Synthese (2021) 199:1245–1267
https://doi.org/10.1007/s11229-020-02784-z

Set-theoretic justification and the theoretical virtues

John Heron[1]

Received: 30 September 2019 / Accepted: 7 July 2020 / Published online: 14 July 2020

Abstract
Recent discussions of how axioms are extrinsically justified have appealed to abductive considerations: on such accounts, axioms are adopted on the basis that they constitute the best explanation of some mathematical data, or phenomena. In the first part of this paper, I set out a potential problem caused by the appeal made to the notion of mathematical explanation and suggest that it can be remedied once it is noted that all the justificatory work is done by appeal to the theoretical virtues. In the second part of the paper, I appeal to the theoretical virtues account of axiom justification to provide an argument that judgements of theoretical virtuousness, and therefore of extrinsic justification, are subjective in a substantive sense. This tells against a recent claim by Penelope Maddy that such justification is "wholly objective".

Keywords Set theory · Gödel · Russell · Axiom choice · Penelope Maddy · Justification · Theoretical virtues · Abductive inference

1 Introduction

Many statements of set theory are impossible to resolve unless the ZFC axioms are extended. Gödel's program involves resolving these open questions by extending standard set theory via the inclusion of well-justified axioms (Gödel 1947 [1990]). This program suggests natural questions. How are candidate axioms selected? What makes an axiom justified? These are questions that are important not only for those of us interested in pursuing Gödel's program, but more generally for those interested in set-theoretic justification.

A traditional distinction is drawn between intrinsic and extrinsic justifications for admitting an axiom. An axiom is intrinsically justified if it is contained within the concept of set, or is self-evident. An axiom is extrinsically justified if it is fruitful, or successful. This paper concerns the latter kind of justification. A recently dis-

✉ John Heron
heron.h.john@gmail.com

[1] London, UK

cussed account of extrinsic justification—drawing on Russell's regressive method and Gödel's remarks about the methodological continuity between science and mathematics—spells the notion out in terms of abductive inference, or explanation. On this sort of account, an axiom is extrinsically justified if it figures in the best mathematical explanation of some mathematical data. In this paper I argue that this account's focus on mathematical explanation opens it up to worries and I suggest an alternate, but related, view that fixes this vulnerability. I then put the resulting account to work in assessing another open question about extrinsic justification: whether or not judgements of extrinsic justification are wholly objective, or if there is an important sense in which they are subjective, or subject-sensitive. Using the account of extrinsic justification defended earlier in the paper, I argue that there is an important sense in which comparative claims of extrinsic justification are subject-sensitive. This speaks against a recent claim made by Penelope Maddy: namely, that extrinsic justification (or, as she has it, "mathematical depth") is wholly an objective matter.

A road map of the paper: in section two I briefly recount the distinction between extrinsic and intrinsic justification in mathematics. In section three I discuss a way of understanding extrinsic justification which has roots in the work of Russell and Gödel, and which has been recently discussed in the literature, an account that foregrounds the notion of explanation. I raise a problem for the account and suggest that it can be remedied once it is recognised that all the justificatory work is being done by the (implicit or explicit) appeal to the theoretical virtues. In section four I consider the ramifications of the theoretical virtue account of extrinsic justification for debates about the status of such justification: in particular, as to whether such judgements are objective or subjective (in a sense to be made precise). After rejecting an argument given by Maddy for objectivism, I appeal to the theoretical virtues account to provide an argument for subjectivism. In section five, before concluding, I illustrate the wide-ranging consequences of paying attention to the objectivism/subjectivism question by providing an argument from the truth of subjectivism to the conclusion that there are absolutely undecidable mathematical statements.

2 Intrinsic and extrinsic justification

The distinction between intrinsic and extrinsic justification in mathematics is a familiar one, so this section will be brief. I'm concerned here with the justification for set-theoretic axioms, rather than justification for our beliefs concerning more humdrum mathematical claims.[1]

In his 1947, Gödel famously distinguishes the two forms of justification. If an axiom is intrinsically justified, then it is self-evident or is part of the concept *set*:

> These axioms show clearly, not only that the axiomatic system of set theory as used today is incomplete, but also that it can be supplemented without arbitrari-

[1] The justification we provide for non-axiomatic mathematical beliefs is likely to be motley: we take some mathematical beliefs to be justified on the basis that we (or, more likely, someone else) possess a proof or proof-sketch, some may be non-deductively justified whilst others may be non-inferentially justified.

> ness by new axioms which only unfold the content of the concept of set. (Gödel 1947: pp. 260–261)

The idea, then, is that some claims about sets (in particular, some set-theoretic axioms) are already part of the notion of *set* that we operate with. There is a sense in which we are not making new claims about what sets are like, but merely making explicit what is already implicit in the concept employed in our practice. For current purposes it can be granted that there are some facts of the matter about whether or not (at least some) axioms are part of the concept *set* and that some of these facts are transparent to us. Earlier, Frege appealed to the notion of self-evidence to explicate this kind of justification (Frege 1884; see Shapiro 2009 for a discussion of Fregean notions of self-evidence). Maddy provides an account of intrinsic justification that includes both the notion of self-evidence and that of concept-containment:

> [T]he 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. […] 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).

At first pass, some of the standard axioms may strike us as self-evident. Once we grasp what sort of object a set is, Extensionality (asserting that if X and Y have the same elements then X = Y) or Pairing (asserting that for any *a* and *b* there exists a set {*a, b*} that contains exactly *a* and b) strike us as the kind of thing that couldn't possibly be false of sets. However, there will be disagreement about whether or not an axiom is intrinsically justified.[2] In addition, for some axioms considered plausible candidates, it is unclear whether or not we have stable judgements as to their self-evidence or to their containment in the concept *set*.

Both Maddy and Gödel note that many justifications given in set-theoretic practice for adopting an axiom do not appear to be intrinsic. Sometimes set-theorists take themselves to be justified in accepting a new candidate axiom that fails to be self-evident or fails to suggest itself as already contained within the concept *set*. Maddy claims that in such cases:

> [T]he axioms are evaluated in terms of their consequences, or more broadly, in terms of the type of theory they produce. (Maddy 1997: p. 37)

Gödel makes similar remarks when discussing extrinsic justification:

> [E]ven disregarding the intrinsic necessity of some new axiom, and even in case it has no intrinsic necessity at all, a probable decision about its truth is possible also in another way, namely, inductively by studying its "success". (Gödel 1947 [1990]: p. 261)

Left at this, the notion of extrinsic justification is unclear: what sort of consequence is relevant and why does some axiom having these consequences justify accepting it?

[2] For discussion of potential examples of such disagreements, see Barton, Ternullo and Venturi *ms, pp.* 6–8; Clarke-Doane (2013, pp. 473–474).

What exactly is it for an axiom to be, as Gödel puts it, successful? It is the job of the next section to discuss answers to this question.

Choice is a paradigm example of an axiom that is best justified extrinsically, rather than by marshalling intrinsic considerations (for more detail see Maddy's discussion in Maddy 2011: pp. 32–36, 45–48; Imocrante 2015: pp. 230–231). Indeed, on a pre-theoretic notion of fruitfulness, Choice appears very fruitful. It has applications not only in set theory but also in algebra, topology, analysis (Imocrante 2015: p. 230), and in geometry (Maddy 2011: pp. 34–35). It is this kind of fruitfulness that favours accepting Choice.

In addition to playing a role in admitting new axioms, Maddy has suggested that judgements about extrinsic justification also play an important role in the introduction of new mathematical concepts, like the introduction of the concept of a group:

> What guides our concept formation, beyond the logical requirement of consistency, is the way some logically possible concepts track deep mathematical strains that the others miss. (*ibid*: 79)

How do intrinsic and extrinsic justifications relate? Maddy goes so far as to say that:

> [I]ntrinsic 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. (*ibid*: 136)

If either the weaker claim (that intrinsic justification must be complemented by extrinsic justification) or Maddy's stronger claim (that extrinsic justification ought to take some form of priority over intrinsic justification) are true, we would do well to try and answer the pressing philosophical questions about extrinsic justification.

