Note: Snapshot PDF is the proof copy of corrections marked in EditGenie, the layout would be different from typeset PDF and EditGenie editing view.

Author Queries & Comments:

Q1 : Please provide a short abstract of no more than 200 words.

Response: The paper contains responses to three sets of comments on my book 'The Birth of Ethics'. The first set is from Adam Lerner, the second from Tristram McPherson and David Plunkett, the their from Manuel Vargas.

Q2 : Please provide missing institution for the affiliation.

Response: Australian National University and Princeton University

Q3 : Please provide missing postal address for the corresponding author.

Response: Apt 1308, 19 Marcus Clarke St, Canberra, ACT 2601, Australia

Q4 : Please check that the heading levels have been correctly formatted throughout.

Response: Correct

Q5 : The disclosure statement has been inserted. Please correct if this is inaccurate.

Response: Correct

Q6 : Please provide missing page range for reference "Pettit 2017" references list entry.

Response: 249-59

Q7 : Please provide missing city for the reference "Williams 2002" references list entry.

Response: Princeton: Princeton University Press

Reply to critics the birth of ethics[Q1]

Recto running head : INQUIRY

Verso running head : P. PETTIT

Philip Pettit

Philosophy, Australian National University, Australia, Canberra[Q2]

CONTACT Philip Pettit philip.pettit@anu.edu.au[Q3]

History : received : 2023-11-26 accepted : 2023-12-20

KEYWORDS

- ethics; genealogy; naturalistic realism; expressivism; error theory

Reply to commentaries

The critiques of *The Birth of Ethics* (henceforth BE) that my co-symposiasts have provided are of the very highest quality and I have benefitted enormously from thinking about them and considering how to respond and what to concede. In the following reply, where the responses are arranged in alphabetical order, I seek to keep up my end and do the best I can by the claims made in the book. But I hope it will be clear that I have learned much from each of the commentators: perhaps more, by some metrics, than I ought to have needed to learn. My thanks to all for the care they took over the book and the insights they have shared.

These responses will only make sense against the background of the claims I make in the book and the criticisms I address here. The precis of the book that accompanies the symposium offers a condensed account of the main theses of the book and indeed, anticipating some of the criticisms below, a statement of the lesson I take it to support. I can only refer the

reader to the critiques themselves for an account of the criticisms I address, but I do try to summarize these in the course of my responses.

Response to Lerner

While Adam Lerner advertises a positive view of the project in *The Birth of Ethics*, and in particular of my defense of a naturalistic version of moral realism for ethics, he raises significant challenges for three theses that he finds in the book. First, that moral concepts ascribe natural properties; second, that a 'full-scale' version of moral realism is true; and three, that there is a satisfactory answer to the question, 'Why be moral?' I try to argue in what follows that I can make a reasonable response to each of these challenges **[Q4]**.

First thesis: moral concepts ascribe natural properties

Lerner focuses on the argument within my genealogy for thinking that the Erewhonians would develop a way of thinking and speaking that directs them, not just to what is desirable, but to what is desirable from all points of view: to what is multilaterally desirable, as I put it. I maintain, as he summarizes the argument (L), that their discourse of the multilaterally desirable is 'expressively equivalent' to our thought and talk about moral desirability, and that since their 'concept of multilateral desirability ascribes natural properties', so 'our concept of moral desirability' does so too; this conclusion supports the first thesis that he wishes to question. He allows the other premises to stand but questions the conclusion on the grounds that the claim about expressive equivalence is suspect.

The claim is that we will take the Erewhonian to be speaking 'in broadly the same terms as us' about matters of desirability insofar as we find that 'we are using those terms under the same prompts and to the same purposes' (BE 25). Lerner questions this claim in the first place on the general ground, supported in social psychology, that we are subject to the egocentric bias of 'assuming that other people's mental states resemble' our own when they may not do so, and that 'this finding extends to the domain of linguistic processing'. As a philosopher with a penchant for the a priori, I can only hope that this finding is consistent with the presumption, surely borne out in experience, that frequently, if not always, we do manage to converse to our informational benefit with one another. That presumption is all I need. And to my mind, it is supported by the fact that while we do sometimes conclude that we are not speaking in the same terms as another – that our differences are merely verbal – we treat that sort of case as exceptional.

Lerner also questions the equivalence claim, however, on more familiar philosophical grounds, as we must now see.

He begins by endorsing, at least for purposes of argument, the assumption I make that a thought or utterance will ascribe the property of being desirable to an option or arrangement – even desirable from a merely personal or religious or national or indeed neutral perspective – just insofar the three constraints mentioned in the precis are satisfied. He focuses on moral desirability in particular and is also prepared to endorse the two further assumptions that I take to mark it off; one requires the content to be non-indexical, the other requires the truth-value to be non-relative. He agrees that if Erewhonian talk of multilateral desirability satisfies all five, then it will be expressively equivalent to our talk of moral desirability, but he argues that the non-indexicality assumption raises problems for the equivalence claim.

Lerner begins his critique by observing that on a view I have endorsed in a number of writings, including *The Birth of Ethics*, there are different ways we may think of a moral property like that of moral desirability. Given that the property is associated with fulfillment of the role defined by the five constraints listed, the moral desirability of a given option or arrangement may be identified with one of two distinct properties. Either the ground-level property in virtue of which it realizes that role or the higher-level property of having some ground-level property, maybe this, maybe that, that realizes the role.

Now according to the genealogy, the Erewhonians will target properties in the category of multilateral desirability, as I describe it, insofar as the acts of making commitments to one another, some individual, some joint, involve them treating certain desiderata as 'attractive robustly over variations in any factors other than the relevant … desiderata' (BE 163). If those desiderata are to make something multilaterally attractive, then they must ensure that the option or arrangement they support satisfies the three constraints associated with desirability of any kind. But, even more important, they must also be properties that matter on all sides, not just in an indexical or relative way. A speaker who avows the desire must do so on the manifest presumption that they speak for others as well as themselves: that they are entitled to commit to that desire not just in a personal or sectarian way but on a multi-lateral basis.

But what property will constitute the desirability of the option or arrangement at issue? One possible candidate is the ground-level desideratum that actually attracts the avower. And another is the property of the target that consists in its

having a realizer, maybe this, maybe that, to play such a role. Whether the desirability is identified with the realizer or the role property, so Lerner argues, there are problems in prospect.

Suppose first that the desideratum is the ground-level property: for example, the property of reducing pain or ensuring peace. And suppose that the Erewhonians differ in the desiderata that move them when they each commit multi-laterally to a desire for something, agreeing that it is desirable in the multi-lateral way; one may be drawn to the target because of its reducing pain, for example, the other because of its ensuring peace. In that case, Lerner argues, they will be talking past one another and not speaking in a non-indexical fashion; multi-lateral desirability – and the moral desirability it is supposed to model – will be the property of reducing pain for one, the property of ensuring peace for the other.

My response to this claim is that while it is true in this case that the truth-maker for a claim of multi-lateral desirability will be the target's reducing pain in the one case, its ensuring peace in the other, the truth-condition of the claim in both their mouths will be the same, so that they will not be talking past each other. The reason for this is already stated in an observation that Lerner quotes from a joint paper I published with Frank Jackson in 1995. In that paper, we argued that while a property like rightness might be identified with the realizer property that makes it right – say, its maximizing happiness – it might equally be identified with the role property of its having such a realizer property. But does that mean that people who focus on different desiderata ascribe different properties, when they describe some choice as right? No it does not, we say, because 'the question does not bear on the conditions for being right but on the metaphysical matter of what rightness, the moral property is' (Jackson and Pettit 1995, 28).

I take the same line in responding to the charge that Erewhonians who are moved by different attractors – the reduction of pain, perhaps, or the promotion of peace – when they treat something as multi-laterally desirable will talk past one another, addressing different properties, when they ascribe such desirability. That the target they address is desirable in the multi-lateral – and presumptively the moral – way will be ensured by its being robustly attractive on all sides, even when those sides differ about which property makes it attractive in that way. As this is put in the book (BE 163): 'claims about what I ought to believe and desire' are 'general truths that hold in abstraction from the precise identity of the relevant data or desiderata'.

So much for the difficulty alleged by Lerner if multi-lateral desirability is identified with the relevant realizer property. What of the difficulty he thinks will follow from identifying it with the corresponding role property? The problem he alleges, in this case, is that if we human beings are to be said to identify moral desirability in a manner that parallels the Erewhonian concern with multi-lateral desirability, then it cannot be a property that satisfies non-indexicality in our case.

The Erewhonians take the desiderata that make something multi-laterally desirable to be 'attractive robustly over variations in any factors other than the relevant … desiderata' (BE 163), and attractive both for themselves and for others too: the open range of others in whose name they presume to speak in co-avowing the corresponding desire. But that something is robustly attractive does not mean that it is necessarily or inevitably attractive, only that it is routinely so. Thus, on the line taken in the book, the Erewhonians will come to recognize that there are certain factors, labeled distractions and disturbers, that may lead an individual not to live up to the avowal of a desire, multi-lateral or otherwise, or to an avowal manifestly made in their name. And equally they will recognize that something will truly prove to be multi-laterally desirable, only if there are desiderata to ensure that it has a grip on those on every side: only, as I suggest, if none of the desiderata are rancorous in imposing harm on some or rivalrous in restricting the benefit on offer.

