take to be an intrinsically necessary truth?" ([Tait, 2001], reprinted in [Tait, 2005], p. 284)

In sum, we have a distinction between intrinsic and extrinsic justification by which the former takes axioms to be following from and being justified in view of the concept of set, whereas the latter views justification as resulting from having desirable consequences. Many authors have seen this taxonomy as providing *distinct* and possibly *competing* kinds of justification.

In the next section, we will raise two problems for the distinction (tractability and demarcation), before providing what we take to be the beginnings of a solution.

2 Difficulties with the distinction: Epistemic Usefulness, Tractability, and Demarcation

In this section, we want to address two main problems with the distinction between intrinsic and extrinsic justification. Although these can be described separately, one could, in fact, view them as originating from a *single* incorrect attitude, namely that of taking 'intrinsic' and 'extrinsic' to be fundamental and distinct kinds of justification.

2.1 Epistemic usefulness

First, however, we need to set up the driving force behind our objections. The key problem concerns what we should expect from a theory of mathematical justification.

On the one hand, we expect such a theory to fulfil certain descriptive tasks, for example we would ideally like it to provide us with a general conception of what it means to have justified knowledge of a statement. On the other hand, in a scientific context in which we look to select from many distinct available hypotheses (some of which may be in conflict with one another) we expect a little more. In particular, we expect our account of justification to have normative force, and to be of use in deciding between various possible axioms. We therefore put the following desideratum on accounts of justification:

> **Epistemic Usefulness.** We would like our account to be *epistemically useful* in that it should be usable in either justifying new principles, or explaining why justification is not possible. A theory of justification that can be put to use in actually analysing the justification of scientific claims is preferable to one that cannot be so used.

This simple requirement concerning justification will form the basis of our criticisms of the intrinsic/extrinsic distinction.

2.2 Tractability

With the idea of Epistemic Usefulness in play, we first consider a problem of *tractability*. Simply put, the accounts of intrinsic and extrinsic justification are not sufficiently tractable so as to be epistemically useful. In particular, it is very hard to tell when a principle is or is not intrinsically/extrinsically justified. We examine each kind of justification with respect to tractability in turn.

2.2.1 The intractability of 'intrinsic justification'

We begin with intrinsic justification. There appears to be no systematic way to ascertain whether an axiom *conforms* to the concept of set, as would seem to be implied by

the notion of 'intrinsic'. One construal of 'conformity' evokes the idea that new axioms could be literally *derived* from the concept of set *analytically*.[6]But the concept of analyticity is clearly a difficult one to work with. This can be brought into sharper focus by considering justification under the iterative conception[7]. There is a good deal of agreement on the fact that the **ZFC** axioms are true of such concept (or at least heuristically justifiable on the basis of it). Certainly this has been suggested by the classical [Boolos, 1971][8], [Parsons, 1983], and [Wang, 1974]. However, in spite of the general agreement that the **ZFC** axioms are all true under the iterative conception, some authors have shown scepticism. Potter, for one, has presented arguments that Replacement enjoys a special status, which requires alternative justificatory strategies.[9] Elsewhere, Feferman has expressed worries about the power-set operation underlying the iterative conception.[10]

As it stands, when considering what 'follows from' a concept, we appear to have a stalemate: it is not clear what criteria one could appeal to in trying to convince a different party who bluntly disagrees with what follows from their concept of set. Without some methodology for making progress on these debates, it is hard to see how intrinsic justification can gain traction and be epistemically useful.

There is a more general difficulty concerning what our concept is like at a given point in time. It is common in the literature (or, at least, this is the impression one might get from some set theory textbooks) to introduce the axioms in close connection to, sometimes even motivated by, the iterative concept of set.[11] But this should not be taken to be inevitable. As we know, the iterative concept fully emerged *only after* the set-theoretic axioms were formulated by Zermelo and Fraenkel (with the significant contribution of Skolem).[12] Of course, many authors have retrospectively tried to reconstruct set theory in light of the emergence of the iterative conception, but this is both historically and conceptually inaccurate. As Potter remarks:

> "In contrast with the limitation of size conception, [the iterative conception] took a long time to emerge [...] However, in an attempt to make the history of the subject read more like an inevitable convergence on the one true religion, some authors have tried to find evidence of the iterative conception quite far back in the history of the subject." ([Potter, 2004], p. 36)

It is perhaps useful to briefly contrast the various routes that our conception of sets might have taken before the iterative conception was settled upon. The details will be familiar to specialists, but some remarks help to illustrate the difficulty.

The *limitation of size* conception, under which sets are all those *collections* which are not 'too big' (in a sense that can be made precise in close but inequivalent ways) was, as pointed out by Potter, a lot closer to the early set-theorists' intents and ideas.[13] This may partly be an element of sociological luck, but may also be mo-

[6]For example, in the Gibbs lecture (1951), Gödel says: "I wish to repeat that 'analytic' here does not mean "true owing to our definitions", but rather "true owing to the nature of the concepts occurring [therein]", in contradistinction to "true owing to the properties and the behavior of things" ([Gödel, 1990], p. 321).

[7]Under the iterative conception (very much the 'standard' choice of conception of set), sets are formed in stages by iterating some well-defined operation (usually power-set) along the ordinals.

[8]Though it is possible that Boolos later changed his mind, see [Boolos, 1989] and [Boolos, 2000].

[9]See, in particular, [Potter, 2004], p. 211-237.

[10]See [Feferman et al., 2000], pp. 405-6.

[11]See, for example, [Drake, 1974], or [Enderton, 1977]. These examples can be multiplied.

[12]For a careful reconstruction of the emergence of the **ZFC** axioms, see [Ferreirós, 1999], in particular, Ch. 9 and 11.

[13]See here [Potter, 2004], §13.5.

tivated by the fact that early set-theorists had no developed conception of the well-founded hierarchy. Detailed discussion of well-founded sets appear for the first time in [Mirimanoff, 1917], there are shades of the iterative conception in [Zermelo, 1930] and Gödel's presentation of L, but a full and precise account of the iterative conception was not isolated until [Shoenfield, 1967] and [Boolos, 1971]. This contrasts sharply with the limitation of size conception, which had already appeared in the work of Cantor.[14]

Moreover, these are not the only conceptions of sets, historically. Gödel is also credited to have occasionally expressed the view that all axioms of sets should reduce to just one, Ackermann's Axiom, that the 'Absolute is unknowable'.[15] If this represents Gödel's thought correctly, then it would seem that Gödel's views further evolved to include mentioning of the absolute infinite as part of the concept of set (something which is not directly implied by the iterative conception). It is, also, plausible, then, to conjecture that Gödel saw Reflection Principles, which can be accounted for very easily using such a conception, as the most general axioms of set theory.[16] A fourth salient alternative is the conception of sets as given by the extensions of definite concepts (the so-called *logical conception*), which again appears in the work of Cantor[17] and Frege, whose conception, in turn, has been recently taken up and further investigated by other authors (who have tried to isolate its consistent fragments and base several theories upon it).[18]

So we have four main specifications of the concept of set at hand here: one is based on the 'construction' of all sets iteratively, one on the idea that sets are entities that are not 'too big', another one takes sets to be the knowable portions of an unknowable 'absolute', and one views sets as given by extensions of definitions or properties.

Of course, alternative conceptions may motivate alternative axioms. For example, Reflection Principles appear naturally suggested by the absolutist conception, whereas it is not clear that other conceptions, such as the 'logical conception', sanction their introduction, and the 'limitation of size conception' seems to altogether discourage their use (since they imply that there are 'V-like' sets in a certain precise sense). Similar remarks can be made about other axioms.[19] Looking forward, there are competing ideas we might use in sharpening the iterative concept of set: for instance, our conception might develop in such a way as to include some idea of the universe as being 'orderly' and 'L-like' (as in Woodin's Ultimate-L programme),[20] or, alternatively, as being 'forcing-saturated'[21], or, finally, it might incorporate stabil-

[14] See, in particular, Cantor's famous 1899 letter to Dedekind, [Cantor, 1899].

[15] See [Wang, 1996], p. 283.

[16] Gödel's intentions seem to be oscillating on this point: it seems that, earlier, he had claimed that the iterative conception alone could already justify Reflection Principles and thus the existence of several 'small' large cardinals including inaccessibles, Mahlos, etc. However, the later Gödel, on the contrary, may have held that the existence of an absolute infinite represented the most genuine justification of Reflection Principles.

[17] Cf. Cantor's remark that a set is a "many, which can be thought of as one, i.e., a totality of definite elements that can be combined into a whole by a law" in [Cantor, 1883], p. 916.

[18] See [Incurvati and Murzi, 2017] (though the authors also present difficulties for the 'logical conception') as well as the $\mathbf{NF}(\mathbf{U})$-based theories in [Holmes, 1998] and [Forster, 1995].

[19] We lack the space to give a full survey of case studies here, but note, by way of example, that: (1) the Power-Set Axiom seems problematic from a 'limitation of size' perspective, follows immediately from the iterative conception, and it is unclear whether it could be justified by the 'absolutist' and 'logical' conceptions; (2) Replacement seems more dubious from the iterative conception, positively implied by the limitation of size conception, and unclear on the absolutist/logical conceptions. See [Hallett, 1984] and [Potter, 2004] for discussion of some of these issues.

[20] See [Woodin, 2017] for the state of the art.

[21] See [Magidor, U] for exposition of this conception.

ity of structure existence under extension[22]. It is unclear how we could select one of these sharpenings on the basis of our current concept of set; all seem like natural and legitimate possible future trajectories for our set-theoretic practice.

So we have two main issues making the notion of 'intrinsic' intractable: one is the absence of a clear methodology through which one can ascertain that a statement is 'derived' from the concept of set; the second one is the fact that there are many conceptions of set, all of which may give rise to different versions of 'intrinsicness', and, thus justify some axioms rather than others on intrinsic grounds.

2.2.2 The intractability of 'extrinsicness'

The problem with extrinsic justifications is that practically *all* new axioms are 'successful' somehow. As it turns out, 'success' may be a very volatile criterion. For example, consider:

Axiom of Constructibility. $V = L$.[23]

As is well-known, the power of $V = L$ is stunning. Under $V = L$, the Generalised Continuum Hypothesis (and, thus, the Continuum Hypothesis) are decided, Suslin's Hypothesis is decided, important combinatorial principles (such as $\diamondsuit$) hold, and $V = L$ also implies that there are no *measurable* cardinals. $V = L$ has far-reaching consequences, and imposes a clear structure and conception of V as given by iterated definability.