3 Explanation and the theoretical virtues

This section is about the best way to understand extrinsic justification. I trace the development of an account of extrinsic justification that understands it as an essentially abductive affair: that, in one way or another, axioms are justified on the basis that they best explain, organise and systematize non-axiomatic mathematical claims. In the following section I raise some problems for this way of understanding extrinsic justification. I then suggest a related account that avoids these problems, one that places the theoretical virtues, rather than explanation, at the forefront.

3.1 The explanatory account of justification

In 1907, Russell defended a regressive method for mathematics, putting it to use to argue in favour of the much-maligned Axiom of Reducibility. On this account of extrinsic justification, the proper method for making decisions about the axioms has much in common with the received picture of scientific practice. On this thumbnail sketch of scientific practice, we begin with some facts known via observation and form

beliefs about fundamental scientific principles by selecting those which best account for, or systematize, those facts known via observation. Much the same, says Russell, goes for the foundational principles of mathematics. This reverses the Euclidean picture on which whatever justification we have for the non-fundamental mathematical beliefs stems from our justification about the axioms. As Russell says:

> [W]e tend to believe the premises because we can see that their consequences are true, instead of believing the consequences because we know the premises to be true. But the inferring of premises from consequences is the essence of induction; thus the method in investigating the principles of mathematics is really an inductive method, and is substantially the same as the method of discovering general laws in any other science. (Russell 1907: pp. 273–274)

Consider, also, Gödel's reflections on the regressive method. He says that Russell:

> [C]ompares the axioms of logic and mathematics with the laws of nature and logical evidence with sense perception, so that the axioms need not necessarily be evident in themselves, but rather their justification lies (exactly as in physics) in the fact that they make it possible for these "sense perceptions" to be deduced; which of course would not exclude that they also have a kind of intrinsic plausibility similar to that in physics (Gödel 1944 [1990]: p. 121)

Gödel's account foregrounds methodological similarity between the foundations of set theory and theory choice in science.[3] A final historical precedent: although Zermelo is largely discussed in relation to his notion of self-evidence as unconscious use (e.g., Shapiro 2009), he also seemed to hold that extrinsic justification of this sort plays an important role:

> Actually principles must be judged from the point of view of science, and not science from the point of view of principles fixed once and for all (Zermelo 1967: p. 189)

In addition to having substantive historical pedigree, such an account has received recent interest in the literature. In the process of making the case that theory choice in philosophy does (and should more systematically) proceed abductively, Williamson has recently presented an account of axiom justification that appears to have much in common with Russell's regressive method (Williamson 2016). For Williamson, the sciences, philosophy and (the foundations of) mathematics have significant methodological continuity in that they all substantially rely on abduction:

> There is no methodological firewall between philosophy and real life experiments. However, that does not mean that once philosophy becomes a more systematically abductive inquiry, its current methodological differences with the natural sciences will fade away altogether. For natural science is not the only form of systematic inquiry in which abduction plays a significant role. It also does so in at least one highly successful form of "armchair" inquiry: mathematics. (Williamson 2016: p. 269)

[3] Of course, it's unlikely that in scientific practice, observation statements are strictly *deduced* from the theories accepted.

Russell, Gödel and Williamson, then, all recognise methodological continuity between science and the foundations of set theory—in particular, between choosing between (empirically equivalent) theories in the former and between axioms in the latter. One currently open question is whether or not we should understand this methodological similarity to involve the notion of *explanation*. What is important for Russell and Gödel seems to be the fact that the non-axiomatic mathematical claims can be *derived* from the more general principles, much as individual observations are to be subsumed under general physical laws. Williamson, however, is explicit that the common methodology of science, philosophy and (the foundations of) mathematics is the use of the abductive method. He is similarly explicit that he is using 'abductive' in the sense where it is "approximately equivalent to "inference to the best explanation", when "explanation" is understood to cover non-causal as well as causal explanations" (Williamson 2016: p. 263). On this sort of picture, scientific theories, philosophical theories and set-theoretic axioms are selected on the basis of offering the best explanations of the relevant data.[4] For Williamson, much as theory choice in science proceeds by inference to the best scientific explanation, theory choice in the foundations of mathematics proceeds by inference to the best mathematical explanation (and theory choice in philosophy proceeds by inference to the best philosophical explanation).

In the next section I will argue that an account of axiom justification that builds in the notion of explanation is subject to a number of worries. In response, I will offer a related account that captures the methodological continuity that Russell, Gödel and Williamson note, without being subject to the worries discussed. It is this account of axiom justification that will guide the remaining discussion.

3.2 Mathematical explanation and the scarcity problem

One initial worry with the kind of explanationist story presented by Williamson concerns descriptive adequacy, at least insofar as the picture is offered as spelling out the sort of account endorsed by Russell and Gödel. As noted above, on Russell and Gödel's picture axioms are selected on the basis that from them we can derive the non-axiomatic claims (the data points, as it were) that are, in some sense, epistemically prior. Someone who is inclined to see axiom choice as an abductive, or explanatory, affair can of course simply label this reverse-Euclidean relationship as an explanatory one. However, this does not show what this relationship of derivability has in common with the other kinds of relations between facts that we more readily label as explanatory (such as causal relations).[5] In response, the defender of the kind of explanationist picture offered by Williamson can always accept that the account is not the account offered by Gödel and Russell but *is* the account we should accept. However, there are some non-exegetical worries lurking.

[4] This sort of explanationist account (offered as an account not of extrinsic justification, but axiom justification *simpliciter*) is spelled out in more detail in work by Barton, Ternullo and Venturi (Barton, Ternullo & Venturi *ms*). I take the considerations raised against such accounts in the next section to tell against any account of axiom justification that places explanation at its heart.

[5] Thanks to an anonymous reviewer for pressing me to be clearer about the degree to which the explanatory account of axiom justification captures Russell and Gödel's regressive method.

What might be wrong with introducing the notion of explanation into one's account of axiom justification? One might harbour worries about the notion of explanation at play in the first place. There has been a flurry of work on explanation within mathematics (e.g., Baker 2009; Mancosu 2008; Lange 2009; Lange 2017; Dougherty 2018): however, unsurprisingly, no consensus has been reached. Perhaps more worryingly for the explanatory account, the vast majority of this work concerns providing an account of the difference between explanatory and non-explanatory *proofs*. The (now familiar) thought here is that some proofs of a theorem *explain why* the theorem is true whilst others merely demonstrate *that* it is, and the search is on for the characteristic(s) of proofs that render them explanatory. This, however, is not the kind of mathematical explanation that must be being appealed to by the explanatory account. On this account, an axiom is justified if it offers the best explanation of the relevant mathematical data: the explanatory relation, as it were, must hold between the candidate axiom (or collection of candidate axioms) and the relevant data. There is nothing to say that this relation is one of *proof* however. This is especially clear in the case of a new mathematical concept (like a group) being introduced: the concept of a group isn't the right kind of thing to stand in a proof relation. Let's distinguish, then, between explanatory proofs and axiom-data mathematical explanations. It is axiom-data mathematical explanations that are relevant to the appeal to "the best explanation" in the statement of the explanatory account above. Just as mathematical explanations are often appealed to in order to justify a bifurcation of explanation into its causal and non-causal varieties, the kind of explanations required by the explanatory account can be used to justify a bifurcation of the notion of an explanation in mathematics into those that are constituted by proofs and those that are not.[6]

Just as there are some that have doubted the existence of explanatory proofs (Zelcer 2013) one might doubt the existence of the form of mathematical explanation that holds between the axioms and the relevant mathematical data. Certainly, anyone with the sort of monist picture of explanation that holds that all explanations of particular facts are either causal or (more weakly) that all explanations of particular facts involve providing information about counterfactual dependence (Pincock 2018; Reutlinger 2016) will have trouble fitting in these kind of axiom-data mathematical explanations. On this sort of picture—increasingly popular in the literature on non-causal explanation—information is explanatory in virtue of providing the answers to *what-if-things-had-been-different* questions. However, mathematical statements are standardly understood as being necessarily true. Accordingly, Choice cannot feature in the best explanation of some mathematical data on the basis that it tells us how things would have been different, were the relevant mathematical facts different: because these facts are metaphysically necessary.[7]

Nevertheless, even if one grants the proponents of the explanatory account that there *are* axiom-data mathematical explanations, a deeper problem lurks. This is a problem that is familiar from the literature on the use of inference to the best explanation (IBE) in the scientific context: that, because we lack an account of the kind of explanation

[6] There is, of course, a prior bifurcation into explanations *within* mathematics and mathematical explanations of non-mathematical facts.

