

Is Responsibility Essentially Impossible?

Susan Hurley

Philosophical Studies vol. 99 (2000), pp. 229-268.

A revised version appears in my Justice, Luck, and Knowledge (Harvard University Press 2003)

Abstract:

Part 1 reviews the general question of when elimination of an entity or property is warranted, as opposed to revision of our view of it. The connections of this issue with the distinction between context-driven and theory-driven accounts of reference and essence are probed. Context-driven accounts tend to be less hospitable to eliminativism than theory-driven accounts, but this tendency should not be overstated. However, since both types of account give essences explanatory depth, eliminativist claims associated with supposed impossible essences are problematic on both types of account.

Part 2 applies these considerations to responsibility in particular. The impossibility of regressive choice or control is explained. It is argued that this impossibility does not support the claim that no one is ever responsible on either context-driven or theory-driven accounts of 'responsibility'.

Part 1.

1. Elimination vs. Revision
2. Essence and Reference
3. Context, Theory, and Examples
4. Impossible Essences?

Part 2.

5. Elimination, Revision, and the Essence of Responsibility
6. Responsibility: Context and Theory
7. Meaning and Disagreement about Responsibility
8. Can the Essence of Responsibility be Discovered from Contexts of Application and Causal History?
9. Is Responsibility Essentially Whatever (if Anything) Occupies a Theoretical Role?

10. Context and Theory in Reflective Equilibrium

Part 1.

1. Elimination vs. revision

If people can be responsible for what they do, they can deserve praise or blame, reward or punishment, gratitude or resentment for what they do. The term 'responsible' is often applied to human actions with these implications, and has long been. People also hold various beliefs about the general conditions for an action to be responsible and its agent correspondingly deserving. A set of beliefs about the general conditions for responsibility can be regarded as a theory of responsibility. The actual contexts in which 'responsible' is and has been applied may or may not be correctly described by a widely held theory of responsibility.

Suppose contexts of application and theoretical description are mismatched. Does this show that the applications are wrong, because they don't satisfy the theory, or that the theory is wrong, because it doesn't describe the contexts of applications? Suppose a theory requires for responsible action that people control not just their actions but the causes of their actions, and their causes in turn, all the way back. Suppose this condition is not met in any of the contexts in which 'responsible' has ever been applied. Indeed, suppose it could not be met, because it is nonempirically impossible to satisfy. Does this show that all those applications were in error, that no one is ever responsible for anything? Or does it show that the theory is in error, that it misdescribes the conditions for responsible acts?

The structure of this issue about responsibility is parallel in some respects (though not all) to familiar issues about eliminativism in other areas of philosophy. Most familiar these days are issues about thought and internal syntactic causal structure. Terms such as 'thought' and 'belief' have been applied

in a variety of human contexts for a long time. According to one type of theory, such terms refer to cognitive states with certain internal causal structure. Suppose it turns out, as some connectionists argue, that cognitive states do not have the relevant internal causal structure, so that in fact nothing satisfies the theory. Does this show all those applications were in error, that there is no such thing as thought or belief? Or does it show that the theory is in error, that it misdescribes what thought and belief are?

The general issue is: when does a theory change revise or resolve our beliefs about some entity or property, and when does it eliminate that entity or property? That general issue is closely linked to another, the issue of what determines the necessary, or essential, properties of some entity or kind. Do the contexts in which a term has been applied and its causal history determine what is essential to its referent, or does some theoretical role assigned to it? The answer might differ for terms of different kinds. When essence is context-driven, we can be very wrong in our theoretical descriptions of a given entity or kind. We can discover surprising things about what is essential to that stuff we've been talking about. On the other hand, when essence is theory-driven, we can be very wrong in our applications of a term. We can discover to our surprise that nothing occupies the theoretical role essential to the entity or kind in question, that there is no such thing as what we took ourselves to be talking about. Entities or properties with context-driven essences are more hospitable to revision than to elimination. Eliminativism tends to assume theory-driven essences. What isn't clear in general is what counts as horse and what as cart. Are issues about eliminativism responsible to issues about essence, or vice versa?

These large questions are, to put it diplomatically, still unresolved. Do not imagine that the little tail of responsibility is going to wag this monster. The purpose of this article is relatively modest. It is to speculate on how issues about responsibility fit into the broader debate, and at the same time to engender some

skepticism about any position that takes responsibility to be essentially impossible while blithely ignoring this hornet's nest. In the first part, general issues about elimination and essence will be sketched with a broad brush, with no pretense at exhausting the possibilities. In the second part, connections with responsibility will be added to the picture.