However, $V = L$ is not regarded as correct by many set theorists. One reason is that it is incompatible with certain large cardinal hypotheses, and the latter are also supposed to be very successful and fruitful axioms.

Another more controversial example is the:

Axiom of Determinacy (AD**)**. Every two-player game G_A on $A \subseteq \omega^\omega$ is determined.[24]

AD once again might be viewed as very fruitful. In particular, under AD, sets of reals are all *well-behaved*, that is, they are Lebesgue measurable, have the Baire property and all have an uncountable perfect subset, that is, under AD we have an optimally informative picture of the real continuum. However, as is known, AD is incompatible with the Axiom of Choice, so $\mathbf{ZF} + \mathsf{AD}$ is highly non-conservative over $\mathbf{ZFC}$ (in fact, inconsistent with it!). Now, AC also seems to have fruitful consequences, so how should one, *in practice*, make a choice between these two, based on purely 'extrinsic' considerations?

There are thus different incompatible candidates for axioms, all of which can be viewed as fruitful from some perspective. This calls into question the epistemic

[22] See [Arrigoni and Friedman, 2013] for a survey of this idea.

[23] L is the *constructible* universe, which has the same ordinals as V, but wherein, at successor levels, only *definable* (in a technical sense) subsets of the previous level are formed (contrary to what happens in V, where, at successor levels, *all* subsets of the previous level are formed) and, at limit levels, unions of all previous levels are formed.

[24] For details here, see [Jech, 2002] Ch. 33. A quick recap of the relevant definitions: let G_A be the following game on $A \subset \omega^\omega$: two players, I and II, play, in turn, natural numbers. The resulting sequence of the choices of I and II may or may not be in A. If the former is the case, then I wins, otherwise II wins. A winning strategy is a strategy which makes one of the two players win. A game is determined if there is a always a winning strategy. AD is the statement that every game G_A is determined. The Axiom of Projective Determinacy is AD restricted to projective sets of reals (for a definition of a projective set, see Jech, [Jech, 2002], p. 144), $\mathsf{AD}^{L(\mathbb{R})}$ is AD restricted to $L(\mathbb{R})$, the smallest inner model of V containing all reals.

usefulness of solely extrinsic justification, as we can nearly always find a way for a particular axiom to be successful. We therefore need to provide a more detailed philosophical account of how and why these notions should provide justification.

One attempt to deal with this problem, probably one of the most concerted attempts to build a clear philosophical underpinning for extrinsic justification, comes from [Maddy, 2011], who remarks that, as it stands, extrinsic justification is often applied "willy-nilly" to any justification that is not clearly intrinsic, and thus more precision is required.[25] Recalling an earlier quotation, Maddy describes extrinsic justifications precisely as those which identify that mathematics is *effective*, *fruitful*, and *productive*.[26] These are then taken to track facts concerning *mathematical depth*, which are, in turn, supposedly *not* subjective:[27]

> "It also bears repeating that judgments of mathematical depth are not subjective: I might be fond of a certain sort of mathematical theorem, but my idiosyncratic preference doesn't make some conceptual or axiomatic means toward that goal into deep or fruitful or effective mathematics" ([Maddy, 2011], p. 81)

Perhaps, then, the non-subjectivity of these depth-facts can rescue the friend of extrinsic justification from the charge of intractability?

We find Maddy's appeal to mathematical depth to be at least as problematic as the notion of merely 'fruitful' or 'successful' mathematics was in the first place. What appears *deep* seems, to us, to be a highly agent-sensitive matter.[28] While it might be the case that this is simply a situation in which our intuitions and Maddy's clash, there is some evidence from the cognitive sciences that seems to indicate that the difficulty might run deeper.

For instance, Inglis and Aberdein have found that, when presented with a proof, mathematicians' ascriptions of evaluative terms for proofs vary along four main dimensions in a similar way to judgements of personal character vary across five dimensions.[29] They term these dimensions 'aesthetic', 'utility', 'intricacy', and 'precision'. These terms were chosen by the authors, but are fairly illustrative of the kinds of terms included in each (for example 'beautiful' was an aesthetic term). 'Deep' correlated strongly with the aesthetic dimension, and 'effective' and 'fruitful' strongly with the utility dimension.

On these grounds, we can raise two problems for Maddy's conception of extrinsic justification as fruitful mathematics. First, the empirical evidence may suggest that 'fruitfulness' and 'depth' are not even measuring the same dimension of human thinking (at least as far as research-level mathematicians' use of language is concerned). Second, further work indicates that mathematicians strongly disagree with one another on whether particular instances of proofs are fruitful. When presented

[25] See [Maddy, 2011], p. 130.

[26] For the sake of the reader we repeat it in this footnote:

> "It has become customary to describe these two rough categories of justification as 'intrinsic'—self-evident, intuitive, part of the 'concept of set', and such like—and 'extrinsic'—effective, fruitful, productive." ([Maddy, 2011], p. 47)

[27] In fact, for Maddy, intrinsic justifications also track facts about mathematical depth: sets are understood as tracking the 'topography' and 'contours' of mathematical depth. However, as we've noted, Maddy is circumspect about the extent to which the intrinsic justifications can take us beyond the most basic of axioms, so we focus on mathematical depth as a possible source of support for the use of 'extrinsic' justifications.

[28] We thank [name removed for blind review] for discussions concerning the possible agent-sensitivity of mathematical depth.

[29] See [Inglis and Aberdein, 2015]. For the literature on personal character, see [Donellan et al., 2006].

with an (anonymised) proof from *Proofs from THE BOOK*, mathematicians appraisal of the proof varied wildly.[30] In Inglis's and Aberdein's words:

> "We found a remarkable level of disagreement between our participants' ratings of the proof. For each of the four dimensions of proof appraisal there were participants who thought the proof should score high on that dimension, and there were participants who thought the proof should score low on that dimension. Furthermore, neither research area nor career stage seemed to be predictive of mathematicians' appraisals on any of the four dimensions." ([Inglis and Aberdein, 2016], p. 10)

Thus, even if we could settle on a particular phenomenon that is being picked out by the term 'fruitfulness' in Maddy's characterisation of extrinsic justification, it is highly unclear that there is a non-subjective sense to the term 'mathematical depth', or 'fruitfulness'. At the very least, current research-level mathematicians' usage of the terms is not close to being coextensive, at least insofar as proofs are concerned.

Of course, we should be careful as to what we take the above observations to have established. Any strong philosophical conclusion extracted from empirical data needs to be treated with scrutiny, both with respect to the experimental methodologies employed, and to how these might be taken to connect with the philosophical phenomena. It may be that the tension in Maddy's account can be resolved, or that more work would result in a satisfactory sharpening. However, as it stands, it is hard to see how we can use claims of mathematical depth to move forward on the 'extrinsic' justification of new axioms if there is (as a matter of empirical fact) little agreement on whether particular proofs are deep or not.

To sum up, intractability seems to affect both intrinsic and extrinsic justification, when conceived of as fundamentally distinct and competing kinds. We now turn to the problem of *demarcation*.

2.3 Demarcation

An other problematic aspect of the standard view on justification is that the force of an intrinsic justification cannot be neatly separated from that of an extrinsic one. In the set-theoretic context even in cases where one might have the impression that such a demarcation has been clearly drawn are, on closer inspection, difficult to demarcate. Whilst there is a sharp distinction between a principle *itself* and its *consequences*, this is not so for the justificatory force attaching to intrinsic and extrinsic justifications.

We could summarise the situation in the following way:

> **Demarcation Problem in Set Theory.** Is there a sharp boundary between the justificatory force associated to intrinsic, on the one hand, and extrinsic, on the other, justification? If so, how should this boundary be characterised?

Now, as we'll outline in more detail below, we believe that the answer to the first question is negative. The key issue is that the justificatory force of intrinsic and extrinsic justifications seems, to our minds, not to be sufficiently well-separated. Thus an account of justification that sees one kind as privileged fails to be transparent

[30] See [Aigner and Ziegler, 2009]. *Proofs from THE BOOK* is a volume of the supposedly most 'beautiful' proofs of various theorems. Interestingly, if participants were informed of where the proofs came from, their appraisals were modified (as one might predict).

with respect to the grounds of asserting the 'axiom'. In this way, such accounts fail to be epistemically useful. In this section, we'll make a case for this in more detail, presenting historical case studies in which extrinsic claims seem to be intimately related to intrinsic considerations.

2.3.1 Zermelo on the Axiom of Choice

Even if Zermelo did not go as far as Cantor in asserting that the *well-ordering theorem* was a *law of thought*[31], he did find the success of AC to be linked to its *intrinsic* justification. In the 1908 paper, where he discusses the objections that his explicit use of AC raised, Zermelo says that:

> "..this axiom, even though it was never formulated in textbook style, has frequently been used, and successfully at that, in the most diverse fields of mathematics, especially in set theory, [...]. Such an extensive use of a principle can be explained only by its *self-evidence*, which, of course, must not be confused with its provability. No matter if this self-evidence is to a certain degree subjective – it is surely a necessary source of mathematical principles." [32]

Zermelo thus seems to surmise that the *success* of the Axiom of Choice is precisely the indication that the axiom is *self-evident*. Therefore, one could indirectly reconstruct Zermelo's ideas on justification in the following way: one starts with conjecturing the self-evidence of some principle, and then verifies that the principle is successful, something which is, ultimately, taken as an indicator that the principle is *really* self-evident (which was precisely what we were trying to establish). Here we have a lucid case where there is no obvious distinction between the justificatory force of intrinsic and extrinsic justifications: because of this symbiotic back and forth between success and self-evidence, the two seem inextricably linked, and so it is impossible to demarcate the justificatory force of each from the other.[33]

2.3.2 Measurable cardinals and $V = L$

As seen in the previous subsection, $V = L$ can be understood (roughly put) as the claim that every set can be constructed via iterated definability (incorporating a restriction on the parameters used).[34] Now, whilst Gödel did consider the suggestion that $V = L$ should be added as an axiom (referring to it as "natural"), he ultimately felt that it was not justified, and in fact should be regarded as false for intrinsic reasons.[35]

[31] Cf. Cantor, [Cantor, 1883], in Ewald, [Ewald, 1996], p. 886: "In a later article I shall discuss the law of thought that says that it is always possible to bring any well-defined set into the form of a well-ordered set—a law which seems to me fundamental and momentous and quite astonishing by reason of its general validity."