[7] However, Baron et al. (2017) do some of the groundwork for an account of mathematical explanation that appeals to counterfactuals, leaving the door open to a genuinely monist account of explanation.

we're interested in, IBE is in some sense indeterminate. In the context of scientific IBE, the issue presents itself as a problem of plenitude (Cabrera 2020; see, also, Salmon 2001). Given that there are many rival accounts of scientific explanation (such as those that foreground causal or counterfactual information, those that involve the notion of unification, those that involve the notion of deduction and subsuming under a general law, and so on) and given that these accounts will provide inconsistent verdicts about what the best explanation of a given explanandum is, IBE is unable to provide a verdict unless we have good reason to think that we've identified the correct account of scientific explanation. In the case of mathematical IBE, and axiom-data explanations, the same fundamental worry presents itself as a kind of *scarcity* problem. The problem isn't that we have *lots* of accounts of axiom-data explanation, and that these given inconsistent verdicts about what the best explanation of some data is, but rather that we currently have *no such* accounts. But the main worry remains: as Cabrera says regarding the plenitude problem, "if it is unclear what it means to explain some phenomena, then what IBE consists in will likewise be rendered indeterminate" (Cabrera 2020: p. 729). We can be unclear because we have *many* accounts, but we can also be unclear because we lack an account.

Here are the two most obvious responses to this worry. The first is to try and settle the question regarding the nature of axiom-data mathematical explanations and 'plug' this account into the model of IBE endorsed by the explanatory account of axiom justification. If we know what it is for a (collection of) axiom(s) to explain some mathematical data, then we can use this account to select axioms via identifying the best explanation of the data. However, we seem far from settling this matter. For the reasons canvassed above, the standard accounts of scientific explanations seem ill-suited to understanding axiom-data explanations and, similarly, it is not obvious that extant accounts of explanatory proofs can be extended to accommodate axiom-data explanations. The second is to abandon wholesale the kind of picture that the explanationist account gives us. However, this would be to throw the baby out with the bathwater. Indeed, there is an account that has much in common with the explanatory account but does not suffer from its flaws. In the next section I set out and further motivate the acceptance of this account.

3.3 The theoretical virtues conception of axiom justification

As we saw in the previous section, it has been suggested that Gödel and Russell's account of extrinsic justification should be read as having a crucial role for axiom-data explanations. A development of Gödel and Russell's account in this direction faces the twin worries discussed above: it is unclear if reading the notion of axiom-data explanations into the regressive method is exegetically sound and, more importantly, such an account suffers from a crucial lacuna, absent an account of axiom-data explanation.

Fortunately, highlighting the similarities between axiom choice in set theory and theory choice in science also highlights potential solutions to problems in the philosophy of mathematics that can be borrowed from solutions to problems in the philosophy of science. In this section, I'll suggest that carrying out such a borrowing leads to an account of axiom justification that has commonalities with the explanationist story,

but is both faithful to the remarks made by Russell and Gödel and avoids the scarcity problem.

First, note that the explanatory account appeals to the notion of there being some *best* explanation of the relevant mathematical data. Standardly, abductive inference uses the theoretical (or 'explanatory') virtues to choose between rival potential explanations (once clearly unsatisfactory potential explanations are ruled out). In the scientific context, for example, empirically equivalent theories are chosen between by appealing to properties of the theories other than their consistency with the evidence. The standard theoretical virtues include things like simplicity, being non-ad-hoc, unifying power, coherence, fertility, and so on. Nothing in the current discussion turns on our having an exhaustive collection of the relevant theoretical virtues, nor having settled whether the set of theoretical virtues is common across scientific, philosophical, and mathematical contexts.[8]

The suggestion here then, is a simple one: that the proposed account of extrinsic justification ought to jettison the notion of *explanation* and simply make do with the theoretical virtues. As Cabrera says:

> It is not the case that a hypothesis needs to satisfy the conditions of some one particular model of explanation in order to be explanatorily virtuous. Indeed, what is common in all of the many instances of explanatory inference cited across a wide array of different domains is not some model of explanation, but rather a set of widely applicable explanatory virtues. In my view, this suggests that what does all the justificatory work in any application of IBE is just that *H* does well with respect to the explanatory [theoretical] virtues relative to some evidence *E*. Hence, whether a hypothesis satisfies the conditions of any particular model of explanation is, epistemically, beside the point. (Cabrera 2020: p. 743).

The thought, then, is just that the focus on *explanation* is a red herring. Suppose that a set-theorist pursuing Gödel's program realises that a particular candidate axiom does particularly well with respect to the theoretical virtues. That is, the axiom is non-ad-hoc, powerful, has high predictive power (in the sense spelled out above), and so on. It is unclear that if our set-theorist learns that (in addition to scoring highly when it comes to the theoretical virtues) the axiom would *explain* the relevant set-theoretical data, that this would be taken to have any additional justificatory power over and above that assigned to the axiom in virtue of it being theoretically virtuous.

A consequence of adopting the suggestion made in this section is that it provides someone attracted to the kind of explanatory account suggested by Williamson (via Russell and Gödel) a solution to the apparent indeterminacy created by the lack of an account of axiom-data explanation. A second consequence is that it suggests that we hone in on judgements about the extent to which a (collection of) axiom(s) is (are) theoretically virtuous. This is the task of the next section.

[8] Barton, Ternullo and Venturi suggest additional theoretical virtues (though they do not use this name to refer to them): theoretical completeness (where a theory T_1 is more theoretically complete than another theory T_2 *iff* there is a sentence implied by T_1 that is not implied by T_2 and not vice versa), predictive power (in terms of a candidate axiom having consequences that are already provable without the candidate axiom)—and there could well be more.

4 Objectivism versus subjectivism

In the last section I suggested an alternative to the explanatory account of axiom justification, one that places the theoretical virtues at its heart. In this section I consider the question as to whether or not value judgements (or, more generally, facts about ourselves as cognitive agents) play an indispensable role in assessing the extent to which an axiom is theoretically virtuous. After drawing the distinction between objectivism and subjectivism about extrinsic justification, I consider and reject an argument given for objectivism by Maddy. I then provide some considerations that speak in favour of subjectivism.

4.1 Objectivism and subjectivism

What sort of judgement is a judgement about the virtuousness of an axiom or collection of axioms? What sort of property do such judgements identify and is one (collection of) axiom(s) more theoretically virtuous than another solely in virtue of some mathematical facts or in virtue of both mathematical facts and facts about us? Or, as Ernst, Heis, Maddy, McNulty & Weatherall recently put it: is theoretical virtuousness "in the math" or is it "in us"? (Ernst et al. 2015a: p. 157)

Penelope Maddy, in *Defending the Axioms*, provides the most thorough articulation of one kind of answer to these questions, what I will call *objectivism*. I'll first explore what Maddy says about the objective nature of extrinsic justification, before trying to provide explicit formulations of a couple of different claims regarding this objectivity.