This qualification means that the recommendation to do that which is multi-laterally desirable is conditional: it holds if and only there are no distractions or disturbers present and support is provided by non-rancorous and non-rivalrous desiderata. But any recommendation of this kind may presuppose either a procedural or a substantive conception of the conditioning factors. Think of the adage: in Rome, do what the Romans do. This may he offered as advice in the absence of a conception, shared between adviser and advisee, about what the Romans do: the idea may be that the advisee should find out what they do and act likewise. Or it may be offered as advice on the assumption that the two each know what the Romans do. In the first case the recommendation is conditioned procedurally, in the second substantively.

On Lerner's reading, the recommendation in the case of multi-lateral desirability is intended to be understood substantively rather than procedurally. And so he says that while the Erewhonians may agree on what are distractions and disturbers and what are to count as non-rancorous and non-rivalrous considerations, ordinary human beings do not. But if humans are to mean the same thing in ascribing multi-lateral desirability to anything, that 'requires them to agree' on such matters (L). And since they don't agree, as he assumes, they must ascribe different properties in such a case and talk past one another.

The response to this line of criticism, of course, is that for both Erewhonians and humans, the recommendation to do that which is multi-laterally desirable is to be understood procedurally and so in the same way on all sides. The idea is that if any

subjects find that they differ on such a recommendation, they should follow a procedure for negotiating their disagreement: a procedure that might be expected to identify what we theorists would cast as a distraction or disturbance on one or another side, or the neglect on one or another side of considerations that are non-rancorous and non-rivalrous (BE 100-01). Thus, in words that any Erewhonian or human subject might use, 'the criteria by which something counts as a disturber' or a distraction are 'internal to my practice', where this is 'bound to be available to me and to others' (BE 98).[1] And while each of them will have to filter out desiderata that 'put people in inescapable competition with one another' (BE 184) when they judge of multi-lateral desirability, they need not share~~have~~ any substantive conception of such rancorous or rivalrous considerations.

Second thesis: 'full-scale' moral realism is true

I allow that even if the desiderata supporting a claim of multi-lateral desirability are non-rancorous and non-rivalrous, still they 'may be weighted differently' among the Erewhonians and 'may block the emergence of full-scale convergence' (BE 191). I argue, however, that the Erewhonians 'may expect to achieve actual agreement over a broad range of issues'. For example, so the line goes, they will probably 'agree that it is multi-laterally desirable' to abide by socially essential norms – 'to tell the truth, abstain from violence, avoid fraud, and so on' – when 'their demands are clear and conformity is not likely to trigger any exceptional, multilaterally undesirable costs' (BE 191).

Lerner suggests that in making this concession, I undermine what he describes as 'full-scale realism'. By such realism he appears to mean a view under which matters of multi-lateral or moral desirability are not subject to 'widespread moral indeterminacy' (L).[2] He holds that widespread moral indeterminacy is a real prospect, on the view I adopt, unless at least many of the differences of judgement that I allow can be shown to be 'the products of distracting or disturbing influences'; he sets aside for purposes of the argument another possible source; 'a failure to filter out rancorousness or rivalrousness' (L). But in running this line, he again intrudes a misreading of my view – an intelligible misreading that I ought to have done more to combat – of a kind with that which affects his criticism of the first thesis.

As I see things, the Erewhonians will identify the influences that we theorists describe as distractions or disturbances on a procedural basis: they will be those factors that show up within their practice of negotiating differences, when they seek to explain why one or another side should concede. They need not have an agreed list of those factors that they can introduce in negotiation, even if a list of likely factors emerges over time; they will identify them case by case, trying to explain why they should disagree with one another despite being presumptively responsive to desiderata that are robustly attractive for all.

On Lerner's reading of my position, however, the Erewhonians will be able to achieve convergence only to the extent that there is a plausible 'set of influences' that can be singled out in advance of the negotiation of this or that difference as 'distracting or disturbing' factors (L). He is rightly dubious about this being possible, however, for two reasons. One, because of the 'stark moral disagreement about a number of moral questions' among our own kind, 'including same-sex marriage, patriotism, abortion, euthanasia, and the death penalty' (L). And two, because people's positions on such issues are likely to dictate what they see as distractions and disturbances, rather than the other way around (L).

I agree that the chances of an advance agreement on what are to be treated as distractions and what as disturbances is unlikely among Erewhonians, as it is unlikely among us, and that it will not offer a means of reducing the disagreements that surface on issues of multi-lateral desirability. But the claim I make is simply that insofar as the Erewhonians have practices under which they seek to resolve differences, and insofar as common norms will argue for the importance of achieving agreement on various issues, we need not despair of their coming to a single mind on at least some such questions. Those practices will push them to identify obstacles that may be getting in the way of agreement: obstacles, in our terms, like distractions and disturbances, or the failure to put aside the rancorous or the rivalrous. Thus, the availability of the practices can nurture a hope among the Erewhonians of achieving convergence in this or that case, enabling them to stay in the business of negotiating differences rather than becoming resigned to irresoluble widespread disagreement: enabling them, in other words, to think as realists about the exercise.

Third thesis: there is a sufficiently good answer to the question, 'Why be moral?'

Lerner begins the last, long section of his paper by distinguishing between the traditional question as to why an agent should behave morally, on a presumptively independent account of what morality requires, and the question as to why they should be behaviorally faithful to their own judgment of what morality requires. He justifiably criticizes me for not paying proper attention to that distinction in my discussion of the issue in the final chapter of the book. And then he lays out important and detailed challenges. Rather than deal with these individually, however, I would like to address the main

criticism in this response. I do so by offering a restatement of my view but one, I think, that does not revise the overall position defended.

The discussion in the book does not distinguish between the specific question as to why an Erewhonian agent should be moral – why they should act on their moral values – and the more general question as to why they should act on their values, moral or non-moral. It will be useful to say something about this general question before looking at the specific.

In looking at both the general and specific questions, however, it is important to distinguish two forms each may take. The question, in the first form, is whether an Erewhonian agent should act on what they actually value as distinct from what they think they value. The question, in the second form, is whether they should act on what in some sense they ought to value as distinct from what they actually value. The distinction between these forms reveals an ambiguity in the issue formulated by Lerner in the moral case, at least as things appear within the context of our genealogy. It will be useful to look at the questions in turn. I begin with a consideration of the general question in its first form, then look at the specific question in that same form, and turn finally – but briefly – to both questions in the second form.

Should such an agent act on what they actually value or on their conception of what they actually value? Relating the robustly persuasive to belief, the robustly attractive to desire – and not yet attending to that which is robustly attractive on a multi-lateral basis – the book takes a clear line on this issue, arguing that the Erewhonians should act on their actual values: that is, on the desiderata that are robustly attractive for them as a matter of fact, whether or not they recognize this. The Erewhonians 'ought to believe that which is robustly persuasive, desire that which is robustly attractive, and avoid beliefs and desires that fail these conditions', and they ought to do so 'even if do not currently recognize that it is persuasive or attractive' (BE 163).

It may be clear why people ought to act on their beliefs about what they actually value; neglecting such beliefs would be a failure in them as intentional systems. But why ought they act on what they actually value: on what they find robustly attractive and fit to support their commitments? Because, in virtue of being commissive agents, as noted in the precis, it will be a functional failure on their part to act on a conflicting desire instead; if they are to be true to the commitments they make, they must be behaviorally responsive to the robust desiderata supporting the commitments (BE 165).

What establishes that certain properties are robust attractors for an individual, given that the agent may not recognize them as such? The fact that they routinely sustain that agent in the commitments they make, keeping them true to the attitudes they avow and pledge. Certain properties may engage someone as robust attractors, constituting their values, without the individual recognizing that they play that role. Thus, the individual may overlook the presence of those values and fail to make a commitment they support; or they may make that commitment and mistake other features for the values that support it.[3]

The motive that any Erewhonian will have to identify robust attractors and live up to them in the commitments they make and keep is connected with the status they will have as persons. In making commitments, actively and virtually, they will invite others to rely on the persona or image they project and authorize. In doing that sincerely, of course, they will also rely on themselves to live up to that persona. Having 'learned what it is to be on the hook with others' – and why they should avoid that hook by being true to their authorized persona – they will recognize the hook they must also avoid in their relationship with themselves (BE 164–165).[4]

In virtue of identifying and investing in a persona – in virtue of personating, as Hobbes puts it – Erewhonian agents will count as persons rather than just agents on a par with other animals or with robots. They will live under 'a dual commitment: as the person spoken for, to conform to what the person speaking says; and as the person speaking, to ascribe only such attitudes as the person spoken for is likely to be willing to display' (BE 123). Let them fail in that commitment, and they will fall short of the ideal of a whole or integrated person. They will be in danger of becoming wantons in Harry Frankfurt's (1971) use of that term: creatures who act on whatever desires or inclinations or impulses assail them at any moment rather than investing in the identity portrayed in their commitments.