[32] [Zermelo, 1908], p. 187. Maddy also quotes this reference, but holds that it is indicative of Zermelo's despairing of giving a precise content to intuitive justification (see [Maddy, 2011], p.46 and an earlier (if more neutral) use in [Maddy, 1988a], p.487). Interestingly, in [Maddy, 1988a], Maddy regards this as providing only *intrinsic* considerations, despite the mention of 'success'.

[33] One objection might be that to argue that Zermelo meant to assert that the fact the Axiom of Choice is *widely* used rather than *successfully widely* used is indicative of its self-evidence. On this objection, it is simply the breadth of the use, rather than the successfulness, that indicates self-evidence. (We thank [name removed for blind review] for pressing us on this point.) We find this elimination of any notion of success problematic since it fails to exclude intuitively attractive but false principles like Naive Comprehension.

[34] See footnote 23.

[35] Cf. here [Gödel, 1938], p. 557: "The proposition [$V = L$]. . . added as a new axiom, seems to give a natural completion of the axioms of set theory, in so far as it determines the vague notion of an

Gödel's sentiment is one shared by many set-theorists. L is the 'smallest' possible inner model (i.e. model containing all ordinals) and in this sense, does not seem to mesh with the idea that the power-set operation should be as 'rich' as possible. Moreover, the restriction on *bounded* parameters in the definition of the hierarchy seems at odds with the impredicative intuitions underlying higher set theory. In this sense, $V = L$ seemed intuitively *restrictive*, and should be rejected on intrinsic grounds.

In Gödel's own words:

> "..from an axiom in some sense opposite to this one [i.e. $V = L$], the negation of Cantor's conjecture could perhaps be derived. I am thinking of an axiom which (similar to Hilbert's completeness axiom in geometry) would state some maximum property of the system of all sets, whereas the axiom A [i.e. $V = L$] states a minimum property. Note that only a maximum property would seem to harmonize with the concept of set explained in footnote 14. [i.e. the iterative conception]" ([Gödel, 1964] pp.262–263)

Shortly afterwards, *consequences* of $V = L$ were discovered, that seemed to reinforce the idea that it represented a restrictive principle. Consider:

Theorem 1. [Scott, 1961] Assume there is a measurable cardinal. Then $V \neq L$.

Thus the intrinsic idea that $V = L$ should be rejected as it is restrictive could now be made fully persuasive by using extrinsic considerations, since $V = L$ prevents the existence of certain large cardinals. But intrinsic considerations (such as those appearing in Gödel's quote) already suggested that $V = L$ would naturally clash with 'maximum' principles (such as the existence of a measurable cardinal). Intrinsic and extrinsic considerations are thus interacting and reinforcing one another here, and cannot be neatly demarcated.[36]

2.3.3 Large cardinals and the case for axioms of definable determinacy

More recent work on axioms of definable determinacy yields further examples. As we shall argue below, here we have a case where after one has found good extrinsic grounds for the justification of a set-theoretic principle/axiom, then one finds that there may have been an *intuitive* conception underwriting the relevant set-theoretic principle/axiom from the beginning.

Axioms of definable determinacy and their connections with large cardinals constitute one of the most thriving areas of research in set theory.[37] As seen, these ax-

arbitrary infinite set in a definite way.". For more on the notion of 'naturalness' in set theory, also see [Venturi, 2018]. For his change of mind, see [Gödel, 1964] (pp.262–263), quoted below, where he suggests that $V = L$ is minimising and only a maximising principle would harmonise with our concept of set.

[36]There may even be a case to be made that Gödel's views on $V = L$ and its intrinsic plausibility were directly influenced by the Scott result. In particular, if one looks at sections 3 and 4 of the [Gödel, 1947] (see pp. 179–186 of [Gödel, 1990]), we find a mention that perhaps an axiom opposite to $V = L$ would result in a proof of $\neg$CH. However, in the same part of [Gödel, 1964] (pp. 257–264 of [Gödel, 1990] esp. footnotes 20 and 23), he mentions the Scott result, and then goes on to claim that only an axiom opposite to $V = L$ would harmonise with the concept of set (a stronger claim than in 1947). Without further textual evidence, we relegate this point to a footnote. Nonetheless, future developments (such as a greater availability of Gödel's unpublished papers) may deliver additional evidence for this claim, which would greatly strengthen our case.

[37]For an overview of the main mathematical results and concepts, see [Koellner, 2006] and [Koellner and Woodin, 2010].

ioms, such as Projective Determinacy and $\mathsf{AD}^{L(\mathbb{R})}$, prescribe that there are winning strategies for two-player games of perfect information.[38]

The pioneering work of set-theorists in the 1980s showed that $AD^{L(\mathbb{R})}$ is implied by strong large cardinal hypotheses at the level of many Woodin cardinals. The direct converse does not hold, yet definable determinacy axioms imply the existence of inner models with comparable large cardinals. These facts have been interpreted in many ways. No one has questioned the fertility of the hypotheses under consideration, but as to their intrinsic justification, opinions have been somewhat pessimistic:

> "Is PD true? It is certainly not self-evident." ([Martin, 1977], p813)[39]

However, ways to provide intrinsic support for those axioms have been hinted at by some authors. For example, [Koellner, 2014] points out that there are close relationships between determinacy axioms and large cardinals (since one can prove determinacy axioms from large cardinals, and reverse from determinacy axioms to inner models with large cardinals). This idea has also occurred in [Hauser, 2001], which, in addition, argues that the fact that two different and powerful strands of set-theoretic research ultimately converged in a *unified* structure theory is an illustration of a criterion he advocates as a fundamental source of 'internal' evidence in favour of a set-theoretic axiom, *identity through differences*. He says:

> "A particularly striking example in set theory is the aforementioned subtle relationship between large cardinal axioms (global existence postulates motivated in part by a priori considerations about the inexhaustibility of the universe of all sets) and axioms of determinacy (local principles justified by their fruitful consequences in second-order arithmetic). [...] Whether this *coherence* necessarily reflects the existence of a mind-independent realm of sets cannot be analyzed any further here. Its main significance from the viewpoint of methodology (our primary concern) is that it confers objective validity to both kinds of axioms, [...]." ([Hauser, 2001], p. 257)

Thus on Hauser's conception of justification, the fact that determinacy *and* large cardinal axioms represent a somehow unified phenomenon confers to such axioms also an intuitive appeal, which is, ultimately, linked to their expressing features of the 'structure' of V. One could push the point even further and imagine that, as in the previous example of Zermelo's AC, the *intrinsicness* of determinacy axioms/large cardinals has, in a sense, been confirmed by the progressive unfolding of their fruitfulness.

2.3.4 Summary

In the above examples there appear to be difficulties in demarcating intrinsic and extrinsic justificatory force. In our view, this amounts to the following preliminary conclusion: looking at the historical concrete cases of the justification of new set-theoretic principles we cannot draw a clear boundary between the force of intrinsic considerations and that of extrinsic considerations. Combined with our observations concerning the tractability of the distinction, this questions the epistemic usefulness of regarding the two kinds justification as neatly separable and competing.

[38] See footnote 24.

[39] Similar comments by Martin (and others) can be found, for example in [Martin, 1976] (p. 90) and [Moschovakis, 1980] (p. 610). See [Maddy, 1988a] and [Maddy, 1988b] for discussion.

3 Justification and explanation

We now find ourselves in something of a predicament. On the one hand, set-theorists and philosophers of set theory seem to use something like intrinsic and extrinsic reasons in providing justification for the use of different axiom systems, but on the other hand the distinction between the two kinds of justification suffers from problems, both at practical and theoretical levels.

Our solution will be to provide an account of justification in set theory that recognises the presence of different sources for the force of an argument, but that does *not* regard the intrinsic or extrinsic as fundamental regarding justification. The problem of demarcation, thus, will become feature (rather than weakness) of this account, and a response to the problem of tractability will be facilitated, offering some conditions for the selection of new axioms, that in our view perform better then the competing accounts.

Our position can be summarised in the following way:

> **The Explanatory Account of Justification.** Justification of axioms in set theory consists of finding the best explanations for relevant mathematical data.

Of course, this needs to be made far more precise to not suffer from similar problems as we identified for the standard accounts of intrinsic and extrinsic justification. It is to this task that we now turn. First, we will explain the details of the view, and why set-theoretic justification has this nature. Then we will explain how it can be used to address the problems of demarcation and tractability, providing case studies for the latter.

3.1 Beginnings of the explanatory account of justification

In this subsection we describe what a theory of explanatory justification in set theory amounts to and how it produces a cogent picture of set-theoretic justification. Our basic contention is that we can make sense of axioms as the best explanations of particular mathematical *data*, and thus that the best-justified axioms are the most *explanatory* ones. We thus have two tasks before us:

(1.) Outline and articulate a conception of mathematical data.

(2.) Provide an account of explanation in relation to this data, how it is related to justification, and how it avoids the problems of tractability and demarcation.

We tackle these problems in order.

3.1.1 Mathematical data

An essential component of our view is a notion of *mathematical data*. Accounts of data in the philosophy of science often presuppose that data are 'made'. For example, Hacking argues that data are made with recording marks that are obtained by human interactions with various kinds of devices.[40] Rheinberger takes data to be the result of subsequent manipulations of Hacking's 'marks', rather than the marks themselves.[41]

[40] See, for example [Hacking, 1992].

[41] See here [Rheinberger, 2011].

All this raises a challenge for the philosophy of mathematics, since there are not clearly any 'marks' that are obtained by recording devices in the mathematical context. However, a wider view of data (one that leaves it open whether or not the data are obtained via marks) is available in the work of Sabina Leonelli, who defines data as follows:

> "I propose to view data as any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, which is collected, stored, and disseminated in order to be used as evidence for knowledge claims...Hence, any object can be considered as a datum as long as (1) it is treated as potential evidence for one or more claims about phenomena and (2) it is possible to circulate it among individuals." ([Leonelli, 2015], p. 817)

Leonelli's view is designed to account for the *mobility* of data, in that some data may be processed, modified, and manipulated into different formats, and then shared and used across multiple communities. Further, she argues, there is no clear difference between 'data' and the traces from which they are obtained.

Her loosening of the definition of data though also permits application to a mathematical context. For, in mathematics and its philosophy we also have products of research activities that may be used in evidence of knowledge claims; namely the theorems we have derived so far from our currently accepted axioms. Thus we propose that the mathematical data available to an agent or community comprises at least the following:

> **Mathematical Data.** The *mathematical data* available to a community or agent at a time consists of (at least[42]) the body of axioms and theorems accepted by that community at that time.