Describing the "topography of mathematical depth", Maddy says:

> This topography stands over and above the merely logical connections between statements and furthermore, *it is entirely objective* […] [I]t's not up to us that appealing to sets and transfinite ordinals allows us to capture facts about the uniqueness of trigonometric representations, that the Axiom of Choice takes an amazing range of different forms and plays a fundamental role in many different areas, that large cardinals arrange themselves into a hierarchy that serves as an effective measure of consistency strength. […] [These facts] *are not traceable to ourselves as subjects* (Maddy 2011: pp. 80–81) [emphasis mine]

Maddy also says the following when considering the relationship between extrinsic justification and us, the users of mathematics:

> 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; for that matter, the entire mathematical community could be blind to the virtues of a certain method or enamoured of a merely fashionable pursuit without changing the underlying facts of which is and which isn't mathematically important (*ibid*: 81)

In the first passage, Maddy makes the claim that it is not up to us that Choice has wide-ranging applicability within mathematics, or that large cardinals can serve as a measure of consistency strength, and so on. It is important, though, to distinguish this

from a second claim which Maddy also seems to be making: that it is similarly not up to us that Choice's applicability is sufficient for it being a piece of deep mathematics (or, in the terminology of the virtues account, sufficient for the axiom being theoretically virtuous). It is clear for Maddy that, in addition to the applicability of Choice being an objective matter so is the fact that this suffices for Choice being a theoretically virtuous piece of mathematics. Maddy also says that, for the objectivist, in addition to the "vast net of logical relations" (Ernst et al. 2015b: p. 246) being objective, so too are the "outstanding mathematical virtues of some strains of these" (*ibid*). This second claim is what is at issue. It is uncontroversial that whether or not the continuum hypothesis is a consequence of some collection of axioms is, in some important sense, objective: but the further question is how to conceive of and adjudicate the mathematical virtues of that collection.

The strength of Maddy's objectivism in Maddy (2011) is brought into relief by her treatment of the continuum hypothesis (CH). Another aspect of the claimed objectivity of mathematical virtue, then, is that it assigns determinate truth-values to currently open mathematical questions—the truth-values that result from the deepest mathematics. Maddy makes this explicit when discussing CH:

> [W]e might never come to formulate or to appreciate the virtues of some set-theoretic axiom that would settle CH. In this way, our Thin Realist might never come to know whether CH is true or false, despite it having a determinate truth value. (Maddy 2011: p. 81)[9]

This passage further demonstrates that Maddy thinks the facts about mathematical virtuousness are independent from the goals and practices of actual mathematicians. These facts are *already there*, and it is contingent only whether or not mathematicians successfully identify this or that strain of 'deep' mathematics. On this picture, the truth-value of CH is determined independently of whether or not the mathematical community correctly identifies as deep the set-theoretic axioms that determine this truth-value. Maddy allows, also, for error in identifying virtuous mathematics, claiming that it is possible for mathematicians to "wander off the path of mathematical depth"—if this happens, then said mathematicians are "going astray, even if no one realises it" (Maddy 2011: p. 82)

[9] Maddy's Thin Realist accepts what set theory says about sets but demurs from attempting to answer any further questions about what sets are like. These questions are not settled by set-theoretic practice and so, the story goes, for the properly naturalistic philosopher of mathematics they do not require settling. Maddy holds that a descriptive account of mathematical practice is consistent with both Thin Realism and what she calls Arealism. The Arealist agrees with the nominalist that we lack good reason to think that sets exist—but thinks that this is so because "well-developed methods of confirming existence and truth aren't even in play" (Maddy 2011: p. 89). Maddy holds that Arealism should be thought of as distinct from standard forms of nominalism because such positions appeal to "a priori prejudice against abstract objects" or "preconceptions about what knowledge must be like" (*ibid*: 97). I'm inclined to think that whether Maddy's Arealist is a kind of nominalist will bottom out as a terminological dispute. If nominalism is the view that we lack good reason to believe in mathematical objects, then the Arealist is a nominalist in a sense that the Thin Realist (as well as, obviously, the Platonist) is not. If nominalism, however, must incorporate substantive views about knowledge, or the nature of abstract objects, then the Arealist is not a nominalist. I see no substantive reason for holding that one or other of these candidate meanings of 'nominalism' is to be preferred to the other. Thanks to an anonymous reviewer for pushing me to be clearer on how the relationship between nominalism and Maddy's Arealism is to be understood.

How should we understand Maddy's claim that the facts of mathematical virtuousness are objective? There is an important distinction, one that is sometimes occluded in Maddy's discussion. This is between a stronger and weaker objectivity claim.[10]

Weak objectivity: A (collection of) axiom(s) instantiating a particular theoretical virtue depends only on mathematical facts, and not facts about us.

Strong objectivity: Facts about comparative virtuousness between (collections of) axiom(s) depend only on mathematical facts, and not facts about us.

Maddy's remarks about CH suggest that she endorses both the weak objectivity claim and the strong objectivity claim. The notion of CH having a determinate truth-value (that truth-value assigned to it by the uniquely virtuous (or 'deep') axioms), and it having this determinate truth-value independent of any facts about us, seems to require strong objectivity—this is because it requires there to be a unique collection of axioms that is most virtuous, or deep, and for this fact to be one that holds independently of any judgements we make. Independent of such exegetical concerns, however, what we are interested in is whether either of the objectivity claims are true.

4.2 The phenomenological argument for objectivism

So: should we be objectivists or subjectivists about extrinsic justification? Do judgements about extrinsic justification turn, in any substantive way, on facts about us? Arguments in favour of objectivism and subjectivism are few and far between. In *Defending the Axioms*, however, a germ of an argument for objectivism can be located. In this section I briefly set out the argument and then raise some problems for it.

The argument sketch runs as follows. Maddy notes that proponents of Platonism (or, in Maddy's terminology, Robust Realism) claim that their view accounts for certain facts about "the pure phenomenology of" mathematical practice (Maddy 2011: p. 114) and that this is an advantage of the view. She notes that when doing mathematics, we are not free to "follow our personal or collective whims" (Maddy 2011: *ibid*) and that, instead, mathematics is "an objective undertaking par excellence" (*ibid*). The Platonist tells us that what constrains mathematical practice in this way is a "world of abstracta" that our mathematical claims attempt to say true things about. Maddy suggests that the topography of mathematical depth might be a better ground for this feature of mathematical practice—not least because the "phenomenon of mathematical fruitfulness" (*ibid*: 116) more closely resembles the constraints felt by practicing mathematicians (when compared, for example, to constraints generated by a world of abstracta, which seem like the wrong kind of thing to have an effect on mathematicians). This, then, is a possible argument for the objectivity about facts about extrinsic justification: it is the best explanation of certain features of mathematical practice—namely that it is constrained in some sense by a further set of facts.

[10] There are open questions about exactly how these two kinds of objectivity are to be spelled out and understood. The formulations above are sufficient for capturing the distinction appealed to in the following three sections: in Sect. 4.5 I turn in more detail to the question as to how exactly we ought to understand, in particular, the strong form of objectivity.

Unfortunately, it's currently hard to assess the argument. For one thing, it's unclear what the explanandum is intended to be. It could either be that what needs explaining is the *fact that* mathematics is constrained by some set of objective features, or the fact that the phenomenology of mathematical practice has certain distinctive features—the *feeling that* one's practice is constrained by a set of facts, facts that one's practice ought to be faithful to. Canvassing the candidate explanations, and then choosing amongst them, will be a task of different character depending on which of these explananda is chosen. At a first pass, if what is to be explained is the *fact that* mathematics is constrained by some set of objective features, then the argument is question-begging. What is at issue is whether or not either of the objectivity claims are true, so an argument for (some form of) objectivism cannot take as a premise the supposed fact that mathematical practice is constrained by some set of objective facts. On the other hand, if what is to be explained is the fact about the phenomenology of mathematical practice, then it is unclear why the truth of either form of objectivism would have a good claim to being the best explanation of the target phenomenon. As it stands, the argument looks unsuccessful.