This line of thought supports a clear answer to the general question of why the Erewhonians should act on that which they actually value. It will damage their very standing as persons – they will fail the aspiration to integrity that personhood entails – if they do not live up to what they actually value, at least in circumstances where there are no red flags: no indication, for example, that this would entail special costs. Living up to what they actually value in such circumstances – making and keeping the commitments that such values support – is a constitutive ideal for persons, on the personation theory of personhood.

What of the specific question, again in this first form, of whether the Erewhonians should act on that which they actually

value from a moral perspective? Well, by the argument that applies in the case of all values, it suggests that the Erewhonians should act on what they actually value in this way, assuming that the relevant multi-lateral values are not outweighed by other values. But is there something special about the multi-lateral, and presumptively moral, values that the Erewhonians will embrace? Is there something that might argue in support of the view that there is a particularly strong reason why they should act on their moral values?

Yes, there is. The concept of multi-lateral desirability will enable the Erewhonians to foster mutual reliance among themselves, invoking 'considerations that are relatively unrestricted in the range of interests invoked, relatively unrestricted in the standpoint adopted, and consequently fit to play an authoritative role in adjudicating certain clashes among other judgments of desirability (BE 194). And with such a concept established, the Erewhonians will be fit to be held responsible by others in general for how they behave, at least when multi-laterally desirable standards get to be accepted as 'shared and routine' (BE 197). Since these facts are going to be manifest, it will be clear to each that their status as a person is particularly at risk in how they perform on this front. Thus, the considerations that argue for an individual living up to what they actually value in any mode – that their wholeness or integrity as a person is at stake – argue with particular force for living up to what they value on a multi-lateral or moral basis.

Returning now to Adam Lerner's question, it should be clear that in one sense the book asks after whether the Erewhonians should act according to their moral values – that is, according to what they actually value in a moral way – not whether they should act according to their beliefs about those values. But as we saw, there is another question that remains unaddressed. To take that question in its specific form, should they act according to what they ought to value in a moral way? And, turning to the general form, should they act according to what they ought to value in any mode of valuing?

The book does not address the general form of this question, since it has little or nothing to say on what the Erewhonians ought to value in non-moral modes of valuing. For all it assumes, there may be nothing to determine what an agent ought to value in their own individual life, or what they and some others ought to value in their particular relations. But what of the question in the more specific form? What of the question that Lerner probably has in mind when he asks whether the Erewhonians have a reason to be moral? The book says nothing explicitly on this question but it does have an implication that bears on it.

As just mentioned, it assumes that the Erewhonians will have an especially strong reason to abide by standards that they actually value in a moral mode, provided that those standards are shared and routine in the community. And it argues that this proviso is satisfied by standards that are bound to emerge in Erewhon 'like those exemplified in social norms against lying, violence, and fraud, and against infidelity in avowals or pledges' (BE 197). This claim has an implication for the question raised. Standards that are accepted in this sense, and are robustly attractive on all sides, plausibly count, not just as standards that each actually values, but as standards that they ought to value. They ought to value those standards in view of the goal that would drive them, according to the genealogy, to invoke multi-lateral values in the first place: the 'prospect of overcoming various conflicts over desirability, as the master category of common credibility entrenches a prospect of overcoming conflicts on issues of credibility' (BE 197).

The upshot is that according to the genealogy, the Erewhonians will at least have a reason to be moral in a restricted range of cases: they will have a reason to follow standards that are multi-laterally desirable, are recognized as such, and are established as shared and routine.[5] That answer to the question is restricted insofar as it says nothing on whether someone has a reason to follow a standard that is multi-laterally desirable, when that standard is not generally recognized by others as multi-laterally desirable or, if recognized, is not established as shared and routine. Standards in that category may be such that their full desirability depends on being generally recognized and established, as in the case of a standard against free-riding in any domain. Or they may not be dependent in that way; indeed they may depend on others failing to act appropriately, as in the case of a standard of benevolence towards those abused or neglected by others. The book says nothing on why the Erewhonians should abide by such standards as distinct from standards that exist as social norms.

I see that as a shortfall in what the project achieves, to be sure, but not as a problem in how the project is conducted. And there is also something to say that may compensate for the shortfall. If various Erewhonians are likely to hold that it would be multi-laterally desirable for individuals to satisfy certain non-established standards, then we should expect them, assuming some chance of at least partial success, to advocate for the recognition and acceptance of those views. And if they do campaign in this way for those standards, this will presumably lead them to use the ought-language available and to say that people ought to recognize those standards as multi-laterally desirable and ought to comply with them. If advocates campaign for a given standard in this way, will they be disposed to hold others responsible for abiding by it (rather than treating it as supererogatory)? By the account in the book, they may do so if they can invoke a shared standard – say, that all humanoids count as equals – in the act of exhorting those others ex ante to conform and in holding them

responsible to that standard ex post.

Regardless of whether this idea is accepted, however, the shortfall we have registered may help to address a complaint that Lerner makes in the conclusion to his piece. This is that ~~because~~ I do nothing in the genealogy to explain why there is so much dissension among our own kind about issues of morality: strictly, issues where we take morality, despite that dissension, to rule determinately (L). But if I do nothing to explain why there is no such disagreement in Erewhon, it is still intelligible that such disagreement should emerge. If some among the Erewhonians believe that certain unrecognized and unimplemented standards are multi-laterally desirable – and important for that reason – it is robustly probable that they will campaign for their acceptance, generating dissension with those who disagree and campaign for rival standards instead.

Response to McPherson and Plunkett

The paper by Tristram McPherson and David Plunkett (henceforth M&P) has an eye-catching clarity of structure but, as I shall try to show, is based on a reading of my narrative about the imaginary Erewhon and its inhabitants that overlooks an important aspect of the enterprise. I shall make a case against seeing the project as they do; argue on that basis that their main line of criticism misfires; comment on some less central issues they raise in passing; and conclude with a qualified welcome for their suggestion that the genealogy can be reinterpreted as an account of how we ought to think about our ethical practices and concepts rather than as a proposal about how we actually do so.

M&P's reading of the genealogy

The reading that M&P offers of my argument might be taken as a gloss on the following statement early in the book (BE 22):

> The argument involves two claims. First, that insofar as the terms or concepts that emerge in the story respond to the same sorts of prompts, and serve the same sorts of purposes, as our actual ethical terms, the properties they predicate are good candidates for the properties we ourselves predicate with our terms. And second, that since the appearance of those concepts in a predicative role is naturalistically explicable, the properties they ascribe – and the properties ascribed by our counterpart concepts – must be naturalistic, too.

They endorse the idea that these are the two claims crucial to the argument. But they reverse the order of the claims and give a rather different account of their content. They use the term 'ethical' in describing the claims and we may take this to register the evaluative character of the claims about desirability – perhaps moral desirability, perhaps not – and the reprobative character of the claims about responsibility. On their reading, the argument turns on the theses, understood in the following way (M&P)

> **1** 'Erewhonian naturalism ~~naturalism~~: Naturalistic realism is highly plausible as an account of Erewhonian "~~'~~ethical"~~'~~ concepts'.
> **2** 'The linking thesis: Erewhonian "~~'~~ethical"~~'~~ concepts are relevantly similar to our ethical concepts': that is, similar enough 'to allow us to conclude that naturalistic metaethical realism is true'.

Naturalistic metaethical realism – or just naturalistic realism – can be characterized roughly as follows. Those who employ ethical concepts in their thinking and speech use the concepts to ascribe properties in a regular truth-conditional manner; those properties are broadly of the descriptive kind that can figure in natural science (Jackson and Pettit 1996); and the properties are not imaginary or illusory.

Such naturalistic realism may be rejected by those in three salient camps, as M&P say. Expressivists deny that the ascription is truth-conditional, or at least truth-conditional in the regular way. Non-naturalistic realists deny that the properties that ethical concepts are used to ascribe are of a kind with those of natural science. And error theorists deny that the properties ascribed are bona fide properties, treating them as fictions or illusions.

M&P's dilemma for the genealogy

M&P argue that the two-stage argument they find in my book is exposed to a damaging dilemma, which turns on the question of whether theories like expressivism, error theory and non-naturalistic realism offer accounts of Erewhonian ethical concepts that compete for plausibility with naturalistic realism, contrary to thesis 1: 'Erewhonian naturalism'. The dilemma is this (M&P). If those theories are reasonably plausible in the Erewhonian case, then there is no point in looking at that counterfactual possibility: 'it is hard to see how the appeal to Erewhon makes any dialectical progress'. And if they are

not reasonably plausible in that case, then there is nothing to learn about the actual world from the counterfactual possibility: Erewhonian ethical concepts will likely be 'quite different concepts' from the ones we have.