Several remarks are in order concerning this account of mathematical data:

First, the mathematical data may differ between agents and communities, and indeed across time. This is a *good thing*, we should not expect the justificatory challenges and relevant evidential base to be invariant across diverse contexts. The questions of justification are very different for those studying subsystems of second-order arithmetic as compared to those working on resolutions of the Continuum Hypothesis. Similarly questions of justification were very different for the early set theorists as compared to our current set-theoretic epistemological state.[43]

Second, we note that an individual datum is defeasible. We may take something to be a datum at one particular time that is subsequently removed since it is shown to be false given our other more entrenched theoretical commitments. This can happen, for example, when a theorem (or indeed axiom) is shown to be false, or the proof flawed. A good example of the former is the rejection of Naive Comprehension. The latter phenomenon is common in mathematics, but a good example from set theory is [Džamonja and Shelah, 1999] which claimed to have shown that there are models of

[42] We leave it open that there may be more data. For example, in particularly applied areas of mathematics, non-theorem-like observations may play a role.

[43] A possible objection here is that it is unclear how mathematics got going given this account of mathematical data. What about the earliest mathematicians, who had no clear body of theorems that were taken as accepted by any particular community? We set this aside for two reasons: (1.) We think that even for these mathematicians it is likely that they possessed a body of mathematical or quasi-mathematical data, even if it was relatively primitive. These might include components of basic computations in arithmetic (e.g. "one object taken together with a different object always yields two objects"), and (2.) Even if our account does not apply to *those* mathematicians, the fact still remains that agents and communities since at least the scientific revolution *do* have a core bank of accepted mathematical premises to work from.

set theory in which both ♣ (a particular combinatorial principle) is true but there are no Suslin trees. Their methods, however, contradicted a well-established theorem (namely Miyamoto's Theorem) and so the proof was recognised to be flawed[44], and the (now open) question removed from the established mathematical data.[45]

Third, it is important that the data need not be *interpreted* (or indeed even truth-evaluable). There is not (without further argument) any a priori reason why the *acceptance* of a particular mathematical datum (say that the power set of the natural numbers has a greater cardinality than the natural numbers) implies the *truth* of any claim or that the datum expresses a particular *fact*. Rather, data encapsulates what is accepted as requiring systematisation by a foundational framework. So, for example, a Platonist (who actually believes there are sets) and a fictionalist (who believes that strictly speaking our mathematical claims are all false, but *accepts* set theory as a correct fiction) can coherently have a conversation about mathematical justification on the basis of a shared data set, even if they vehemently disagree on the *interpretation* of that data. Again, we regard it as a desirable feature of our view that it is not beholden to a particular account of mathematical truth and ontology.

For now, we will assume that the data pertaining to set-theoretic mathematics and justification constitutes *at least* **ZFC** and the currently accepted theorems (we will discuss some extensions shortly). Explanation can then be understood as an epistemic virtue attaching to particular coherent pictures systematising our data. Thus justification is a particular kind of process on which certain statements are taken as *basic*, and explanations thereby sought.

3.1.2 Explaining the data

With an account of mathematical data in hand, we turn to the notion of *explanation*. There is already a substantial literature on mathematical explanation in the philosophy of mathematical practice[46], largely centring on the explanatoriness (or otherwise) of particular proofs. From the outset, we should emphasise that the kind of explanation we will be considering is rather different. Instead of concerning particular proofs, it will rather be similar to the notion of explanation found in the philosophy of science, where there is a rich literature on what the explanation of data might be (such as in the discussion of laws of nature).[47]

We can begin to flesh out our account by defanging a natural immediate objection, given the initial bare-bones statement of our account. It goes as follows: If axioms are to be inferred by inference to the best explanation, then we might be subject to a tu quoque, namely that explanation is just a problematic notion as depth (let alone 'best' explanation). This is especially so when we bear in mind that [Inglis and Aberdein, 2016] showed that the notion of a proof being *explanatory* exhibited a similar level of diversity in appraisal by research-level mathematicians as that of depth. So, in what sense is incorporating explanation into our account better than depth?

Two points should be made immediately: First (as noted above) we are aiming at a notion of explanatoriness independent of particular proofs (in this sense

[44]It bears mentioning that [Džamonja and Shelah, 1999] contains much useful material, even if one result fails to go through.

[45]See [Brendle, 2006], p. 45, footnote 1 for some discussion and further references. We thank [name removed for blind review] for bringing this example to our attention.

[46]See, [Mancosu, 2008], for a general presentation.

[47]There is, however, a deep question of how a notion of explanation independent of proof might be related to the idea that certain proofs merely prove, whereas others also explain. [Lange, 2017] represents an in depth study of some of these ideas.

[Inglis and Aberdein, 2016] figures more into questions concerning proofs). It is thus open whether or not *our* account of explanatoriness is open to the same kinds of issues in appraisal as proofs, given that it is of a different kind.

If we can provide such an account of explanation, it might have certain advantages over a theory of axioms based on depth. First, one might think that mathematical explanation is epistemically useful for the selection of new axioms, whereas depth might not be. This goes for many ways of interpreting explanation: Philosophers often distinguish (in the broader context of scientific explanation) between *ontological* explanation (features of the world that explain each other) and *epistemic* explanation (how our mathematical experiences can be organised in terms of explanation).[48] Given either kind of explanation we can see that either interpretation of explanation is epistemically relevant for the choice of axioms; the first identifies the most fundamental features of mathematical reality and the second ascertains the key features of our practice that helps us to organise our experience of mathematics. Either way, and whether or not one leans more towards one of the two kind explanation (or a mixture of both), we have salient features of mathematics that are useful for the selection of axioms. It is comparatively unclear how *depth* is meant to figure into an epistemic story. The same is not so clear in the case of mathematical depth where it is at least possible (without further argument) that a piece of mathematics be deep without being particularly epistemically significant. [49] So, in this sense *even if* explanation is a slightly problematic notion, if we can address these philosophical problems and provide a sharper characterisation, we will automatically have something epistemically useful.

This of course is only to offer the possibility of a way out, rather than actually providing one. Whilst we acknowledge that explanation is a difficult notion to pin down, we have *far more* tractability on the notion than that of depth, and *can* sharpen it into something epistemically useful. Whilst we do *not* have necessary and sufficient criteria for when a statement is the best explanation (and indeed these might not exist), we *can* at least point to precise features that increase our *confidence* that it has the character of a good explanation. Two obvious examples, which will serve to illustrate the general strategy, are:

(1.) The sentence (or scheme) should obviously be consistent in the background logic.

(2.) Given some especially entrenched mathematical data, the principle should not contradict these. For example, basic number-theoretic facts or simple (and now known) propositions of analysis should not be contradicted.

Whilst these two examples of criteria do not get us very far (they are in some sense the *minimal* requirements on a sentence providing an explanation of some accepted mathematical data), they serve to outline the broad strategy, namely: There is no problem of tractability with constraints of the above kind. Consistency is a technical notion that can be unambiguously defined. Of course it may be that we can never be *certain* that a theory is consistent (given Gödelian considerations), but nonetheless there is no ambiguity as to *whether* a theory is consistent, and often we have various reasons for accepting the consistency of theories. For example, the existence of a rich structure theory (in the case of **ZFC** and its extensions, the ones provided by L and

[48] See, for example, [Salmon, 1984].

[49] Of course if one is already on board with [Maddy, 2011] that sets *just are* the markers of mathematical depth, then of course depth is epistemically relevant. We see no clear reasons to accept this claim.

other fine-structural inner models[50]) and intuitive motivating picture (such as stage theory for **ZFC**) give us confidence that a theory is consistent. Concerning the second constraint, once we have settled on what the more basic mathematical data are, consistency with these data can also be assessed on similar grounds. Of course, there may be some debate as to what are to count as the basic data around the peripheries (or across research communities and time), but for the purposes of mainstream classical mathematics and set theory we can settle on a core of widely accepted axioms and theorems. This is quite similar to theoretical physics, where the interpretation of experimental data can be challenged around the peripheries, such as in the recent debate (and subsequent identification of experimental error) with respect to faster than light neutrinos, but there is nonetheless a core of accepted physical data. Similarly we may remain circumspect as to whether a purported proof in mathematics contains a subtle flaw. Despite these wrinkles we can take some data to be basic (currently this probably comprises at least **ZFC**) that can only be challenged in extreme circumstances.

Certainly, however, these criteria only serve to weed out the really *bad* putative explanations. Can we go further?

3.2 Prediction and explanation

The core claim we shall argue for is that there is a precise sense of prediction and verification in set theory that counts in favour of principles and is related to explanatory considerations. Shades of this idea in fact already appear in [Maddy, 1988a] regarding reflection (p.503), in [Maddy, 1988b] in a summary of different kinds of evidential support (p. 758–759), in [Maddy, 2011] (Ch. V, esp p. 127) discussing some remarks of [Martin, 1998] (p. 224) concerning the Cone Lemma, and in [Koellner, 2010] (§1.5 and p. 204) for the justification of determinacy axioms. We will develop these ideas, in particular providing the following additional contributions: (1.) We identify an additional case in which a notion of prediction and confirmation occurs, strengthening the case that this is an integral part of set-theoretic justification, and (2.) We will argue that, contrary to previous accounts, prediction and confirmation are not solely 'extrinsic', but mainly *explanatory*, integrating our justificatory enterprise with the account of mathematical data provided above.

We can simply state our third condition thus:

(3.) A principle has a better claim to being an axiom (i.e. good or best explanation) if it predicts (new) mathematical data.

Again, what 'predicts' comes down to here is a difficult question. However, we can provide a *fully* technically precise account of it by revisiting Gödel talking about the consequences of a new axiom:

> "...in particular in "verifiable" consequences, i.e., consequences demonstrable without the new axiom, whose proofs by means of the new axiom, however, are considerably simpler and easier to discover, and make it possible to condense into one proof many different proofs." ([Gödel, 1947], p. 182)

[50]These models provide contexts of study for large cardinal principles which yield a vast amount of information, for example, most fine-structural inner models satisfy principles such as GCH as well as combinatorial principles like $\diamondsuit$ and $\square$. This is argued (e.g. by [Steel, 2014]) to constitute evidence in favour of the consistency of the relevant principle, since we have a detailed picture of a structure in which it holds. In the words of Steel "a voluble witness with an inconsistent story is more likely to contradict himself than a reticent one." ([Steel, 2014], p. 156).