4.3 The weighting argument for subjectivism

In this section I present an argument for subjectivism about extrinsic justification, influenced by Kuhn's later work on theory choice in science.[11]

A decade after the publication of *The Structure of Scientific Revolutions*, Kuhn argued for the existence of a set of theoretical virtues used (apparently *across* paradigms) to select between rival scientific theories (Kuhn 1977). These virtues include (but are not intended to be exhausted by) empirical accuracy, consistency, scope, simplicity, and fertility. Whilst, Kuhn says, even though these virtues are universally used to assess scientific theories, there is no universally used *weighting* of the virtues. Kuhn concluded from this that there is, as he put it earlier, "no neutral algorithm" for theory choice (Kuhn 1962: p. 199).[12] If there were such an algorithm, then it would be possible to move (without any input from us) from the facts about scientific theories and the extent to which they are consistent with the evidence, simple (and so on) to a conclusion about which of the theories we ought to accept. The conclusion we're invited to draw from the 'no neutral algorithm' argument, Kuhn clearly states, is *not* that scientific theory choice is irrational, or that "anything goes" when it comes to choosing between scientific theories. Indeed, as Schindler says:

> The standard criteria of theory choice are not projections; they map into actual theory properties. Theories really are accurate, consistent, fertile, and so on, or they are not. (Schindler 2018: p. 53)

Nevertheless, theory choice involves input from us at a crucial stage: the selection of a weighting of the virtues. Perhaps it is the case that some potential weightings

[11] As should be clear from the below, none of the argument requires accepting any of Kuhn's substantive claims about scientific progress, etc.

[12] See, also, discussions in Sober (2001a, b) and Hesse (1974) about how the theoretical virtue of simplicity might be balanced with the other virtues.

can be intuitively discarded such as those that sacrifice all other virtues in favour of maximising one of them[13]—but there will nevertheless be very many plausible weightings that seem unable to be ruled out.

Kuhn argues for the no neutral algorithm claim, in the context of scientific theory choice, by gesturing towards some examples of the history of science in which disagreement among scientists is best explained by positing disagreement about the weighting of the theoretical virtues (Kuhn 1977: 357–359).[14] I think there is an easier route to securing the no neutral algorithm claim for a domain in which theory choice is guided by use of the theoretical virtues—one that suggests that Kuhn's argument via case studies is potentially misguided. Even if all practising scientists in some domain agreed about the weighting of the theoretical virtues (at least to the extent that their individual weightings were sufficiently similar to never lead to disagreements about which theories merited acceptance, which is the level of agreement relevant in this context), this agreement would not demonstrate that the particular weighting was privileged in virtue of some set of biological (physical, cosmological, etc.) facts.[15] Indeed, if the candidate scientific theories in some domain *did* have ramifications for how we ought to weigh up the theoretical virtues, this, I suggest, would strike us as strange. The theories are about the phenomena of scientific interest, and not about theory choice and virtue weighting. We should not expect our candidate scientific theories to tell us which weighting function to select. The same goes, I maintain, for the use of theoretical virtues to justify axiom choice in set theory. If, to use Maddy et al.'s phrasing, the weighting function is to be located either "in us" or "in the math", then the answer must be that it is located in us—set theory is the wrong place to look for a function for weighting the theoretical virtues.[16] So, even assuming that there *is* some norm about how the virtues are to be weighted that is shared (however implicitly) by

[13] Although we should be careful to not infer from the fact that all of us would discount a weighting that (for example) prioritised simplicity above all other virtues to the conclusion that the discounting of such a weighting stems directly from the scientific matter being dealt with.

[14] The focus of recent critical discussion of the no neutral algorithm claim has concerned whether or not there is *any* algorithm for theory choice that satisfies certain plausible criteria (in short, whether Kuhn thought he saw a vast bounty of potential algorithms when instead there is a lack). Okasha (2011) argues, co-opting Arrow's impossibility theorem from social choice theory, that there is no such algorithm. See Bradley (2017) and Marcoci and Nguyen (2019) for replies.

[15] An anonymous reviewer raises a potential consequence of such consensus: namely, if one thinks that agreement *qua* consensus (amongst some subset of agents) can *in and of itself* be reason to think that this agreement tracks truth, then there is at least one potential sense in which consensus about virtue-weighting and that particular weighting being connected, in the right way, to the relevant set-theoretic facts. For this sort of argument to be made good on, one would have to make the case both that the relevant sociological thesis is true (that such consensus exists) and that the purported connection between consensus and truth can be made good on, two matters that are beyond our current scope.

[16] Consider, as an analogy, Field's informal argument for thinking that mathematics is conservative over physical theories (put loosely, that if a nominalistic statement is derivable from a scientific theory with both mathematical and nominalistic content, then it could have been derived from the nominalistic content of the theory alone—that adding mathematical content to a nominalistic scientific theory doesn't introduce any new derivable nominalistic consequences). Field notes that "it would be extremely surprising if it were to be discovered that standard mathematics implied that there are at least 10^{60} non-mathematical objects in the universe, or that the Paris Commune was defeated […] *good* mathematics *is* conservative" (Field 2016: p. 13). Field's thought here is that it is constitutive of mathematics, or at least of *good* mathematics, that it doesn't (by itself) have consequences for non-mathematical affairs like the Paris Commune. Similarly, some set theory that implied that one particular weighting function was to preferred over another would be

all competent set theorists, this seems to be a norm that must have its origin in us.[17] Indeed, it seems that the user of the theoretical virtues to choose between theories (or axioms, etc.) outside of scientific contexts will have fewer resources at hand to rule out particular weightings. The use of the theoretical virtues in scientific contexts is subject to empirical feedback, allowing us to fine-tune the balance of the virtues against our predictive successes and failures. This is, at least, one way in which candidate weightings might be ruled out by the subject matter of our theories, at least in the scientific context. However, no such calibration seems possible in the case of axiom justification.[18,19]

Footnote 16 continued
bad set theory—weighting functions are not in its proper domain. One should be careful, here: there are some stronger claims in the vicinity that are plausibly false. Some badly-conceived weighting functions might be ruled out for mathematical or logical reasons and, in this sense, mathematical facts can have bearing on the choice of weighting functions. However, barring some extremely surprising result that there is in fact only *one* mathematically possible weighting function that meets a set of criteria, mathematical facts alone are insufficient to single out a particular weighting—and it is *this* that one would require to get the result that the weighting function is found 'in the math' rather than 'in us'. One final consideration (raised by an anonymous reviewer) is that one should leave open the possibility that a weighting function is determined by some non-set-theoretic but also non-subject-sensitive facts—that is, some facts from some other scientific domain. This suggestion does seem to complicate the bifurcation between the weighting function being located either in some facts about us as subjects or in some set-theoretic facts: however, if one is convinced by the argument above (that it is constitutive of good set theory, of good cosmology, etc. that it is silent on weighting functions), then one should be similarly convinced that it is constitutive of other branches of scientific endeavour that they are silent on how virtues ought to be weighted in the foundations of set theory.

[17] In Sect. 5.1. I suggest that the falsity of this claim about a universal norm would constitute a good explanation for some facts about set-theoretic practice.

[18] Saatsi (2017) makes a similar point regarding the use of explanationist considerations to choose between rival metaphysical theories (pushing back against the kind of anti-exceptionalism about theory choice in philosophy that is put forward by Williamson). There is an interesting sense in which the claim about the lack of empirical feedback—at least when it comes to the use of the theoretical virtues in set theory—might be pushed back against, however. As an anonymous reviewer notes, for some of those invested in pursuing the program of extending ZFC, there are constraints on the program being well-founded. That is, the program must involve conjectures such that (if these conjectures fall a certain way), the program is ended. This clearly has similarities with the kind of feedback we may get regarding *scientific* theories when they generate a prediction that fails to pan out. Whether or not the set-theoretic constraint therefore counts as properly *empirical* seems like another semantic choice-point regarding broadening the extension of a term.