This is a nicely framed dilemma but it obviously starts from M&P's account of the argument, and not from that which I offer in the passage quoted from the book, and indeed in many similar passages elsewhere. The account differs from my own in its version of both the crucial claims.

The less significant difference in their reading of the claims affects the second 'linking thesis' and while it is important to notice this, I will not give it much attention in this response. The linking claim that I make is merely that insofar as Erewhonian ethical concepts play broadly the same role as ours – insofar as they are applied in response to the same prompts and serve the same purposes – 'the properties they predicate are good candidates for the properties we ourselves predicate with our terms' (BE 53). The idea, canvassed at various points in the book, is that we should be at least ready to credit and explore the hypothesis that the sorts of properties that the Erewhonian concepts predicate are the sorts of properties that our concepts also ascribe.[6]

There is a highly significant difference, however, between the Erewhonian naturalism that M&P ascribe to me and what I actually say in the book. And if we reject their reading, as I shall argue, then we can deflect their main line of criticism.

Resisting the dilemma

According to them, the claim I make is that naturalistic realism is a highly plausible account of the ethical concepts employed by Erewhonians and, in particular, that it is more plausible than expressivism, error theory or non-naturalism. The assumption is that in thinking about ethical thoughts and utterances in Erewhon we face the same interpretive problem – which of the four theories to endorse – that we confront in thinking about our own ethical thoughts and utterances. And the thesis ascribed to me is that this problem is readily resolved in the counterfactual, Erewhonian case in favor of naturalistic realism, by contrast with what is true in the actual case of our own ethical concepts: there, by presumption, the debate between such realism and its three rivals remains relatively open.

With things set up in this way, M&P can give life to their dilemma. If naturalistic realism is not more plausible in the counterfactual than in the actual case, then there is nothing to be gained by looking at the counterfactual Erewhonians. And if it is more plausible in the counterfactual than the actual case, then that undermines the reason, which is registered in the linking thesis – and this, even on my weaker version of that thesis – to think that the counterfactual world is sufficiently like the actual world to support naturalistic realism about that world.

This dilemma loses its force, however, once we recognize that on my version of the crucial claim, Erewhonian thought and language does not present, side and side with actual counterparts, as something to be interpreted in the same way. It is not taken as a framework of thinking and speech that raises the same interpretive questions about the relative plausibility of naturalistic realism in comparison with expressivism, error theory and non-naturalistic realism.

On my account, there is indeed a framework apparent in how we human beings think and speak and there is a live question about the relative plausibility of these four theories. But the genealogy does not present a comparable Erewhonian framework that is up for interpretation in the same way. Rather, it does something quite different. First, it gives an account of how, on purely naturalistic assumptions, certain social practices would be robustly likely – that is, likely independently of lucky circumstances (BE 38) – to arise among the Erewhonians, despite their lacking certain evaluative and ethical concepts. And second, it argues that those practices would be robustly likely to lead the Erewhonians to value as well as desire things, judging them to be desirable in one or another mode; to value them in a distinctively ethical fashion, judging them to be morally desirable; and to hold one another responsible – and to hold themselves responsible – for living up to shared values.

Think by analogy of the case where it is unclear how to interpret the goal of a presumptive artifact. Take a little device that operates as follows when put on a table with bottles strewn about: it appears to scan those objects with antennae on its head that swivel in different directions; it fixates on any bottle on its side, moving towards it on little wheels; and then it uses levers on its side to lift the bottle into an upright position, dropping it off the table if it is near the edge. Is the device built just to put bottles upright, knocking them off the table by accident if they are near the edge? Or is it built either to put bottles upright or to knock them off the table?

If we know nothing of the designer, these interpretations are both open. But if we learn that the designer constructed the device in order to ensure that the bottles are all upright, and that they were unable to prevent it from knocking those near the edge off the table, then we know that the first interpretation is correct. To know how something is made is to know what it is, as the so-called *Verum factum* principle holds (Pettit 2008, 20–22).

In a similar way, I take it that we know from how the Erewhonians construct their evaluative and reprobative language, albeit they construct it without planning or intention, that it is truth-conditional, that the properties it posits are really there for them to track, and that those properties are constituted by the descriptive, natural character of the world with which they interact. We know from its mode of construction that the Erewhonian language has a character that resists interpretation in expressivist, error-theoretic or non-naturalistic terms.

The genealogy is reconstructive, not reductive

This two-part story, as emphasized in the book, is meant to be reconstructive but not necessarily reductive, at least in the sense that a reduction of a claim would provide an alternative way of making the claim. That is to say, it aims at naturalistically explaining how the Erewhonians might come to register certain properties from within the practices they develop, introducing terms or concepts to single them out. And it aims at doing this without aspiring to be able to offer a naturalistic analysis or paraphrase – in effect, a reduction – of any of those concepts. As the book says, 'it does not require us to provide naturalistic paraphrases of problematic concepts, only a naturalistic account of how people might have come to have a referring use for similar concepts' (BE 301).

The line taken is that the story told about the Erewhonians provides a naturalistic account of the conditions under which they would be able to have a proper, primary sort of access to properties of desirability and responsibility. But the story will not necessarily provide such an equivalent form of access itself; it will not offer an analysis of those evaluative and ethical concepts in naturalistic terms, delivering naturalistic translations or paraphrases. It 'identifies their acquisition conditions, not their application conditions' (BE 263). The idea is that by evolving naturalistically intelligible practices of commitment to others and of exhorting them to live up to their commitments, the humanoids would be enabled 'to access properties like those of desirability and responsibility'; and this in such a way that it is hard to see how they 'could properly or non-parasitically grasp those properties without familiarity with such practices' (BE 261).

How might we human beings be persuaded that the practices invoked in the genealogy would enable the Erewhonians to gain a proper, non-parasitic form of access to evaluative and ethical properties? According to the book, 'the reconstructive analysis will have to appeal for its confirmation to our capacity to simulate how things would present in the wake of evolving developments of practice, and how insiders to those developments would find certain patterns salient and could evolve terms or concepts to articulate them' (BE 27). The idea, for example, is that as the Erewhonians come to see a property as robustly attractive enough, other things being equal, to support a commitment to acting on the desire to realize that property, they would treat the property as we treat desirability, responding 'to the same sorts of prompts', and invoking it to 'the same sorts of purposes' (BE 22).

The idea that the genealogy may naturalistically explain how the Erewhonians can access properties like those of desirability and responsibility without offering naturalistic analyses or reductions of corresponding concepts should not be unfamiliar. We may explain in non-indexical terms how subjects can get to speak and think indexically without being able to analyze or reduce indexical claims in non-indexical terms. The impossibility of such an analysis is well documented in the literature on *de se* attitudes: believing that PP is in danger may not cause me concern – I may not know that I am PP – but believing that I am in danger certainly will (Lewis 1979; Perry 1979). But nonetheless, we can readily explain in non-indexical terms how someone may get to speak and think in terms of 'I' or 'now' or whatever: all they need to do is let 'I' refer to the speaker, let 'now' refer to the time of utterance, and so on. In this case too, the philosophical account gives us acquisition conditions for indexical concepts but not applications conditions.

We recognize the conditions provided by the account as acquisition conditions in virtue of simulating the position of those who learn the rules that determine how indexical concepts gain referents. And knowing what it means to think indexically, we can recognize in our simulation that by learning how to use 'I' and 'now' appropriately, those subjects would be led to think in broadly that way too. Similarly, when we simulate how the Erewhonians would see things from within the practices of commitment, exhortation and the like, we will be able to recognize from our own habits of thinking in evaluative and reprobative terms, that they would be led under the influence of their practices to think in those same ways.

This observation on the way in which Erewhonian thought and speech are presented in the book, and on the contrast between the framework ascribed to them and the framework we confront in our own evaluative and ethical thinking, should help explain why the approach taken is not exposed to the dilemma posed by M&P. The book does not invite us to consider Erewhonian language in the same way we might consider an actual language, as if it left open issues like those raised by expressivism, error theory and non-naturalism. Rather, the assumption is that knowing how the Erewhonian language emerged, we know that it is truth-conditional, that it posits only bona fide properties, and that those properties are revealed to its speakers in the social world they construct and confront. That being so, the approach taken is not exposed to the

dilemma posed.

On expressivism, error theory and non-naturalistic realism

That line on the genealogical project should also explain why, as M&P complain, I pay little attention to the possibility that other theories besides naturalistic realism might hold of the Erewhonian framework of thought and speech.