Though Maddy and others do mention prediction, they often take the consequences Gödel is interested in to be "nice" in some appropriate sense, constitutive of a notion of extrinsic justification which we argued earlier to be problematic. However, notice that Gödel here talks about consequences "*demonstrable without the new axiom*". We can then come to a precise account of prediction and confirmation; the prediction of a stronger and more controversial theory should be verified by an accepted and weaker one. For example, we might confirm some axiom Ψ by proving some unknown statement ϕ from $\mathbf{ZFC} + \Psi$, and then subsequently verifying ϕ in **ZFC**. This would yield confirmation of Ψ, and is a fully technically precise notion.

Does such prediction occur in mathematics? We will now argue that it does using two case-studies.

Our first example concerns the *Cone Lemma* and is in fact discussed in [Maddy, 2011], but fits especially well with our current purposes. The Cone Lemma states that from AD one can prove that for any set A of Turing degrees either A or its complement contains a cone. This holds, mutatis mutandis, for Projective, Open, and Borel determinacy. In the case of PD, we have that PD implies that every projective set of Turing degrees either contains a cone or its complement contains a cone. Now, during the study of determinacy axioms, Martin in fact tried to show that a *contradiction* with **ZFC** resulted from projective determinacy, as he discusses:

> "When I discovered the Cone Lemma, I became very excited. I was certain that I was about to achieve some notoriety within set theory by deducing a contradiction... In fact I was pretty sure of refuting Borel Determinacy. I had spent the preceding five years as a recursion theorist, and I knew many sets of degrees. I started checking them out, confident that one of them would give me my contradiction. But this did not happen. For each set I considered, it was not hard to prove, from the standard **ZFC** axioms, that it or its complement contained a cone...
>
> ...I take it to be intuitively clear that we have here an example of prediction and confirmation. What was predicted, moreover, was not just individual assertions. Though there had been much work on the structure of the degrees, no attention at all had been paid to the notion of a cone. There was one known theorem (Richard Friedberg's 'criterion of completeness'), which we would now describe as showing that a certain set contains a cone. Afterwards cones and calculations of 'vertices' of cones became significant in degree theory. In determinacy theory, the Cone Lemma became an important tool. What was predicted by the Cone Lemma was thus a whole phenomenon, not merely isolated facts. The example seems fully analogous to striking instances of prediction and confirmation in empirical sciences." ([Martin, 1998], pp. 224–225)

Martin's point is that in these contexts PD actually *predicted* phenomena (the existence of certain cones on Turing degrees) that were subsequently *verified* in **ZFC**. This then increases our confidence (ceteris paribus) that PD should be added to our axiomatic framework. Moreover, it does so completely *precisely* in terms of prediction of phenomena (namely the existence of cones) by a stronger, more controversial theory (namely $\mathbf{ZFC} + \mathsf{PD}$, and susequent verification by a weaker one (namely **ZFC**).

Our second and perhaps less well-known example is represented by Dehornoy's work on *braids* (expanding on previous work by Laver, Martin and others). Braids can be defined as collections of disjoint polygonal arcs $\gamma_1, \ldots, \gamma_i$ in $\mathbb{R}^2 \times [0, 1]$.

As it turns out, the study of these simple, *finitary* objects can be more easily carried out using methods involving large cardinals, in particular *elementary embeddings*.[51] Many kinds of large cardinals are most naturally defined in terms of elementary embeddings.

Now, earlier work on a very strong large cardinal hypothesis made it possible to study algebraic operations associated to collections of elementary embeddings, such as *composition* and *product*.[52] Let $\mathcal{E}_\delta = \{j : V_\delta \prec V_\delta\}$ be a collection of such embeddings: one can define operations on elements of $\mathcal{E}_\delta$, in a way which is fully analogous to how one defines operations on elements of the braid group B_n.[53] Dehornoy, then, puts the analogy to work by managing to prove further crucial theorems on braids in an extension of the infinite braid group B_∞ using **ZFC** alone, thus eliminating the need for large cardinals.

In his comprehensive monograph on braids ([Dehornoy, 2000]), Dehornoy acknowledges and emphasises the connection between large cardinals and braids, as well as the crucial role this connection played for his work, using the following unequivocal terms:

> "It seems to us that the role of set theory in such cases is quite similar to the role of physics when the latter gives heuristic evidence for some statements that mathematicians are to prove subsequently. In both cases, the statements are first established rapidly but at the expense of admitting some additional hypotheses or approximative proof methods — observe that adding a set theoretical axiom is nothing but adding a new proof method — and the subsequent task is to give a proof that does not use the additional hypotheses any longer." ([Dehornoy, 2000], p. 600)

Thus, it really seems that we have another example of the notions of *prediction* and *confirmation* at hand. Results on braids, predicted using a stronger theory (**ZFC**+large cardinals) are then verified in **ZFC**, increasing our confidence that large cardinal axioms are well-justified.

One might think that considerations of parsimony could then be brought against such a picture. For, our account of prediction depends upon a particular prediction being subsequently confirmed through proof in a weaker already accepted theory. But if the prediction can be proved in the weaker theory, shouldn't we simply plump for the weaker theory over the stronger one on the basis of explanatory parsimony?

We think that the correct response to this claim is to point out that there are different kinds of parsimony. We acknowledge that in terms of *logical* parsimony, it is the weaker theory that receives the highest praise. However, the ability to systematise wide ranging data, an ability manifested through prediction, tells strongly in favour of the predicting principle in terms of *conceptual* parsimony.[54] This is evidenced by

[51] It is useful to recall some foundamental notions here. Given two structures $\mathfrak{A}$ and $\mathfrak{B}$, an elementary embedding of $\mathfrak{B}$ into $\mathfrak{A}$ (denoted $\mathfrak{A} \prec \mathfrak{B}$) is the isomporphism of $\mathfrak{B}$ into a submodel $\mathfrak{B}' \subset \mathfrak{A}$. Large cardinals of the same strength as, at least, *measurables* may be defined in terms of elementary embeddings. In particular, the existence of a measurable cardinal is equivalent to the existence of an elementary embedding $j : V \to M$, where the least κ such that $j(\kappa) \neq \kappa$, called the critical point of j, is a measurable cardinal. The embedding notions related to braids are stronger refinements of the definition above. Recent work has looked at embedding characterisations of smaller large cardinals, see [Holy et al., S].

[52] The strong large cardinal hypothesis mentioned is I3: 'For some δ there is a $j : V_\delta \prec V_\delta$'. I3 is currently not known to be inconsistent with ZFC. See also [Kanamori, 2009], p. 325.

[53] Full details may be found in [Kanamori, 2009], pp. 329-331.

[54] Ideas similar to this appear in [Cartwright, 1980] and Essay 8 of [Cartwright, 1983]. There Cartwright argues that the truth (which may be messy) does not explain in science, rather we need conceptually simple simulacra that help us to systematise the phenomena. The application there is somewhat different, since this idea of explanation is elucidated in terms of the number of bridge principles employed and we

Dehornoy's remarks above where it is the large cardinals that show the *conceptual* route to the proofs, even if it is subsequently eliminable in **ZFC**.

In other terms, the examples above not only offer a more precise account of prediction and confirmation, but they also show the unificatory power of this perspective on explanation; again along the line suggested by Gödel in [Gödel, 1947][55].

We thus have a further precise criterion on when we might take a principle to be explanatory; namely prediction and confirmation. However, we also need to make convincing the claim that such prediction and confirmation should figure into an account of explanation driving justification. We will argue that there are at least two senses in which this is so.

(1.) Prediction and confirmation shows how a principle can systematise a wide variety of data, and thus increase the chance that the principle *itself* forms a part of explanations.

(2.) A natural explanation of the fact that the principle makes verifiable predictions is that it is *correct*.

We start again by considering Gödel, who was sensitive to the explanatory role axioms have through systematisation. For example, Mehlberg writes the following reporting on Gödel:

> "According to Gödel, an axiomatization of classical mathematics on a logical basis or in terms of set theory is not literally a foundation of the relevant mathematics, i.e., a procedure aiming at establishing the truth of the relevant mathematical statements and at clarifying the meaning of the mathematical concepts involved in these theories. In Gödel's view, the role of these alleged 'foundations' is rather comparable to the function discharged, in physical theory, by explanatory hypotheses. Thus, in the physical theory of electromagnetic phenomena, we can explain why the sky looks blue to us under normal circumstances, and we are even able to produce the same phenomenon in the laboratory. Both the explanation of the physical phenomenon under consideration and its production under laboratory conditions are due to the logical fact that the statements describing the blue of the sky or that of an artificially produced area in the laboratory are *theorems provable* within an axiomatic system the postulates of which are concerned with hypothetical laws governing electro-magnetic phenomena, the composition of the atmosphere, etc. It would not occur to a physicist that these electro-magnetic assumptions which enjoy the role of postulates in an axiomatized, or axiomatizable physical theory, are more dependably known to be true than the prescientific phenomena (like the blue of the sky) which are being explained by being shown to be provable theorems in the aforementioned physical theory. Thus, *the actual function of postulates or axioms occurring in a physical theory is to explain the phenomena described by the theorems of this system* rather than to provide a genuine 'foundation' for such theorems. Professor Gödel suggests that so-called logical or set-theoretical 'foundations'

do not have an obvious division between the theoretical and concrete. However, we *do* hold that the brute provable facts of **ZFC** might not be what provide the best explanations, rather we need the more theoretically elegant theories that incorporate principles of greater consistency strength.

[55]On the unificatory power of explanation in science there is a vast literature. We would like to recall here the contribution of Philip Kitcher [Kitcher, 1981] who also believed in a methodological uniformity between pure and applied science and thus suggested that his view on explanation extended easily to mathematics.

for number-theory, or any other well established mathematical theory, *is explanatory, rather than really foundational, exactly as in physics.*" (Emphasis ours, [Mehlberg, 1960], p. 397)

While we might diverge slightly from the author on the use of the term 'foundation', we see clearly here that if we are to systematise data, the fact that the known (and previously unknown) data can be derived from the axioms is important for showing that the axioms are good explanations in themselves, providing a wide systematisation of diverse data. Thus prediction shows how diverse data are importantly similar in being traceable to a common root.