[19] Barton, Ternullo and Venturi mention weighting as one of the questions left open by their explanatory account, asking: "is it possible to come up with a way of assigning different theories [meaning collections of axioms] weights and comparing them satisfactorily?" (Barton, Ternullo and Venturi *ms*: 28). Although they mention assigning weights to *theories* (or collections of axioms), it seems as though what we want to assign weights to is the various good making features of axiom collections (the theoretical virtues) and using this weight-assignment to produce an overall assessment of the rival collections. Their open question, then, is slightly different to the question pressed by the weighting objection. Their question concerns whether or not it's possible to produce some means of weighing up axiom collections that instantiate the various virtues to different extents: in Kuhnian vocabulary, whether or not one could produce an algorithm, neutral or otherwise. The weighting argument isn't an argument that there could be no algorithm: it's an argument, in Kuhn's words, that there can be no *neutral* algorithm. So, a positive answer to Barton, Ternullo and Venturi's open question is fully consistent with subjectivism—the subjectivist about axiom justification merely presses that whatever weighting we produce will come from "in us" rather than "the math".

4.4 The weighting argument and Maddy's objectivism

This weighting argument is consistent with much of what Maddy says in her defence of objectivism, but not all. Part of Maddy's claim is that one's "idiosyncratic preference doesn't make some conceptual or axiomatic means toward that goal into deep or fruitful or effective mathematics" (Maddy 2011: 81). Recall that *weak* objectivity is the claim that whether or not an axiom (or collection of axioms) instantiates a particular virtue is a fact that holds independent of any decisions made by us. The weak objectivity claim is not threatened by the weighting argument—as noted above, the fact that we 'input' the weightings of the virtues doesn't make it the case that whether or not the axioms *in fact instantiate* the virtues depends on us in any important sense. If ZFC unifies disparate phenomena, I cannot make it such that ZFC fails to so unify by pure force of will.[20] However, what is threatened—the *strong* objectivity claim—is the idea that all comparative claims of virtuousness are objective, in the sense of being independent from all facts about us and our value-judgements. Comparative claims of virtuousness require a weighting of the virtues, and the weighting of virtues is not something that is to be found "in the math", as it were.[21] Now, some of Maddy's remarks are ambiguous, and it is possible that she only intended the weak objectivity claim.[22] I maintain that this weak reading doesn't sit easily with the comments that she makes about the truth-value of CH being that which is assigned to it by the collection of axioms that is most "mathematically deep" (in her vocabulary). Nevertheless, whether or not Maddy intends the strong or weak objectivity claim, I have presented an argument here that the strong objectivity claim is false.[23]

The emphasis on theoretical virtues shed light on this route to subjectivism about extrinsic justification. However, I take it that even on any version of the rejected explanatory story, the theoretical virtues will still play the part of determining which of the competing explanations is the *best* explanation. It should also be stressed here that the argument presented here is not one for the conclusion that using the theoretical virtues to choose axioms is not truth-conducive: that, in other words, we lack reason

[20] This isn't, of course, to say that it's altogether straightforward to show that, for each of the theoretical virtues, a (collection of) axiom(s) instantiating that particular virtue is subject-insensitive—just that the falsity of the weak objectivity claim isn't a consequence of the weighting argument.

[21] *Some* comparative claims will, practically speaking, require no weighting of virtues—but those that are going on in active set-theoretic debate will.

[22] It is worth mentioning that in Ernst et al. (2015b), Maddy seems less confident of the claim made in Maddy (2011) that depth-judgements are objective in the above sense.

[23] One might worry that taking 'being theoretically virtuous' and 'being mathematically deep' to be functionally equivalent terms occludes a possible response on behalf of a Maddy-influenced objectivist. If mathematical depth is *the* good-making, justification-conferring property of axioms, then the problem of weighting doesn't arise: as there's nothing to weigh depth against. On this line of response, the weighting problem is a consequence of the theoretical virtues account of extrinsic justification, but it's too quick to sign Maddy up to this account. However, I think it's clear that in Maddy (2011), 'mathematical depth' is a term used to refer to a variety of good-making, justification-conferring properties of axioms (and concepts, etc.): Maddy herself says that she "lumped a number of different notions together under a broad umbrella of 'depth'" (Ernst et al. 2015b: p. 245). I think a case can be made that using the terminology of 'depth' (when, at the very least, judgements about depth will involve judgements about fruitfulness, unifying power, etc.) occludes more than it helps: its aesthetic connotations, ironically enough, making subjectivism about such judgements look more tempting than it ought initially to do.

for thinking that being theoretically virtuous makes an axiom more likely to correctly describe the set-theoretic universe. This is a much larger debate: indeed, plausibly only the Platonist owes us some sort of story that connects up the theoretical virtues had by (collections of) axioms to set-theoretic reality. The non-realist (and, perhaps, idiosyncratic forms of realist) can simply accept that (a) mathematicians justify their adoption of set-theoretic axioms on the basis of judgements about theoretical virtuousness and (b) that doing so involves making value-judgements about weightings.

4.5 Comparative and non-comparative conceptions of objectivity

In this section I'll consider a worry about the weighting argument against strong objectivity. In short, the worry says that the argument depends on a particular way of spelling out exactly what the strong objectivity claim says and that if one instead opts for a different, but equally plausible, way of spelling the claim out, the argument fails to get off the ground.[24] The distinction between the different ways of spelling out objectivity concerns whether strong objectivity is spelled out comparatively or non-comparatively.

Recall the spelling out of weak and strong objectivity from Sect. 4.1. The weak objectivity claim says that a (collection of) axiom(s) instantiating a particular theoretical virtue depends only on mathematical facts, and not facts about us. The strong objectivity claim, as understood in 4.1., says that facts about comparative virtuousness between (collections of) axiom(s) depend only on mathematical facts, and not facts about us. Someone sceptical of the scope of the argument could note that weak objectivity is spelled out non-comparatively and that strong objectivity is spelled out comparatively and, further, that the weighting argument against strong objectivity *depends on* strong objectivity being spelled out comparatively. For the weighting argument, recall, says that in order to make claims about comparative virtuousness of collections of axioms, a weighting of the virtues must be selected—and this selection of weighting is inputted by us, rather than being found 'in the math'. If there are plausible non-comparative understandings of strong objectivity, then it seems as if the weighting argument is ineffective against strong objectivity so understood.

There are (at least) two possible routes to assuaging this worry. The first is to accept, for the sake of argument, that the weighting argument depends on the comparative unpacking of strong objectivity but to press that only the comparative unpacking genuinely captures the notion. The second is to show how a user-inputted 'weighting' is required (or, weaker, that there is a substantive subject-sensitive element involved) even on plausible non-comparative versions of the strong objectivity claim. I'll pursue the latter, by presenting two formulations that seem to capture the spirit of strong objectivity but fail to explicitly mention comparisons between collections of axioms, and demonstrating that there are subject-sensitive elements present nonetheless.