Consider the proposal that expressivism might be true of Erewhonian discourse with all evaluative claims, and with properly moral claims in particular. As M&P say, a standard argument in favor of expressivism is that any property such as N that is offered in a realist spirit as a naturalistic, analytically defensible counterpart for the property of desirability or moral desirability will raise the question of whether N really is desirable: it will leave that question open, contrary to the assumption that N is analytically tied to desirability. Is this open-question argument likely to pose problems for a naturalistic realism about Erewhonian discourse? Surely not. The genealogy invoked in favor of such realism does not purport to provide an analytically derivable counterpart of desirability. As emphasized, it merely explains why the concept of desirability should gain useful employment from within independently intelligible – indeed naturalistically intelligible – Erewhonian practices.[7]

But not only does this conception make the genealogy safe against the open-question argument for expressivism; it also reduces other temptations to think of Erewhonian evaluative thought or talk in expressivist terms. Or at least it does so, on an assumption I am happy to make, unlike perhaps M&P. This is the assumption that if we can take Erewhonian evaluative and ethical talk to be action-guiding on the one side, truth-conditional on the other, then we should do so; we would have no reason to pay attention to a 'debunking' theory like expressivism that 'struggles to make sense of some salient phenomena' such as the truth-conditionality of evaluative utterances (BE 252–253).[8]

As the genealogy reduces the temptations to think of Erewhonian discourse in expressivist terms, so it has the same impact on error theory. In treating something as morally desirable, to put the idea in our terms, the property that the Erewhonians will focus on is that its having a property – maybe this, maybe that – that makes it robustly attractive on all sides; this came up in the response to Lerner. This property will reveal itself to them only from within 'the practices of avowal and pledging' (BE 261). But it will still be a bona fide feature in the natural world, not a fiction or illusion. By the account I defend, a set of items will be unified by a bona fide feature or property – it will contrast with a random set – insofar as there is a naturalistic basis on which it is possible to determine other members of the set from a proper subset of members: insofar as members of the set display a naturalistically grounded pattern (BE 255–257). And this will be possible for the humanoids if they are enabled, as under the genealogy they would be, to extrapolate from certain instances of the robustly attractive, personal or multilateral, to others. Those features may be of species-specific interest only – they may be like colors in that respect – but they will be none the less real for that.

Turning to a final point, the genealogy makes sense, not just of why we should not understand Erewhonian thought and speech in expressivist or error-theoretic terms, but also of why we should reject a non-naturalistic account. Here the main consideration is parsimony. On my rendering of the genealogy, it offers a plausible story of how an Erewhonian evaluative concept can designate a naturalistically intelligible, bonafide property. And so we have no good reason to postulate a non-naturalistic property – say, a causally inert Moorean property (M&P) – as the entity that is designated in some more proper way. 'Why multiply entities by positing that property, when by hypothesis the concept makes perfectly good sense in naturalistic terms?' (BE 24).[9]

Some other issues

Since my response has been focused so far on the main line of critique run by M&P, it may be useful to add some comments on some other points where they raise questions or make complaints. One complaint that they make, in their words, is that 'we cannot do what Pettit seems to want to do in the quoted passage, which is to infer naturalistic purport for a discursive practice from the fact of a naturalistic description of that discursive practice' (M&P). Following the criticisms of others, they argue that just as a naturalistic explanation of theistic discourse does not support a belief that the deity it postulates makes naturalistic sense, so a naturalistic explanation of evaluative discourse among the Erewhonians does not support the claim that values are naturalistic in character. My reply to this point is that there is a big difference between the sort of explanation envisaged in the two cases. In the theistic case imagined, the explanation will presumably invoke naturalistically intelligible motives why people should believe in error that there is a god. In the evaluative case, the explanation takes a different, non-debunking form: it offers an account of how properties accessible to them within certain practices would pass, by our own understanding of that concept, as evaluative in character.

Another interesting point M&P raise bears on the relationship between the genealogical story about evaluative discourse,

taken to apply in our world as well as in Erewhon, and the functionalist account of such discourse that I have defended with Frank Jackson (Jackson and Pettit 1995; 2023). That account, they say, allows us to 'apply that theory of content determination directly to our ethical concepts', suggesting as I read them that I might have done this within the genealogy (M&P). This raises an interesting question about the relationship between the two approaches.

According to moral functionalism, the way we human beings operate in identifying exemplars of certain properties (e.g. fairness, virtue, rightness), in linking those properties together, and in taking their presence to support one or another action makes it the case that they count as moral properties. As naturalists, we take the actual features that satisfy these functional role requirements to be natural properties, of course, but that is not strictly required by the functionalism. All that the functionalist theory does is to lay down role-constraints that any candidate realizers should satisfy.

This being so, it should be clear that moral functionalism does not determine the content of any evaluative claim in the sense of enabling someone who understands the theory to have a non-parasitic grasp of that content. It determines the contents of such claims in the sense of constraining them, not in the sense of offering the sort of paraphrase that might make it possible for an outsider to gain the understanding of a native (Jackson and Pettit 1995, 33–38).

On this reading of moral functionalism, it registers in useful detail the prompts and purposes, to recall an earlier phrase, that are relevant to the use of moral concepts. The genealogy is wholly consistent with the functionalist account of such constraints, although it works with a rather sparser characterization of those constraints (BE 270–277). What does the genealogy offer, however, that the functionalist account does not? It provides a naturalistic explanation of why the Erewhonians would give a referential use to concepts answering to the functionalist constraints and, if the actual mirrors the counterfactual, why human beings do so as well. It is one thing to show as functionalism does that using moral concepts does not commit us to non-naturalism, or even to show that naturalism makes good sense of why those concepts work; it is quite another to explain in naturalistic terms why the Erewhonians – or why we human beings – might be attracted to employing such concepts, revealing a goal that the concepts would serve.

M&P raise a third point of interest when they ask: 'why select these precise assumptions about Erewhon's starting conditions? In other words, why are these starting conditions, and not others, fit to illuminate our own ethical practices?' The question surprises me, for as I argue in the book, it is only to the extent that the Erewhonians are like us in a variety of ways that we can expect how they come to think and speak evaluatively to have a potential lesson about our own kind.[10] Had we postulated that as fortune would have it, they came to evaluative discourse under decidedly counter-intuitive pressures and opportunities, then we would hardly have grounds for thinking that something similar must be true of us (BE 33).

What M&P may have in mind, of course, is this thought: that even if there is no objection as such to the sort of genealogical thinking I advocate, still it will be useful only if the sort of genealogy I offer does not have any live competitors. A live competitor would have to satisfy two constraints. It would have to be able to take the Erewhonians to be like us in ways other than those I emphasize. And it would have to be able to show that a different set of evaluative and reprobative practices and concepts – say, one supporting an expressivist view – would be robustly likely to emerge among such counterparts. The existence of such a live competitor would support a serious challenge to the line I take. But it is not clear to me that there are any plausible candidates for that role.

Recasting the genealogy as conceptual ethics

M&P strike a more congenial note, at least to my ears, when they suggest in the final section of their paper that the genealogy might be understood as an exercise in the 'conceptual ethics' of the discourse of desirability and responsibility, rather than 'a descriptive or interpretive project' of identifying what our actual discourse involves. This exercise would consist in exploring how far it might be useful in some way to give up our actual ways of speaking and thinking in favor of a mode of discourse like that which is ascribed to the Erewhonians. As they say, 'the counterfactual genealogy might help us to see how certain sorts of "desirability" and "responsibility" concepts and practices distinctively function, in a way that allows us to see what is valuable in concepts and practices with those functions' (M&P).

The assumption behind this proposal is that there is a contrast between the descriptive or interpretive project of analyzing what is involved in our actual evaluative and reprobative thinking and speech and the revisionary project of identifying how such discourse would be better structured. Based on that assumption, M&P suggest that it might be worth investigating the promise of recasting my genealogy in that revisionary way. I do not share the assumption behind this suggestion but I am quite open, for different reasons, to something like the idea they have in mind.

In an article to which they refer, 'Analyzing Concepts and Allocating Referents', I argue that there are two aspects to philosophical theorizing on almost any topic, exemplifying the distinction in the theory of political freedom (Pettit 2020). A

first consists of identifying commonplace assumptions about the topic such that no theory that did not vindicate at least a bunch of these could claim to be a theory of that topic; it would change the subject. I claim that with almost any topic worthy of philosophical investigation, this first analytical exercise will leave a number of theories in the field: for example, theories of freedom that equate it respectively with non-frustration, non-interference and non-domination (Pettit 2014). The second aspect of philosophical theorizing will consist then in identifying among candidate entities in the domain of the discourse, that which is best taken to constitute the referent of the relevant concept or concepts; the best candidate will be that which best explains the significance of the topic in related areas of inquiry, displaying its claim on our philosophical attention. Thus, I argue for the equation of political freedom in any choice with the absence of domination rather than the absence of interference or frustration, on the grounds that this best explains the significance of freedom in political theory.

As the philosophical theory of freedom has these two aspects, so I think the philosophical theory of desirability and responsibility will also have those two aspects. But how might it display those two aspects if it is developed, as my theory is developed, on a genealogical basis? I argue in 'Analyzing Concepts and Allocating Referents' that such a theory will satisfy the analytical desideratum insofar as the counterfactual language and thought is recognizably of a kind with the targeted discourse: it is subject to the same prompts, and serve the same purposes, as our evaluative and reprobative judgments. And I suggest that insofar as the counterfactual discourse answers to the needs of people in the imagined society – and this, without violating naturalistic constraints – there is ground for thinking that the referents assigned to it in the genealogy are suitably significant (Pettit 2020, Pt 4).