This has implications for explanation in mathematics. [Lange, 2017], for example, argues that explanation occurs when a proof exhibits an explanatory and striking similarity between two domains.[56] In particular, proofs are understood as explanatory when they exhibit 'symmetries' of natural properties between different objects or structures.[57] In the case of prediction, we have multiple similar facts traceable to a common root. This results in an increased likelihood that such similarities will be found, as the Dehornoy already showed. Therefore, systematising vast swathes of knowledge under a single assumption (as is the case with prediction) increases the confidence that such an explanatory symmetry may be found, even if it does not guarantee it. Thus, prediction increases the likelihood that an axiom forms part of an explanation.[58]

Turning to (2.), we might also think that explanatory considerations independent of the proposed axiom should lead us to the acceptance of axioms that predict. The debate over whether principles that predict (i.e. prove some data without that data in mind) as opposed to accommodate (i.e. prove some data with that data in mind) is well-trodden in the philosophy of science.[59] The key thought is that prediction increases the chance that an axiom is, in some sense, 'correct'. Since we appeal here to a notion of *correctness* the exact form of this relationship will depend somewhat upon the underlying ontology.

For the realist about mathematical objects, we can take standard realist arguments for the correctness of predictivism for the philosophy of science. For example, one might argue (in line with [White, 2003]) that a good explanation for why predictive theories predict is simply that they are more reliably aimed at truth than those that accommodate (the so called 'Archer analogy'). Similarly, we can say for the realist that the reason that their principles predict (possibly surprising) results that can be subsequently verified in the weaker theory is that they aim more generally towards truth. In this case, the correctness (i.e. truth) of an axiom can simply be inferred abductively. [60]

The situation is slightly more complex for anti-realists. However, any anti-realist who hold that there is a project for justification of independent sentences requires

[56]See here [Lange, 2017], Ch. 7.

[57]Note here that while we do talk about explanatory proofs, in the context of Lange's account of explanation such proofs are explanatory in virtue of being related to certain explanatory properties in the world. Thus we are not depending here irreducibly on an account of explanatory proof.

[58]One could also consider [Steiner, 1978]'s account of mathematical explanation in terms of characterizing properties, where a property is characterizing if it is "unique to a given entity or structure within a family or domain of such entities or structures" ([Steiner, 1978], p. 151). Since there are several objections to Steiner's account (for a summary, see [Pease et al., 2018], p.4), and we a somewhat constrained by space, we will not consider Steiner's account.

[59]See [Barnes, 2018] for a recent survey.

[60]The idea that axioms should be inferred abductively is suggested by [Williamson,], though not in as much as we do here. For other arguments concerning prediction and abduction, see [Lipton, 2003].

that there is some notion of 'correctness' in foundations, even if it is not truth.[61] For example, we might think of predicting principles for a fictionalist as aiming at *truth within the set-theoretic story*, and a good explanation for this being that they are good candidates for continuation of that story (much as in the Archer analogy).[62] Given then, any such notion of correctness, we can still incorporate axioms as likely correct in virtue of making correct predictions.

To summarise, we have now argued for the following claims:

(A) There is a sense in which set theory has a technically precise notion of prediction and verification.

(B) This notion of prediction and verification can be linked to explanation, supporting our account of the justification of new axioms as an explanatory enterprise.

In the rest of the paper we accomplish the following three tasks:

(1.) We'll explain how our account resolves the problems of tractability and demarcation.

(2.) We'll deal with some natural objections.

(3.) We'll propose some open questions.

3.3 Tractability and Demarcation: Redux

We are thus in a position in which we regard justification in set theory as essentially a matter of assessing explanatory claims given some mathematical data. Before moving on to an analysis of objections, it will be helpful to pause and discuss how our problems of tractability and demarcation are resolved.

We have just finished explaining our response to the problem of tractability. Whether or not a principle figures in a good explanatory story may be difficult to assess, but there are *precise* criteria that we can point to that increase our confidence that a principle figures in mathematical explanations. While we have canvassed several options, there may well be (indeed we expect there to be) many other precise criteria for when we should regard putative axioms linked to explanations. We leave a full analysis to other work, but some possibilities have already been suggested in the literature, without tying their epistemic reliability to explanation. For example, not considered here were convergence by a sufficiently strong theory ([Koellner, 2006]) and maximising interpretative strength ([Maddy, 2011], [Steel, 2014]). Each requires study and a story of why we think criteria should track explanation in mathematics. For instance, Maddy considers the maximising of interpretive strength to be justified by her conception of the foundational goals of set theory and the axiom MAXIMIZE. Since that literature is already well-studied, we do not provide detailed examination here. We do wish to note, however, that Maddy often comes close to advocating

[61] We acknowledge that for those who do not think that there is any notion of 'correctness' in set theory, such as the pure formalist (who holds that mathematics is simply a meaningless game played with symbols), the only notion of 'justification' is one based on making choices of formal expedience, and so set these views aside.

[62] Another possible candidate would be Cartwright's 'simulacrum' account of explanation, that is anti-realist in that she thinks that explanations need not even get the phenomena right (see Essay 8 of [Cartwright, 1983]). However, her account depends upon a clear partition of the world into the mathematical and physical, with mathematical simulacra roughly resembling physical concreta. In this way, it is not clearly applicable to the purely mathematical context, and so we set it aside. For the anti-realist, we should also discuss [Fraassen, 1980]'s suggestion that explanations are context-sensitive answers to why-questions, we will consider this in the section on objections below.

something *like* the conditions we advocate here (for example using her notion of restrictiveness), it is specifically her tying of justification to *mathematical depth* that we take issue with.

Indeed much of Maddy's work could be naturally utilised by the current project. For example, we might consider ideas such as Maddy's MAXIMIZE, which is also a precise criterion, as would notions of theoretical completeness.[63] Detailed examination of these technically precise features and how they relate to explanation would be required for a full answer to the tractability problem, but we hope to have convinced the reader that this is at least a promising line of inquiry where depth seems to be more problematic (we discuss this further in the objections).

The demarcation problem is, however, fully resolved. This is because there is no pressure, on our account, to regard justificatory force as entirely intrinsic or extrinsic, or that the two can be separated in any meaningful way. Rather, we take it that there are various hallmarks of a theory receiving some justification either by being explanatory or correct through explanation. It might be appropriate to call some of these features more 'intrinsic' or 'extrinsic', and we do not want criticise the use of these terms as rough heuristics, but simply to argue that the justificatory force is stemming from the explanatory role of the axioms rather than anything else.

4 Objections

In this section we briefly consider some objections to our account.

Against mathematical explanation. One rejoinder to our account is to argue that there is in fact no good account of mathematical explanation. This has been pressed by Mark Zelcer who summarises his arguments as follows:

> "My claim amounts to the following: philosophical accounts of explanation for mathematics will not satisfy desiderata established for explanation in other domains, like science. Among other things, an account of mathematical explanation would have to, but cannot, show (1) either that there is a solid history of mathematics as a discipline with explanatory concerns (or there is a good reason why these concerns went largely unnoticed), (2) there is a good account of predictions in mathematics (or that prediction is not important for science) that are symmetrical to explanations, (3) the methodological differences between mathematics and science—that mandates that in science and not mathematics we expect everything (beside perhaps certain fundamental facts) to fall under an explanatory schema—can be explained away, (4) reducing surprise (in an objective way) is not a desideratum in mathematics, and (5) that mathematical explanation, despite appearances, does play a significant role in mathematics." ([Zelcer, 2013], p. 3)

Considerations of space prevent a full and detailed rebuttal of Zelcer's arguments. However, we can make a few remarks here. Regarding (1), Zelcer argues that there are not good examples of mathematical explanation either in the historical or contemporary literature, and those that there are (such as in [Hafner and Mancosu, 2005])

[63] A theory $\mathbf{T}_1$ is *more theoretically complete* than another $\mathbf{T}_2$ iff there is a sentence implied by $\mathbf{T}_1$ that is not implied by $\mathbf{T}_2$ and not vice versa.

tend to be rather "exotic". Whether or not one finds the examples "exotic" is something of a matter of taste, but in any case there are studies that suggest that mathematicians as a matter of empirical fact *do* use explanatory terms in their reasoning. [Pease et al., 2018] found that, in a study of the Mini-Polymath projects, mathematicians do regularly use language normally associated with explanation (e.g. "expla*", "underst*", "because", "as"). So, even if it is at the 'back' of mathematics[64], explanation seems to be an important part of mathematical discourse. (2) we take ourselves to already have responded to with our account of prediction. In favour of (3), Zelcer argues that there are some portions of mathematics that do not require explanations, whereas in science even so called 'brute' fact do have explanations. We have argued, however, that foundational axioms can be viewed as explanations of the mathematical data *as a whole*, and so for us every mathematical fact is an explanandum in virtue of being part of this data set. For (4) Zelcer argues that surprise is a result of realising that out of many possible worlds, the actual world is the way it is (or narrow down the space of possible worlds to a smaller set of more probable ones). Since the truths of mathematics are necessary, no such surprise is possible. However, aside from the data point that surprise *just is* a natural part of mathematical practice (the history of mathematics is littered with surprises, but the Martin quotation above is fairly representative), the appropriate notion of possibility for mathematical surprise is epistemic, not metaphysical, possibility. Thus we can perfectly well have Zelcer's notion of surprise in mathematics, if only with merely epistemically possible worlds that tolerate metaphysical impossibilities. Regarding (5) Zelcer argues that mathematical explanations are not employed, and thus considerations of parsimony dictate that we expunge them. Again, we hold ourselves to have flat out argued against this, and in any case the results of [Pease et al., 2018] tell in our favour.

Justification can be wrong. Another objection is the following: Under our conception, justification does not aim at *truth*. Rather, since it relates to what we view as explanatory, what we are justified in asserting might depart radically from the truth.

Before we respond, we should remark that the objection presupposes a very strong version of set-theoretic realism on which all (or at least many/most) set-theoretic sentences are either true or false. On the contrary, we see it as an advantage of our position that it is compatible with many different conceptions of the nature of set-theoretic discourse. We might, for example, have concepts of set that agree on **ZFC** with large cardinal axioms added, but disagree on CH (something like this idea is advocated in [Steel, 2014]). We see it as an advantage of our view that it can be integrated into many different positions, and can serve as a point of common ground between dissenting parties.

Our response to this general question is just to acknowledge that what is regarded as best justified on our framework may be false for a very strong realist. What is regarded as explanatory or inferred abductively may turn out to be false. But this is simply an epistemic fact of life, and is not a problem limited to mathematics. We are open, in principle, to our justifications failing. But failure does not undermine the theoretical value of the guiding principles that lead our research and justification.