Someone sceptical of the scope of the weighting argument could propose the following as an equally plausible spelling out of the notion of strong objectivity: facts about overall virtuousness of (collections of) axiom(s) depend only on mathematical facts and not facts about us. Here, overall virtuousness is understood as distinct from

[24] Thanks to an anonymous reviewer for raising this worry.

the extent to which a (collections of) axiom(s) instantiates some particular virtue (e.g., simplicity)—it is concerned with the extent to which that collection is theoretically virtuous *simpliciter*. On this non-comparative reading of strong objectivity, the objectivity claim is true if facts about the overall virtuousness of a (collection of) axiom(s) depends only on mathematical facts, and requires no 'input' from us—that there is no subject-sensitive component of strong objectivity. The thought, I take it, would be that one can then move from these objective facts about the virtuousness of rival collections of axioms to a belief about which collection we ought to adopt. However, pressure can be put on the claim that one can move from a set of facts about a collection of axioms each individually instantiating this or that particular virtue (the kinds of facts relevant for *weak* objectivity) to a sort-of global fact about the virtuousness of that collection. Moving from the facts about the extent to which the collection is simple (unifying, and so on) to the fact about global virtuousness is going to require a function from individual virtues to global virtuousness—the same kind of weighting we saw in the Kuhnian argument above. Is simplicity as important as unifying power, when coming to a judgement about the virtuousness *simpliciter* of the collection? Are there thresholds past which increases in the ability to unify disparate phenomena become less important, all things considered—some sort of amount of unification that is 'good enough'? These are all questions that need answering when moving from the local facts about this or that virtue to the global fact about overall virtuousness—and the answers, it seems, must come from us rather than from the math. It is important to note that even the most natural response to this question (namely that all of the virtues contribute equally to overall virtuousness and there are no thresholds of the kind previously discussed) is *still* an answer inputted by the person making the judgement!

Here is a second potential way of drawing the distinction between weak and strong objectivity which, also, doesn't seem to consist in strong objectivity being understood comparatively and weak objectivity being understood non-comparatively.[25] On this understanding, weak objectivity consists in (judgements about) (collections of) axiom(s) instantiating this or that virtue, whilst strong objectivity consists in taking these facts about virtue instantiation to speak in favour of accepting that (collection of) axiom(s). On this understanding of the distinction between strong and weak objectivity, then, the distinction is between some facts about the instantiation of the theoretical virtues, on the one hand, and what this instantiation means for our epistemic attitude towards the collections of axioms, on the other. Given this way of drawing the distinction, the relevance of the weighting concerns discussed earlier seem less obviously apparent: a weighting of virtues is not required either for the mere recognition that some collection of axioms has, for example, unificatory power—nor is a weighting required in order to judge that a collection of axioms having this feature speaks in favour of accepting it. This, however, is no comfort to someone looking to insulate from subject-sensitive concerns. On this understanding of the weak/strong distinction, too, subject-sensitive elements seem to have been introduced at the move from weak to strong objectivity. Granting for the sake of argument that the extent to which

[25] Thanks to an anonymous reviewer for suggesting this sort of approach to demarcating weak and strong objectivity.

some particular theoretical virtue is instantiated by a (collection of) axiom(s) is settled by set-theoretic facts, this does not similarly secure the result that such an instantiation speaks in favour of acceptance. Indeed, these facts about instantiation of virtues do not, alone, even speak in favour of a collection being considered a candidate for acceptance. Set-theoretic facts do not settle what attitude we ought to have towards some collection of axioms: these facts do not dictate that we take being theoretically virtuousness to be a (defeasible) guide to acceptance.[26]

We have now seen three potential ways of understanding strong objectivity (the comparative understanding in Sect. 4.1, and the two non-comparative understandings above)—on each, the claim that the matter in question depends in no way on facts about us seems to be false. Before concluding, in the next section I'll briefly consider some potential ramifications of this result.

5 Subjectivism and absolute undecidability

That judgements of extrinsic justification are subject-sensitive in a substantive way promises to have consequences for adjacent debates. Below, I consider one such example.

5.1 Absolute undecidability

As mentioned at the start of this discussion, some conjectures are undecidable relative to the standardly accepted axioms. Such is the justification for Gödel's program of searching for well-justified extensions of ZFC. However, more controversially, it is often thought that in addition to these relatively undecidable claims, there are also *absolutely* undecidable claims. The continuum hypothesis (CH), originally taken by Gödel to be the kind of relatively undecidable statement that motivates the program of extending ZFC, is now often thought to be absolutely undecidable. In addition to being undecidable relative to the ZFC axioms, the thought is that CH is undecidable *simpliciter*. Speaking loosely, there is *no fact of the matter* as to whether or not CH is true.

There is a natural conceptual connection between one's stance regarding objectivism and subjectivism and one's stance regarding absolute undecidability. As noted in passing above, it is part of Maddy's picture that CH has a determinate truth-value: this is partially why it seems clear that Maddy endorses the strong objectivity claim. For Maddy, there is a collection of axioms picked out by the "contours of mathematical depth" such that this collection is uniquely picked out and such that the collection settles CH one way or another. Transposed to the key of the virtue account of extrinsic justification, this amounts to the claim that there is a collection of axioms that is uniquely theoretically virtuous. As one might imagine, there is a subjectivist route to the view that there are absolutely undecidable statements (contra Maddy).

[26] Of course, if we had a compelling argument *that* the theoretical virtues are truth-conducive (either *simpliciter* or in local set-theoretic contexts), then this might constitute the beginnings of a case that the mere *fact that* a collection of axioms instantiates the virtues to various extents speaks (defensibly) in favour of its acceptance.

The question as to how we ought to understand absolute undecidability cannot be decided (!) here. However, one independently plausible account appeals to the notion of cognitively flawless agents (Clarke-Doane 2013: p. 473). For some proposition P, an agent is cognitively flawless with respect to P if they have mastery of the relevant concepts, have awareness of all the logical relations that P stands into other propositions, are rational, are imaginative and inventive thinkers, and so on. They are, speaking loosely, cognitively perfect agents.[27]

The corresponding account of absolute undecidability simply says that a proposition (hypothesis, conjecture, etc.) is absolutely undecidable if there could be disagreement regarding the truth-value of the proposition amongst cognitively flawless agents. If we accept this account of absolute undecidability and the subject-sensitivity of extrinsic justification, then the possibility of absolutely undecidable statements seems to directly follow. Recall that the Kuhnian suggestion canvassed above suggests that whenever we make judgements about the extent to which a theory (or axiom, or collection of axioms, etc.) is theoretically virtuous, an important role is played by a value-judgement concerning the selection and weighting of the theoretical virtues. Rationality does not seem to dictate that one particular weighting of the virtues be uniquely selected over the others; nor do any of the other cognitive features that determine whether or not an agent is cognitively flawless. It seems, then, that two cognitively flawless agents could disagree about which collection of axioms we ought to accept and therefore could disagree about the truth-values of those statements which have different truth-values (relative to the two collections of axioms endorsed by each agent). If there could be disagreement between cognitively flawless agents about these statements then, according to the account of absolute undecidability endorsed here, these statements are absolutely undecidable.[28] It is tempting to think that the truth of subjectivism (and the fact that that it is true, at least partially, in virtue of the kind of weighting considerations discussed at length above) *explains* the possibility of absolute undecidability, via it making it the case that disagreement between cognitively flawless agents is possible. Discussion as to what the other, competing, explanations of absolute undecidability are and (to echo the concern from earlier) exactly what model of explanation is in play

[27] We should stop short of building into the notion of a cognitively flawless agent (regarding some proposition P) the idea that such agents only believe *truly* (regarding that proposition). One could be cognitively flawless regarding some proposition, receive exclusively misleading evidence regarding P, and therefore believe falsely regarding P (but not as a consequence of some cognitive flaw).