Thus, taking Herbert Hart's genealogy of law as an example – a model for my own genealogy of ethics – I say in the article that it 'removes any mystery as to how we could get the concept of law going, it makes sense of the role of the concept without debunking it, and it directs us to a plausible property that constitutes its referent: this is what the concept serves to ascribe' (Pettit 2020, 350). As this is true of discourse about law, so it is likely to be true of evaluative and reprobative discourse, as indeed I note (Pettit 2020, 349).

Where does this leave me in relation to the suggestion about conceptual ethics that M&P make? In friendly territory, I think. I diverge from them in thinking that philosophical theory, whether or not it is developed genealogically, has an ineliminably normative aspect; it looks for an analysis of concepts and an allocation of referents – strictly, indeed, a combination of the two – that best satisfies relevant desiderata. Thus, I take the genealogy of ethics among the Erewhonians to identify a useful, and ideally the most useful, way of thinking about what they are doing – and, if the analogy is allowed, about what we too are doing – in evaluative and reprobative discourse. I can only hope that this construal of the enterprise will make it more appealing to M&P.

Response to Vargas

In his commentary on the book, Manuel Vargas addresses the general idea of a counterfactual genealogy, illustrating the different forms it may take, situating the form it takes in my hands, and raising some telling questions about the purposes it may serve. While he is moderately positive about the idea in general, and to some extent about the genealogy I try to provide, he makes a range of criticisms, some applying to any form of genealogy, some to mine in particular.

Forms of counterfactual genealogy

Vargas draws a distinction between more ambitious and less ambitious forms that a counterfactual genealogy may take, understanding genealogy on broadly the following congenial lines. All genealogies focus on familiar concepts that are problematic in some manner: paradigmatically, in ascribing properties that do not seem to fit with naturalism. A genealogy offers a counterfactual story about how creatures like us might develop practices that would make those properties and concepts available to them. If the story has a naturalistic character, it will suggest that among those creatures the concepts and properties available are themselves naturalistic. And so it will hold out the prospect that they are naturalistic among our kind as well.

The most prominent form that a less ambitious genealogy might take, according to Vargas, is that of providing a possibility proof. This is a proof that however we human beings actually come to target a problematic property and form a corresponding concept, all of that might have happened, as the counterfactual story illustrates, without breaching a constraint like naturalism. Such a possibility proof might describe a naturalistic process, however, that does not correspond to what likely happens in the actual world.

As an example of such a possibility proof, Vargas cites Michael Bratman's argument that it is intelligible in broadly individualistic terms how we can act together for joint ends. According to Bratman, we – or perhaps fictional counterparts – might act together in that manner on the basis of complex nested intentions that are individualistically intelligible, though

quite unrealistic; and because that is so, there is no inherent mystery about joint action (Pettit 2017). Another example might be Donald Davidson's(1984) account of how we or creatures like us might come to understand the meanings of a potentially infinite range of sentences by virtue of mastering a Tarskian truth theory for the language. Davidson does not suggest that that's how we do understand the meanings of our sentences; the idea is merely that that possibility renders such understanding intelligible in roughly 'finitistic' terms.

There are other unambitious forms that a genealogy may take in Vargas's view. It may serve just to stimulate our thinking about a topic, for example, or to suggest some novel ideas. Or, an 'especially interesting possibility' is that it might help us to identify 'a *new* target concept (whether this is understood to be a revised or replaced concept), or alternately, a new set of properties to which it would be rationally or practically preferrable to have as its referents' (V). This last possibility is indeed interesting and I discuss how it relates to my own efforts in the response to McPherson and Plunkett.

But what form will an ambitious genealogy take? 'The ambitious form of counterfactual genealogy', Vargas says, 'endeavors to be instructive about the nature of things in a particularly bold way'. Where the possibility proof ' might focus on providing a proof that some concept is compatible with naturalism', the ambitious form of the exercise 'attempts to show that the concept in fact predicates only naturalistically respectable properties' (V).

Vargas has a distinctive conception of what naturalism demands, and of what is required of an ambitious genealogy in this sense, as he makes clear in the later discussion of a particular example. This arises in connection with a sample genealogy that he sketches for our spatial concepts: this, in the spirit of Strawson (1959), though not on the same lines.

Underlining a very strict reading of what an ambitious genealogy should achieve in this case, he writes as follows. 'Even if a full genealogy of spatial concepts gets right all the details of our concept acquisition, and even if the forces that produced it and regulated it were all naturalistic in nature, this does not guarantee that what we predicate of <with?> those concepts will survive in the final accounting ' (V). Why so? Because we now know that there is no such thing as absolute simultaneity to what we would likely have assumed in defending the genealogy. The observation means that such a genealogy would not serve to vindicate naturalism, by his lights. It 'suggests that concepts can serve a naturalistic function while ascribing what, in the end, turn out to be extra-naturalistic properties (i.e. properties not predicated by final fundamental science)' (V).

This makes clear that for him, an ambitious, naturalistic genealogy would be successful only if it could identify the actual properties assigned in a target discourse with properties that are countenanced in the final, fundamental, natural science; only in that case, he assumes, will it 'teach us about the nature of things' (V). On this account, the aim is to show that there are naturalistic properties – fundamental, vindicated properties – that the genealogy presents as the properties predicated. On my account, by contrast, the aim is less ambitious: to show that there are naturalistic properties – maybe these, maybe those, maybe fundamental, maybe not, maybe vindicated, maybe not – that the discourse among the protagonists in the genealogy predicates.

As appears in the exchange with McPherson and Plunkett, a genealogy in my conception will explain how the counterfactual participants in relevant practices of commitment and responsibility would come to employ concepts in description of their world that serve evaluative and reprobative purposes. But it will not aspire to be able to identify the properties in naturalistic terms: to be able to offer naturalistic reductions or analyses of those properties. The aim is to show that the protagonists would think of the properties under the prompts, and to the purposes, of evaluative and reprobative discourse, yet do so in response to naturalistically intelligible inputs (BE 22). Consistently with a successful genealogy, then, the properties ascribed on this interpretation may be ascribed in error, although the genealogy I sketch for evaluative discourse is unlikely to support such an error theory, as I argue in my response to McPherson and Plunkett.[11]

The difference between Vargas and me on how to construe the ambition of a counterfactual genealogy is relevant, as we shall now see, to an assessment of the two constraints he puts on a successful genealogy, using them to raise questions about my own genealogy of ethics. One constraint is that the genealogy should be 'speculatively accurate', identifying 'conditions that are causally, functionally, or explanatorily relevant to the acquisition and current operation of our moral concepts'. And the other is that the genealogy – even if it satisfies speculative accuracy – should also be able to 'secure the naturalistic bona fides it seeks to secure' (V).

Speculative accuracy

Vargas holds that in constructing a counterfactual genealogy, 'the less confident we are about the account's speculative accuracy, the less confidence we should have in the genealogy's capacity to deliver a theory adequate to its aims'. He takes this guideline to require us to seek a counterfactual narrative in which the conditions that lead the protagonists to form their

concepts, and associated practices, should be as close as possible to the conditions relevant in the human case.

I agree with Vargas that if the Erewhonians are to be credited with evaluative and reprobative concepts, then they ought to be responsive to the cues that lead us humans to use such concepts and responsive also to the role that the concepts play among us, say in directing us to what we ought to do or to whether or not we are responsible for what we did. But I do not agree that the genesis of the concepts among the Erewhonians ought to resemble their genesis in our own case. And that is what speculative accuracy seems to require.

It might be in the actual case, for example, that our coming to use evaluative and reprobative concepts was due to a fluke: say, the appearance of an inspiring teacher early in the development of the species. But suppose it turns out, as a counterfactual genealogy might show, that even in the absence of such a teacher, agents of our ilk would have been 'more or less bound' (BE 34) to develop naturalistically intelligible practices of avowal and pledging and to gain access thereby to concepts like those of desirability and responsibility. Surely, that would be of enormous importance.

Plausibly, it would show that there must be a naturalistically role or function played by those practices and concepts that explains why their appearance in Erewhon – and by analogy in the actual world – should be more or less inevitable. And it would thereby indicate that however fluky the origin of our actual concepts in the genius of a primeval teacher, those concepts are nonetheless likely to have the same function as in Erewhon, and to have caught on and stabilized for that reason. The suggestion would be that even if the teacher had not invented and popularized those concepts, still they would have been more or less bound to appear anyhow. They would have been more or less bound to appear in the sense, borne out in Erewhon, that their appearance would have been likely independently of any fluke like the presence of an inspired teacher: its likelihood would have been robust over variations in the presence or absence of such a fortuitous contingency.

Vargas seems hostile, as I am, to giving flukes an important role. But he understands the no-flukes constraint quite differently from me. He takes it to require that if we are to construct an insightful genealogy of any natural development, then that development in the actual world had better not be the result of a fluke. Thus he notes that 'evolution – whether biological or cultural – is fluky' (V), suggesting that that is going to be a problem for the genealogy of any evolved feature, given that speculative accuracy requires it to 'capture the causal, functional, or otherwise explanatory features that matter for the actual concept' (V).