Second, we should emphasise that justification is a fundamentally *dynamic* process. The account depends on a basis of mathematical data that requires systematisation. In this way, we are not providing a static account of justification (on which we

[64]See [Hersh, 1991] for an explanation of the difference between the 'front' and the 'back' of mathematics. Roughly speaking, the 'front' is the part of mathematics that is presented to the public (e.g. in journals, conference presentations etc.) whereas the 'back' is reserved for professionals (e.g. when collaborating on a proof using a blackboard, or on MathOverflow or MathStackexchange).

take ourselves to be supplying an account of how mathematical statements *in general* are justified), but rather how, given some accepted mathematical facts at a point in time, we can come to justify new principles.[65] Justification should thus be understood as a stepwise process of increasing credences, rather than an all-or-nothing matter, fixed eternally.

Is what we have suggested extrinsic? A second line of objection would be to argue that the kinds of criteria we appeal to are actually simply extrinsic. Thus we would not have actually shown a problem with the intrinsic/extrinsic distinction, but have rather advocated a new account of extrinsic justification. To support this, one might point to the fact that one of our examples (the Cone Lemma) is taken up by Maddy[66] as evidence of the priority of extrinsic justifications. Clearly the notion of prediction and verification we have appealed to is tightly linked to the notion of consequence, and hence should be viewed as extrinsic. This is not a criticism of our view per se, but rather a dialectic point about how it sits in the wider context of our criticisms and rebuttals.

The point we wish to emphasise is that while prediction of a datum and subsequent verification is conducted using a notion of consequence, the overall package of prediction and verification involves more. In particular, verification is only possible once a set of *auxiliary assumptions* has already been fixed (in the case of the Cone Lemma and braids **ZFC**). But holding these auxiliary assumptions fixed occurs within a wider justificatory framework, and parts of that involve what we might call 'intrinsic' considerations.

For example, it is partly *because* of the way that **ZFC** interacts with a wider 'intuitive' picture that we take it to be a good theory for assessing verification and fixing auxiliary assumptions, since this intuitive picture increases the confidence that explanation is being conferred. We do not wish to bar the set theorist or philosopher from the use of the terms 'intrinsic' and 'extrinsic' justification, just to insist that the fundamental feature of the world they latch on to is explanation, and there are no solely 'intrinsic' or 'extrinsic' justifications. For example, the existence of a roughly 'intuitive' picture such as the iterative conception, while normally regarded as an 'intrinsic' justification, has also facilitated a particularly 'fruitful' way of thinking about set theory, and so is also 'extrinsic' in some sense. Our point here is just that the existence of an intuitive picture should increase confidence that explanation is being provided.

Fixing the axioms of **ZFC** as auxiliary assumptions is part of the examples of prediction and confirmation we have provided, since it provides the metatheory for the study of braids. However, many of these are naturally linked to more 'intuitive' considerations. For example, concerning the Axiom of Foundation Potter remarks that:

> "Because the axiom of foundation did not have mathematical consequences, mathematicians showed no inclination to adopt it: interest in it was limited to specialists concerned with its metatheoretic consequences.
>
> Matters began to change only when Gödel ([Gödel, 1947], p. 519) presented the grounded collections not merely, as Mirimanoff had done, as a sub-universe of the universe of collections but rather as an independently motivated hierarchy which, as he pointed out, 'has never led to

[65] We thank [name removed for blind review] for pointing out this feature of our account, and subsequent discussion.

[66] See [Maddy, 2011], p. 127.

> any antinomy whatsoever'. Since the 1960s the assumption that every collection is grounded has been adopted enthusiastically by set theorists, and the idea that the only coherent conception is the iterative one has become widespread." ([Potter, 2004], p. 52)

However, of course the Axiom of Foundation has 'extrinsic' support in terms of the systematisation into the iterative picture it provides. Our point is not that the Axiom of Foundation is either clearly intrinsically or extrinsically justified, but rather that it is not clearly either, and that this axiom is needed to fix the background in which we conduct our predictions and verifications. It therefore can't be argued that the notion of prediction and verification we have explored is solely intrinsic or extrinsic.

No best explanation. The next objection is that perhaps there is no best explanation, or that explanatory considerations do not tell firmly in favour of competing axiom systems. Would we not then be stuck in a non-epistemically-useful deadlock?

Our response is that while we may indeed end up in a deadlock this would not be epistemically useless. First, given the emphasis on explanation, we may have already ruled out several putative axioms, and this would at least be somewhat useful. Second, at least we will have outlined in *precise* terms the considerations underlying each axiomatisation and why we take them to be justified. Third, an account of truly distinct competing explanatory frameworks might well provide a justification for genuine pluralism about set theory, and be epistemically useful in telling us why no resolution of certain independent questions is possible. In short, epistemic usefulness need not mandate the provision of definitive yes/no answers, but does require a sufficiently tractable notion to explain why answers to questions are or are not possible.

This relates to a similar criticism from a different angle: Might mathematical explanation be context-sensitive? This question is naturally motivated by viewing explanations as answers to *why*-questions (as proposed by [Fraassen, 1980], taken up by [Lange, 2017], and empirically confirmed by [Pease et al., 2018]) since the why question can vary and hence so can the explanation. But then we would have the criticism that our view might result in the context-sensitivity of axioms.

Two responses are relevant here: First, in the context of the justification of set-theoretic foundations we have quite a restricted class of contexts, namely those in which we look at the entire mathematical data set and ask what axioms concerning sets might systematise that. In this sense, we have a kind of *holism* at play; we should not be cherry-picking individual data-points in looking for foundational axioms. These restrictions increase the likelihood that there will be agreement in explanations. Second, if there truly is disagreement in explanation between differing contexts, then this will provide an epistemically useful underpinning of why there is genuine pluralism in set theory; there are legitimately different mathematical contexts requiring different axiomatisations.

5 Conclusions

In this paper, we've argued for two main claims: First that the appeal to intrinsic and extrinsic justifications as fundamental and competing is problematic, and second that explanatory considerations are the fundamental driver in set-theoretic justifications. Whilst we take ourselves to have taken a first step in this direction, the piece is somewhat exploratory, and there are many open questions to be resolved.

One obvious and key issue is that the notion of explanation in mathematics needs a good deal of further work for a full account to be provided. We canvassed only a few possible constraints on good explanations in this essay, however there is space for an entire literature here. We therefore open with the following question:

Question 2. Can a complete list of *precise* justification-conferring features be compiled?

In particular, while we have focussed on prediction, it is presumably only one explanatory good among many. In discussing justification more generally, there has been much good work done by Maddy, Koellner, Martin, and others in providing a taxonomy and analysis of different kinds of justification, and we do not wish to discredit their work. From our perspective there is much to be done in explaining how the criteria they provide, many of which can be given precise characterisations (e.g. restrictiveness of theories[67], level of theoretical completeness[68], convergence[69]), can be integrated into our own explanatory account.

Further, while we have provided explanation of how our view responds to the demarcation and tractability problems, our response to the latter is only partial. We have shown how certain theories can receive more confirmation than others on the basis of prediction and verification. We have not, however, shown how we can (if at all) *choose between the well-confirmed theories*. To a degree this is to be expected, and reflects a usual problem at the cutting edge of any foundational research (it would be unfair to force physicists to choose between relativity theory and quantum theory because of their incompatability). However, some way of comparing the theories (possibly with a calculus of confirmation analogous to those used in the philosophy of science) is desirable. The following question is thus of interest:

Question 3. Is it possible to come up with a way of assigning different theories weights and comparing them satisfactorily?

One final point of contact is with notions of *grounding*. Often explanation is regarded as a species of this wider dependence relation, and there are several distinctions we have not examined here (e.g. the difference between ontic and explanatory grounding, and the different grounding axioms that might thereby be argued to apply). We therefore ask:

Question 4. Can we make progress on mathematical explanation and the justification of axioms on the basis of the study of grounding?[70]

In sum, we have seen that the distinction between intrinsic and extrinsic justification, when regarded as fundamental and indicative of a conflict in justificatory force, is beset by the difficult problems of demarcation and tractability. A better response is to regard justification as as linked to explanation, with possible precise hallmarks of explanation. We have proposed an initial step in this direction, but much more is still to be done.

References

[Aigner and Ziegler, 2009] Aigner, M. and Ziegler, G. M. (2009). *Proofs from THE BOOK*. Springer, 4 edition.

[67] See [Maddy, 1998].
[68] See [Koellner, 2010], §3.3.
[69] See [Koellner, 2010], §3,4.
[70] We thank [name removed for blind review] for suggestion and discussion of this question.

[Arrigoni and Friedman, 2013] Arrigoni, T. and Friedman, S.-D. (2013). The Hyperuniverse Program. *Bulletin of Symbolic Logic*, 19:77–96.

[Barnes, 2018] Barnes, E. C. (2018). Prediction versus accommodation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition.

[Boolos, 1971] Boolos, G. (1971). The Iterative Conception of Set. *The Journal of Philosophy*, 68(8):215–231.

[Boolos, 1989] Boolos, G. (1989). Iteration again. *Philosophical Topics*, 17(2):5–21.

[Boolos, 2000] Boolos, G. (2000). *Must We Believe in Set Theory?*, page 257–268. Cambridge University Press.

[Brendle, 2006] Brendle, J. (2006). Cardinal invariants of the continuum and combinatorics on uncountable cardinals. *Annals of Pure and Applied Logic*, 144(1):43 – 72.

[Cantor, 1883] Cantor, G. (1883). Foundations of a general theory of manifolds; a mathematico-philosophical investigation into the theory of the infinite. In Ewald, W., editor, *From Kant to Hilbert; A Source Book in the Foundations of Mathematics*, volume 2. Oxford University Press.

[Cantor, 1899] Cantor, G. (1899). Letter to Dedekind. In *[van Heijenoort, 1967]*, pages 113–117. Harvard University Press.

[Cartwright, 1980] Cartwright, N. (1980). The truth doesn't explain much. *American Philosophical Quarterly*, 17(2):159–163.

[Cartwright, 1983] Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

[Dales and Oliveri, 1998] Dales, H. and Oliveri, G., editors (1998). *Truth in Mathematics*. Clarendon Press.

[Dehornoy, 2000] Dehornoy, P. (2000). *Braids and Self-Distributivity*. Springer.

[Donellan et al., 2006] Donellan, M., Oswald, F., Baird, B., and Lucas, R. (2006). Th mini-IPIP scaled: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, 18:192–203.