[28] An anonymous reviewer expresses the worry that if we can sensibly understand some set-theorist who advocates for the acceptance of ZFC without extension as a cognitively flawless agent, then there might be some sense in which absolute undecidability collapses into ordinary undecidability. However, according to the notion of an agent being cognitively flawless set out above, no actually existing set-theorist is going to count as a cognitively flawless set-theorist (even the best amongst us, sadly, are not logically omniscient, for example). At most, some actual agents are going to be *approximations* of cognitively flawless agents (where this might, but need not, involve such actual agents having the same set of beliefs as cognitively flawless agents). What might be behind the thought that some actual set-theorists may be cognitively flawless, however, is the notion that some actual set-theorists might be epistemically *blameless*. I would be a better cognitive agent were I aware of all the logical relations between propositions (and therefore, I fail to be a flawless agent in the sense that is relevant here), but it would be inappropriate to *blame* me for not being aware of these logical relations, in virtue of the fact that I can do nothing to bring it about that I have such awareness.

when we try and explain something like absolute undecidability, must wait for another time.[29]

One final potential ramification concerns debates about realism. It is often thought that there is an important relationship between the possibility of flawless disagreement concerning some domain of discourse and realism concerning the subject matter of that discourse (see Kölbel 2004 for discussion) and, similarly, between absolute undecidability and realism (see Clarke-Doane 2013: pp. 477–478). The relationship between absolute undecidability and these other questions is something that cannot be settled here, just as a thorough defence of the cognitively flawless account of absolute undecidability must also be carried out elsewhere. However, it has been demonstrated that accepting the theoretical virtues account of extrinsic justification has the potential to have widespread consequences.

6 Conclusion

The understanding of extrinsic justification as consisting in (something like) the selection of the most theoretically virtuous collection of axioms is not a new one—important aspects of the account are present in both historical and contemporary contributions on set-theoretic foundations. However, paying close attention to the use of the theoretical virtues in both axiom choice in set theory and theory choice in science leads to some striking results about the former. First, judgements of global comparative virtuousness are subjective in at least one sense: that making such judgements requires a weighting of virtues to be 'inputted' by us. This is a kind of subject-sensitivity that persists even granting that whether or not a (collection of) axiom(s) instantiates a particular virtue, and the extent to which it does, is an entirely objective matter. Second, this subject-sensitivity plausibly has consequences for the possibility of absolute undecidability. If carrying out global comparisons requires a judgement to be made about weightings, and if two cognitively flawless agents could make judgements about weightings that are sufficiently different to result in different global comparison judgements, then cognitively flawless agents could disagree about which collection of axioms ought to be accepted. Further focus, then, on the use of theoretical virtues to carry out Gödel's program promises both to shed light on the nature of extrinsic justification and on adjacent questions of outstanding philosophical interest.

Acknowledgements Thanks to Alex Franklin, Eleanor Knox, audiences and reading groups in London, and two anonymous reviewers for this journal for helpful comments on various earlier versions of this material.

References

Baker, A. (2009). Mathematical accidents and the end of explanation. In O. Bueno & O. Linnebo (Eds.), *New waves in philosophy of mathematics* (pp. 137–159). London: Palgrave Macmillan.

Baron, S., Colyvan, M., & Ripley, D. (2017). How mathematics can make a difference. *Philosophers' Imprint, 17,* 1–29.

[29] Thanks to an anonymous reviewer for this interesting suggestion about explaining the possibility of absolute undecidability.

Barton, N., Ternullo, C. & Venturi, G. (*ms*). On forms of justification in set theory. http://philsci-archive.pitt.edu/15806/.
Bradley, S. (2017). Constraints on rational theory choice. *The British Journal for the Philosophy of Science, 68,* 639–661.
Cabrera, F. (2020). Does IBE require a "model" of explanation?'. *British Journal for the Philosophy of Science, 71,* 727–750.
Clarke-Doane, J. (2013). What is absolute undecidability?'. *Noûs, 47,* 467–481.
Dougherty, J. (2018). What inductive explanations could not be. *Synthese, 195,* 5473–5483.
Ernst, M., Heis, J., Maddy, P., McNulty, M. B., & Weatherall, J. O. (2015a). Foreword to special issue on mathematical depth. *Philosophia Mathematica, 23,* 155–162.
Ernst, M., Heis, J., Maddy, P., McNulty, M. B., & Weatherall, J. O. (2015b). Afterword to special issue on mathematical depth. *Philosophia Mathematica, 23,* 242–254.
Field, H. (2016). *Science without numbers* (2nd ed.). Oxford: Oxford University Press.
Frege, G. (1884). *Die Grundlagen der Arithmetik*. Breslau: Koebner.
Gödel, K. (1944). Russell's mathematical logic. Reprinted in Gödel, K. (1990). *Collected works, volume II: Publications 1938–1974*, (pp. 119–143). New York and Oxford: Oxford University Press.
Gödel, K. (1947). What is Cantor's continuum problem?. Reprinted in Gödel, K. (1990). *Collected works, volume II: Publications 1938–1974*, (pp. 176–188). New York and Oxford: Oxford University Press.
Imocrante, M. (2015). Defending Maddy's mathematical naturalism from Roland's criticisms: The role of mathematical depth. In G. Lolli, M. Panza, & G. Venturi (Eds.), *From logic to practice. Boston studies in the philosophy and history of Science* (Vol. 308, pp. 223–239). Berlin: Springer.
Kölbel, M. (2004). Faultless disagreement. *Proceedings of the Aristotelian Society, 104,* 53–73.
Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Kuhn, T. (1977). Objectivity, value judgement, and theory choice. *The essential tension* (pp. 320–333). Chicago: University of Chicago Press.
Lange, M. (2009). Why proofs by mathematical induction are generally not explanatory. *Analysis, 69,* 203–211.
Lange, M. (2017). *Because without cause: Non-causal explanations in science and mathematics*. Oxford: Oxford University Press.
Maddy, P. (1997). *Naturalism in mathematics*. Oxford: Oxford University Press.
Maddy, P. (2011). *Defending the axioms: On the philosophical foundations of set theory*. Oxford: Oxford University Press.
Mancosu, P. (2008). Mathematical explanation: Why it matters. In P. Mancosu (Ed.), *The philosophy of mathematical practice* (pp. 137–150). Oxford: Oxford University Press.
Marcoci, A., & Nguyen, J. (2019). Objectivity, ambiguity, and theory choice. *Erkenntnis, 84,* 343–357.
Okasha, S. (2011). Theory choice and social choice: Kuhn versus arrow. *Mind, 120,* 83–115.
Pincock, C. (2018). Accommodating explanatory pluralism. In A. Reutlinger & J. Saatsi (Eds.), *Explanation beyond causation: philosophical perspectives on non-causal explanations* (pp. 39–56). Oxford: Oxford University Press.
Reutlinger, A. (2016). Is there a monist theory of causal and non-causal explanations? The counterfactual theory of scientific explanation. *Philosophy of Science, 83,* 733–745.
Russell, B. (1907). The regressive method of discovering the premises of mathematics. In D. Lackey (Ed.), *Essays in analysis* (pp. 272–283). London: George Allen & Unwin Ltd.
Saatsi, J. (2017). Explanation and explanationism in science and metaphysics. In M. Slater & Z. Yudell (Eds.), *Metaphysics and the philosophy of science: New essays* (pp. 162–191). Oxford: Oxford University Press.
Salmon, W. (2001). Explanation and confirmation: A Bayesian critique of inference to the best explanation. In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical approaches and applications* (pp. 61–91). Dordrecht: Kluwer.
Schlindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge: Cambridge University Press.
Shapiro, S. (2009). We hold these truths to be self-evident: But what do we mean by that? *The Review of Symbolic Logic, 2,* 175–207.
Sober, E. (2001a). What is the problem of simplicity? In H. Keuzenkamp, M. McAlleer, & A. Zellner (Eds.), *Simplicity, inference and modelling* (pp. 13–31). Cambridge: Cambridge University Press.
Sober, E. (2001b). Simplicity. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 433–441). Oxford: Blackwell.

Williamson, T. (2016). Abductive philosophy. *The Philosophical Forum, 47,* 263–280.
Zelcer, M. (2013). Against mathematical explanation. *Journal for General Philosophy of Science, 44,* 173–192.
Zermelo, E. (1967). Investigations in the foundations of set theory I. In J. V. Heijenoort (Ed.), *From Frege to Gödel* (pp. 199–215). Cambridge, MA: Harvard University Press.