The example of the inspired teacher shows that I reject that view. I hold that a genealogy can help to identify a naturalistically intelligible function for something that appears in the actual world as the result of a fluke: something like the influence of the imagined teacher, or the development of a benign mutation, biological or cultural. What I require, as Vargas recognizes at other points, is not that the actual genesis of a target be non-fluky, but that its genesis in the genealogy itself should not be fluky: that in that respect the genealogy should be very different from a just-so story (BE 38).[12]

Ethics and money

The disagreement between Vargas and me about the need for speculative accuracy, as he understands it, comes out nicely in the different views we take of an analog to my proposed story about ethics: the economist's counterfactual genealogy of money. That genealogy argues that in the absence of money creatures like you and me – the protagonists in this narrative – might plausibly barter with one another; that that would lead to their prizing for its exchange value any commodity that was manifestly wanted by many; and that this would lead them to treat and think of that commodity, be it gold or tobacco or whatever, as we treat and think of money. Historical and anthropological studies, as I note in the book (BE 50), and as Vargas demonstrates at great length, show that this counterfactual genealogy does not satisfy speculative accuracy. But where he finds fault with the genealogy on that account, I do not.

We may both agree that 'money was not introduced to solve problems in a barter context': that it may have emerged, for example, 'as a way of tracking the antecedent phenomenon of credit and debt' (V). But I argue that the economist's genealogy may still be very useful, while he denies this.

If reliable, by my reckoning, the genealogy shows that even if credit and debt had not already materialized – even if the protagonists in the story, unlike our human ancestors, had been reduced to barter – still money would have been robustly likely to appear; that this would have been likely because of its role as a means of exchange and indeed a metric of price; and that actual money, whatever its origin, may serve that same function – surely a reason to think it a stable phenomenon – and have the same functional character.[13]

By contrast, Vargas thinks that going along with such a purely counterfactual genealogy has 'pernicious' effects, leading us to ignore the dependence of the economy on the state, for example, to think that human beings conform to the image of

homo economicus, and to assign a pre-institutional authority to the economy, taking it to be self-regulating. I reject the idea that the genealogy of money on its own forces us to adopt such views and have argued the point elsewhere, stressing the state-dependence of any feasible economy (Pettit 2023, Ch 6). It may be true, as Vargas comments, that it is 'harder to appreciate' such dependence 'if one accepts the myth of barter origins for money' (V). But to go along with the genealogy is not to believe that money actually had its origin in barter, of course, and in any case empirical claims about what it is hard to appreciate should hardly carry much weight in this context.

Two specific complaints

However differently we understand the role of the no-flukes constraint, and however far we differ on the requirement of speculative accuracy, there are two pertinent complaints about my genealogy of ethics that Vargas makes in discussing that requirement.

One complaint is that in the story about Erewhon I 'delete the fact of our interdependence, altruism, and cooperative impulses from the genealogy of morality' (V). To this I plead not-guilty, as I explicitly assume in the book that the Erewhonians 'are able to rely on others, and able to get others to rely on them', having the capacity 'to exercise joint attention', and 'to act jointly with one another in pursuit of shared goals' (BE 33). True, I assume that while they may be 'moderately altruistic, they primarily desire the promotion of their own welfare'. But if this is deemed excessively pessimistic, still it can be seen as positing a worse-case scenario and then arguing that even under that assumption, evaluative and reprobative concepts would play a useful role. That would seem to support the interest of the genealogy, as indeed the book suggests (pace Vargas, fn). 'What our nature would have generated in the dry wood of Erewhon' – that is, on the assumption of mainly self-interested protagonists – 'it is all the more likely to have generated in the green wood of our actual history' (BE 41).[14]

The second complaint that Vargas raises in relation to speculative accuracy is that I display a questionable anthropocentrism in assuming, first, that among animals language is specific to humans and their Erewhonian counterparts and, second, that other 'animals aren't the right place to look for morality as they lack moral concepts' (V). Here I must indeed plead guilty. I try to show that in creatures otherwise like us language and exchange are enough to make the appearance of ethics and ethical concepts robustly likely. But I would welcome any attempt to show that the nature we share with many other animals would be enough in itself to ensure that result. Certainly nothing I argue should be essentially tied – as admittedly ~~indeed~~ I come close to tying it (BE 13) – to a denial of that possibility.

Wandering predication

Turning to a final point, Vargas suggests that even if a genealogy satisfies speculative accuracy in his sense, still it may be unsatisfying, because of '*predicated drift*': that is, 'that what is predicated can change over time' (V). This is an interesting possibility to register and it is not one I address in the book. But I take it on board as something welcome rather than worrying.

Suppose a genealogy makes it robustly likely that in a suitable context, the protagonists would develop certain practices and evolve a concept that predicates a given property, even a naturalistically respectable property. What Vargas rightly notices is that it might still be the case that with that concept in place, and that property in view, the agents involved would proceed in a further stage to revise the concept and replace the predicated property by something distinct. How should we think of such a possibility?

We might take it to be a development that is robustly likely in the way in which the original development was likely. But in that case, presumably, the genealogy should extend to include the development. The standard genealogy of money does that indeed when it moves from the stage where a money-like commodity has been recognized in a barter society to register a further, robustly likely development: that the fiscal authorities, local or otherwise, would agree to have taxes paid in that currency, and thereby give it an extra credibility and stability.

What, however, if we take the later development not to have been robustly likely in that manner?, What if we recognize that while the first development was robustly likely, the second was not: that it was driven for good or ill by a contingent, unpredictable factor? This would direct us to a use of the genealogy in supplementing actual historiography of a kind that Bernard Williams (2002) recognized, when he argued that it is often useful to put an abstract genealogy to such a concrete use. But the important observation from our viewpoint is that employing a genealogy as part of such a project would not argue against the interest or utility of genealogy as such. To the contrary, it would seem to reveal another appealing possibility that the genealogical approach opens up.

Concluding thoughts

Concluding thoughts

I leave the consideration of these incisive critiques of my book with a deep sense of appreciation. They provide me with a much better understanding of the goals that matter in the sort of genealogy charted and of the various dangers that stand in the way of achieving those goals. There is a follow-up book that should be going to the publisher within the next six months or so and I know that it will be much better for me having been forced by these commentators to recognize the shortcomings of my efforts in this.

In that new book, entitled *When Minds Converse*, I try to provide a genealogy that goes deeper and wider than *The Birth of Ethics*. It identifies six distinctively human capacities: the processing abilities to exercise intentional control in making judgments, in reasoning and in probing perception; and the relational abilities to exercise such control in making commitments, in ascribing responsibility and in interacting as persons. Beginning with humanoid creatures constructed in an image that would leave them within reach of a simple human language, it explores how far the opportunities and pressures created by a linguistic environment would prompt the emergence among humanoids of the six kinds of practices and skills that distinguish humans.

These critiques of *The Birth of Ethics* all directly or indirectly engage with the nature of the counterfactual genealogy employed there and in the book under preparation. It may be useful in conclusion, then, to comment on the benefits that such an approach might have in illuminating the capacity or skill, the phenomenon or institution, addressed. As I now see it, there are three possible benefits that a genealogy – or an analogous project like that of creature-construction, so-called – might hope to generate.

The first benefit is one of demystification. Suppose the target addressed is thought to be mysterious relative to some baseline assumption: say, that the world is naturalistic in the sense explained earlier; or that the attitudes and actions of groups are all ultimately grounded in the attitudes and actions of individuals; or that whatever human minds understand, they must do so in an exercise of finite capacities operating on finite materials. The first aim of any effort like that of a genealogy must be to show that the target does not necessarily breach that assumption.

The methodology will ensure this benefit by providing a possibility proof, as it is sometimes called: an argument showing that the target might in principle be realized without a breach of the baseline assumption. This is what Donald Davidson (1984) claims to do in showing how finite human minds might be able to grasp the contents of an indefinite range of assertions simply by mastering a Tarski-like truth theory for the relevant language. Michael Bratman (2014) claims in a similar vein to show how individual agents might act jointly for shared ends by each implementing sophisticated, coordinated plans for acting individually. Both thinkers claim to demystify the target they address but not to explain how semantic understanding actually materializes or how joint action actually emerges.

A genealogy of the kind I envisage will also serve the purpose of demystification if it shows how in principle a target might be realized without a breach of the relevant assumption. But a genealogy may also help to achieve a second benefit too: that of illuminating the role or function played in human life by the target – say, the institution – addressed. If the psychology and circumstances of the creatures invoked in the genealogy are suitably similar to ours, and if it is robustly probable that the institution would have emerged among those creatures without a breach of the assumption, then the genealogy may serve not just to demystify that target, but to identify an effect that would have made the emergence probable. And if the effect would have made the emergence robustly probable in the counterfactual world, then it is likely to direct us to an effect that the counterpart institution has among humans: an effect that constitutes a function of the institution, giving it a high degree of resilience in the actual world.