[Drake, 1974] Drake, F. R. (1974). *Set Theory: An Introduction to Large Cardinals*. North Holland Publishing Co.

[Džamonja and Shelah, 1999] Džamonja, M. and Shelah, S. (1999). ♣ does not imply the existence of a suslin tree. *Israel Journal of Mathematics*, 113(1):163–204.

[Enderton, 1977] Enderton, H. (1977). *Elements of Set Theory*. Academic Press.

[Ewald, 1996] Ewald, W. B., editor (1996). *From Kant to Hilbert. A Source Book in the Foundations of Mathematics*, volume I. Oxford University Press.

[Feferman et al., 2000] Feferman, S., Friedman, H., Maddy, P., and Steel, J. (2000). Does mathematics need new axioms? *Bulletin of Symbolic Logic*, 6(4):401–446.

[Ferreirós, 1999] Ferreirós, J. (1999). *Labyrinth of Thought*. Birkhäuser.

[Forster, 1995] Forster, T. (1995). *Set Theory With a Universal Set: Exploring an Untyped Universe*. Clarendon Press.

[Fraassen, 1980] Fraassen, B. V. (1980). *The Scientific Image*. Oxford University Press.

[Gödel, 1938] Gödel, K. (1938). The consistency of the axiom of choice and of the generalized continuum-hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 24(12):556–557.

[Gödel, 1947] Gödel, K. (1947). What is Cantor's continuum problem? In *[Gödel, 1990]*, pages 176–187. Oxford University Press.

[Gödel, 1964] Gödel, K. (1964). What is Cantor's continuum problem? In *[Gödel, 1990]*, pages 254–270. Oxford University Press.

[Gödel, 1990] Gödel, K. (1990). *Collected Works, Volume II: Publications 1938-1974*. Oxford University Press. Edited by: Solomon Feferman (Editor-in-chief), John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, Jean van Heijenoort.

[Hacking, 1992] Hacking, I. (1992). The self-vindication of the laboratory sciences. In Pickering, A., editor, *Science as Practice and Culture*, pages 29–64. University of Chicago Press.

[Hafner and Mancosu, 2005] Hafner, J. and Mancosu, P. (2005). The varieties of mathematical explanation. In Mancosu, P., editor, *Visualization, Explanation and Reasoning Styles in Mathematics*, pages 215–250. Dordrecht: Springer.

[Hallett, 1984] Hallett, M. (1984). *Cantorian Set Theory and Limitation of Size*. Oxford University Press.

[Hauser, 2001] Hauser, K. (2001). Objectivity over Objects: a Case Study in Theory Formation. *Synthèse*, 128(3).

[Hersh, 1991] Hersh, R. (1991). Mathematics has a front and a back. *Synthese*, 88(2):127–133.

[Holmes, 1998] Holmes, R. (1998). *Elementary Set Theory with a Universal Set*. Bruylant-Academia.

[Holy et al., S] Holy, P., Lücke, P., and Njegomir, A. (S). Small embedding characterizations for large cardinals. Submitted, arXiv:1708.06103 [math.LO].

[Incurvati and Murzi, 2017] Incurvati, L. and Murzi, J. (2017). Maximally consistent sets of instances of naive comprehension. *Mind*, 126(502).

[Inglis and Aberdein, 2015] Inglis, M. and Aberdein, A. (2015). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica*, 23(1):87–109.

[Inglis and Aberdein, 2016] Inglis, M. and Aberdein, A. (2016). *Diversity in Proof Appraisal*, pages 163–179. Springer International Publishing, Cham.

[Jech, 2002] Jech, T. (2002). *Set Theory*. Springer.

[Kanamori, 2009] Kanamori, A. (2009). *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*. Springer, 2nd edition.

[Kitcher, 1981] Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, 48(4):507–531.

[Koellner, 2006] Koellner, P. (2006). On the question of absolute undecidability. *Philosophia Mathematica*, 14:153–188.

[Koellner, 2009] Koellner, P. (2009). On reflection principles. *Annals of Pure and Applied Logic*, 157:206–219.

[Koellner, 2010] Koellner, P. (2010). On the question of absolute undecidability. In Feferman, S., Parsons, C., and Simpson, S. G., editors, *Kurt Gödel: Essays for his Centennial*, pages 189–222. Association for Symbolic Logic.

[Koellner, 2014] Koellner, P. (2014). Large cardinals and determinacy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2014 edition.

[Koellner and Woodin, 2010] Koellner, P. and Woodin, H. (2010). Large cardinals from determinacy. In *Handbook of Set Theory*, pages 1951–2119. Springer.

[Lange, 2017] Lange, M. (2017). *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford University Press USA.

[Leonelli, 2015] Leonelli, S. (2015). What counts as scientific data? a relational framework. *Philosophy of Science*, 82(5):810–821. PMID: 26869734.

[Lipton, 2003] Lipton, P. (2003). *Inference to the Best Explanation*. Routledge.

[Maddy, 1988a] Maddy, P. (1988a). Believing the axioms. I. *The Journal of Symbolic Logic*, 53(2):481–511.

[Maddy, 1988b] Maddy, P. (1988b). Believing the axioms II. *The Journal of Symbolic Logic*, 53(3):736–764.

[Maddy, 1990] Maddy, P. (1990). *Realism in Mathematics*. Clarendon Press.

[Maddy, 1997] Maddy, P. (1997). *Naturalism in Mathematics*. Oxford University Press.

[Maddy, 1998] Maddy, P. (1998). $v = l$ and maximize. In Makowsky, J. A. and Ravve, E. V., editors, *Proceedings of the Annual European Summer Meeting of the Association of Symbolic Logic*, pages 134–152. Springer.

[Maddy, 2007] Maddy, P. (2007). *Second Philosophy*. Oxford University Press.

[Maddy, 2011] Maddy, P. (2011). *Defending the Axioms*. Oxford University Press.

[Magidor, U] Magidor, M. (U). Some set theories are more equal. Unpublished.

[Mancosu, 2008] Mancosu, P. (2008). *The Philosophy of Mathematical Practice*. Oxford University Press.

[Martin, 1998] Martin, D. (1998). Mathematical evidence. In *[Dales and Oliveri, 1998]*, pages 215–231. Clarendon Press.

[Martin, 1976] Martin, D. A. (1976). Hilbert's first problem: the continuum hypothesis. *Proceedings of Symposia in Pure Mathematics*, 28:81–92.

[Martin, 1977] Martin, D. A. (1977). Descriptive set theory: Projective sets. In Barwise, J., editor, *Handbook of Mathematical Logic*, pages 783–815. North Holland Publishing Co.

[Mehlberg, 1960] Mehlberg, H. (1960). The present situation in the philosophy of mathematics. *Synthese*, 12(4):380–414.

[Mirimanoff, 1917] Mirimanoff, D. (1917). Les antinomies de russell et de buraliforti et le probleme fondamental de la theorie des ensembles. *L'Enseignement Mathématique*, 19:37–52.

[Moschovakis, 1980] Moschovakis, Y. (1980). *Descriptive Set Theory*. North Holland Publishing Co.

[Parsons, 1983] Parsons, C. (1983). Sets and modality. In *Mathematics in Philosophy: Selected Essays*. Cornell University Press.

[Pease et al., 2018] Pease, A., Aberdein, A., and Martin, U. (2018). Explanation in mathematical conversations: An empirical investigation. *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*.

[Potter, 2004] Potter, M. (2004). *Set Theory and its Philosophy: A Critical Introduction*. Oxford University Press.

[Rheinberger, 2011] Rheinberger, H.-J. (2011). Infra-experimentality: From traces to data, from data to patterning facts. *History of Science*, 49(3):337–348.

[Russell, 1907] Russell, B. (1907). The regressive method of discovering the premises of mathematics. In *[Russell, 1973]*, pages 272–283. George Allen & Unwin Ltd.

[Russell, 1973] Russell, B. (1973). *Essays in Analysis*. George Allen & Unwin Ltd. Edited by Douglas Lackie.

[Salmon, 1984] Salmon, W. C. (1984). Scientific explanation: Three basic conceptions. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984:293–305.

[Scott, 1961] Scott, D. (1961). Measurable cardinals and constructible sets. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.*, 9:521–524.

[Shoenfield, 1967] Shoenfield, J. (1967). *Mathematical Logic*. Addison-Wesley Publishing Co.

[Steel, 2014] Steel, J. (2014). Gödel's program. In Kennedy, J., editor, *Interpreting Gödel*. Cambridge University Press.

[Steiner, 1978] Steiner, M. (1978). Mathematical explanation. *Philosophical Studies*, 34(2):135–151.

[Tait, 2001] Tait, W. (2001). Gödel's unpublished papers on the foundations of mathematics. *Philosophia Mathematica*, 9:87–126.

[Tait, 2005] Tait, W. (2005). *The Provenance of Pure Reason: Essays in the Philosophy of Mathematics and its History*. Oxford University Press.

[Tiles, 1989] Tiles, M. (1989). *The Philosophy of Set Theory: An Historical Introduction to Cantor's Paradise*. Dover Publications.

[van Heijenoort, 1967] van Heijenoort, J., editor (1967). *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Harvard University Press.

[Venturi, 2018] Venturi, G. (2018). *Logic and its Philosophy: contemporary trends in Latin America and Spain*, chapter On the naturalness of new axioms in set theory, pages 117–142. Open Court Publishing, Chicago.

[Wang, 1974] Wang, H. (1974). *From Mathematics to Philosophy*. Routledge & Kegan Paul, London.

[Wang, 1996] Wang, H. (1996). *A Logical Journey*. MIT Press, Cambridge (MA).

[White, 2003] White, R. (2003). The epistemic advantage of prediction over accommodation. *Mind*, 112(448):653–683.

[Williamson,] Williamson, T. Abductive philosophy. *The Philosophical Forum*, 47(3-4):263–280.

[Woodin, 2017] Woodin, W. H. (2017). In search of ultimate-l the 19th midrasha mathematicae lectures. *The Bulletin of Symbolic Logic*, 23(1):1–109.

[Zelcer, 2013] Zelcer, M. (2013). Against mathematical explanation. *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie*, 44(1):173–192.

[Zermelo, 1908] Zermelo, E. (1908). A new proof of the possibility of a well-ordering. In van Heijnoort, J., editor, *vanHeijnoort1967a*, pages 183–198. Harvard.

[Zermelo, 1930] Zermelo, E. (1930). On boundary numbers and domains of sets. In *[Ewald, 1996]*, volume 2, pages 1208–1233. Oxford University Press.