David Lewis's (1969) work on Convention is a salient example of a genealogy that does this. According to his narrative, creatures like us who lacked conventions would be faced with coordination predicaments, where each party wants to do whatever other parties do but cannot reliably tell what that is. Those creatures would be motivated and enabled to solve the problems in an unplanned way, so the story continues, by relying on precedents or parallels from elsewhere, or just a sense of what they each find salient. And such solutions would tend to aggregate into regularities or conventions that would make later coordination easy. This narrative not only demystifies conventions in more or less individualistic terms. It also suggests that the function of conventions in human life, as well as in the counterfactual model, is to resolve coordination predicaments.

The third benefit that a genealogy, or any similar exercise, might seek to realize is that of illuminating the origin of the target phenomenon in human life. This is perhaps the most difficult aim of all. Thus, even if Lewis's genealogy does a good job in individualistically demystifying conventions, and in identifying a function that they play in human life, it is hardly a plausible story about how conventions actually emerged among us. But that need not take from the other benefits that the

genealogy provides. As I suggested in response to Manuel Vargas, the genealogy of money commonly offered among economists certainly demystifies money in individualistic terms and surely directs us to a beneficial function that money plays, say in providing a means of exchange and a metric of price. But according to a range of historical and anthropological studies, it does not offer a reliable account of how money actually originated in human societies.

I suspect that the only approach in the general genealogical area that might help to shed light on the actual origin of a target phenomenon is a creature-construction account that would go beyond the demystifying aims illustrated by Davidson and Bratman. It would offer a story about how human beings could have come to generate a practice or a skill but suggest, not just that the story demystifies that target relative to some background assumption, but that it also offers a plausible account of how the practice actually appeared and came to be conceptualized among human beings. A good recent example of such an approach is Barbara Vetter's (2022) sketch of how we gain access to modal concepts. She suggests that we do so by starting from our sense as agents of what we can do and control, moving to the recognition of the affordances that things display when we see them as subject in various ways to our control, and generalizing from that concrete base to more abstract possibilities.[15]

Footnotes

1 This comment is made with reference to avowing a belief but later remarks show that they apply to the avowal of desire as well: see BE 108.

2 His use of the term is different from mine, as I take full-scale realism to involve just the rejection of the error-theoretic view that moral properties are 'illusory', where this rejection 'leaves room for recognizing that there may not be a fact of the matter corresponding to every issue in morality' (BE 324–325). I stick with his usage here.

3 When someone judges that an option or arrangement is desirable, to return to an earlier question – when they avow a desire on the basis of a presumption that it is robustly supported – they will take it to be desirable in the role-property sense: to have the higher-order property of being supported by a robust attractor or value.

4 In *When Minds Converse* (Oxford University Press, forthcoming), I give greater importance to this mode of self-commitment, arguing that once someone has learned to make commitments to others, and have commitments made to them, they will be able to make commitments to themselves, as for example in forming resolutions. Thus, I break with the Hobbesian claim that self-commitment is impossible (BE 249).

5 This range will be significant even among us human beings. In most forms of conversive address, after all, we purport not to be acting insincerely in a way that the rules of the practice would prohibit: not to be speaking carelessly or deceptively, for example, and not to be making false promises. And in such forms of address we also generally purport to be eschewing recourse to coercive forms of influence that the practice precludes: not to be ready, should we not persuade an addressee as we wish, to resort to force. See Pettit (2021).

6 My version of the claim, as presented here, is close to that which M&P later suggest I should consider.

7 Do I go astray if it is true, as M&P hold, that 'Pettit suggests identifying the property of ethical desirability with what he calls multilateral desirability'? I do not think so. Suppose, as the genealogy claims, that the Erewhonians would take some of those properties that they treat as robustly attractive to be robustly attractive on all sides, not just in their personal view. What I say is that they would then view such properties in a prescriptive light – this, as we can see, by simulating their position – that is distinctively multi-lateral and that viewing them in that way would be a counterpart of how we would view properties that we treat as morally significant.

8 Another such salient phenomenon, as I say, is the embeddability of an evaluative assertion in the antecedent of a conditional that Peter Geach (1972) famously emphasized. M&P misrepresent my approach somewhat in citing that feature as one of a number of 'familiar challenges' on which I exclusively rely for dismissing expressivism; the main challenge on which I rely is that the genealogy shows that recourse to expressivism is 'unnecessary' (BE 254).

9 The considerations in the preceding discussion might be taken to argue against expressivism, error theory and non-naturalism in our own human thought and talk. But clearly they have not always been found persuasive, so that there are still adherents of each; that is why the genealogy on offer may do useful work. The point of invoking the considerations here is merely to back up the claim in the previous discussion that we know that the Erewhonian language is not plausibly understood on expressivist, error-theoretic or non-naturalistic lines, given how it is constructed: how it comes into existence.

10 That the humanoids sufficiently like us in suitable ways is defended explicitly in the book (BE 40–45).

11 I am grateful to Vargas for forcing me to acknowledge the possibility of the error-theoretic result envisaged here; I

do not recognize it in the book.

12 Vargas does at one point seem to go along with that reading: see (V).

13 Two things are worth noting. One is that the similarities I assume in the psychology and situation of human beings on the one side, the protagonists in the narrative on the other, are consistent with a contingent difference in how money arises or might arise in either world. The other is another assumption I make: that while anthropology and history may teach us that money actually arose within a world of credit and debt, not a barter world, they do not show that that is the only way in which it could have arisen.

14 I assume that in 'the green wood', our forbears would have been perhaps more altruistic than the Erewhonians in 'the dry wood' but still in good measure self-regarding: for example, self-regarding enough to be concerned with their reputation among others. The point here is that if the Erewhonians had that extra degree of altruism, that would promote, not reduce, the likelihood that certain forms of evaluative and reprobative discourse would emerge among them.

15 I am greatly indebted to all the commentators for the attention they gave to my book, but I am indebted above all to Tristram McPherson and David Plunkett for organizing the symposium and for providing detailed and useful comments on a first draft of my Precis and Reply.

Disclosure statement

No potential conflict of interest was reported by the author(s **[Q5]**).

References

Bratman, M. 2014. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Davidson, D. 1984. *Inquiries into Truth & Interpretation*. Oxford: Oxford University Press.

Frankfurt, H. 1971. "Freedom of the Will and the Conept of a Person." *Journal of Philosophy* 68 (1): 5–20. https://doi.org/10.2307/2024717

Geach, P. 1972. *Logic Matters*. Berkeley, CA: California University Press.

Jackson, F., and P. Pettit. 1995. "Moral Functionalism and Moral Motivation." *Philosophical Quarterly* 45 (178): 20–40. reprinted in F. Jackson, P. Pettit and M. Smith, 2004, Mind, Morality and Explanation, Oxford, Oxford University Press. https://doi.org/10.2307/2219846

Jackson, F., and P. Pettit. 1996. "Moral Functionalism, Supervenience and Reductionism." *Philosophical Quarterly* 46 (182): 82–86. https://doi.org/10.2307/2956309

Jackson, F., and P. Pettit. 2023. "Moral Functionalism." In *Moral Realism*, edited by **D. Copp and P. Bloomfield**, 246–263. Oxford: Oxford University Press.

Lewis, D. 1969. *Convention*. Cambridge, MA: Harvard University Press.

Lewis, D. 1979. "Attitudes de dicto and de se." *Philosophical Review* 88 (4): 513–543. https://doi.org/10.2307/2184843

Perry, J. 1979. "The Essential Indexical." *Nous (Detroit, MICH)* 13 (1): 3–21. https://doi.org/10.2307/2214792

Pettit, P. 2008. *Made with Words: Hobbes on Language, Mind and Politics*. Princeton: Princeton University Press.

Pettit, P. 2014. *Just Freedom: A Moral Compass for a Complex World*. New York: W.W.Norton and Co.

Pettit, P. 2017. "Corporate Agency—The Lesson of the Discursive Dilemma." In *Routledge Companion to Collective Intentionality*, edited by **M. Jankovic and K. Ludwig**. London: Routledge.**[Q6]**

Pettit, P. 2020. "Analyzing Concepts and Allocating Referents." In *Conceptual Engineering and Conceptual Ethics*, edited by **A. Burgess, H. Cappelen, and D. Plunkett**, 333–357. Oxford: Oxford University Press.

Pettit, P. 2021. *A Conversive Theory of Respect*. Respect: Philosophical Essays. Oxford, Oxford University Press. (pp. 29–54).

Pettit, P. 2023. *The State*. Princeton: NJ, Princeton University Press.

Strawson, P. F. 1959. *Individuals*. London: Methuen.

Vetter, B. 2022. "An Agency-Based Epistemology of Modality." In *Epistemology of Modality and Philosophical Methodology*, edited by **A. J. Vaidya and D. Prelević**, 44–69. London: Routledge.

Williams, B. 2002. *Truth and Truthfulness*. Princeton, Princeton University Press.**[Q7]**