But first: Is it inappropriate to treat issues about whether responsibility is impossible under the heading of eliminativism? Does this wrongly imply that responsibility is, if it exists, an entity? No. We can take the same breezy tack that Stich takes about eliminativism for beliefs and desires:

...if it rankles to talk about beliefs and desires as entities, we can construe eliminativism as claiming that there are no such things as *believers*, or *desirers* because predicates of the form '___believes that p' and '___desires that q' are never true of anyone. Viewed this way, what the eliminativist is claiming is that the extensions of these predicates are empty.¹

Similarly, what's at issue here is whether '___is responsible for x' is ever true of anyone, whether the extension of this predicate is empty.

Even back in the days when Stich was an eliminativist about thought, he admitted that the general issue between elimination and revision had no accepted answer. At one extreme, witches had been uncontroversially eliminated (as he then supposed) by better explanations for most of the events in which witches had been implicated. But merely showing that a theory in which a class of entities plays a role is inferior to a successor theory plainly is not sufficient to show that the entities do not exist. Often a more appropriate conclusion is that the rejected theory was seriously wrong about some of the properties of those entities and that the new theory gives us a better account of those very same entities. Ramsey, Stich and Garon point out that it would be a joke to suggest that Copernicus and

Galileo showed the planets Ptolemy spoke of did not exist. They go on to ask: But then why isn't phlogiston just oxygen, and caloric just kinetic energy? They write:

There is, in the philosophy of science literature, nothing that even comes close to a plausible and fully general account of when theory change sustains an eliminativist conclusion and when it does not. In the absence of a principled way of deciding when ontological elimination is in order, the best we can do is to look at the posits of the old theory--the ones that are at risk of elimination--and ask whether there is anything in the new theory that they might be identified with or reduced to. If the posits of the new theory strike us as deeply and fundamentally different from those of the old theory... then it will be plausible to conclude that the theory change has been a radical one, and that an eliminativist conclusion is in order. But since there is no easy measure of how "deeply and fundamentally different" a pair of posits are, the conclusion we reach is bound to be a judgment call.²

Since then Stich has probed the general methodological issue further and has changed his mind about eliminativism for thought several times; lately he is tending toward 'social constructionism'³. Old reference points are crumbling. We used to think eliminativism was on safe ground with witches, but now it is being argued that witches did and do exist but that many people have been mistaken about their properties.⁴ Could the 'social constructionist's' invocation of social and psychological considerations be appropriate to settle these issues in the case of responsibility? The rules of the elimination game aren't clear.

Dennett makes a judgment call in favor of eliminating qualia: the qualitative properties of a subject's mental states that are traditionally taken to be ineffable, intrinsic, private, and directly apprehensible in consciousness. In his view it is a tactical error to argue against the traditional position on qualia by

saying: "We theorists can handle those qualia you talk about just fine; we will show that you are just slightly in error about the nature of qualia". According to Dennett, far better, tactically, to say:: "What qualia?" Our pretheoretical notions of qualia are "so thoroughly confused" as to be "radically unlike" any acceptable account that might be salvaged from them.⁵

But notions of deep and fundamental difference, and radical unlikeness do not give us clear guidance on how to adjudicate between elimination and revision. What would be helpful would be a framework of issues and distinctions within which various candidates for elimination could be located and compared.

2. Essence and reference

First, some independent reading of the essential properties of the kind in question might be useful. If certain properties are essential to qualia, for example, but nothing in the world has those properties, then the eliminativist wins. If the impugned properties are not essential, then the revisionist may be right: we may just have learned something about qualia. But the revisionist needs to have enough of a referential grip on qualia, as opposed to something else, to enable him to deny *of qualia* that they have the impugned properties.

Certain familiar positions about reference and essence have implications about elimination vs. revision. For example:

1) It is necessary in virtue of what anyone must understand if they know what '*F*' means, that any *F* has *P*. *P* is an essential property of anything that would count as an *F*, in virtue of meaning. So if nothing has *P*, then there are no *F*s. If someone says "That *F* does not have *P*", he does not know what '*F*' means. It is not an open question whether *F* has *P*; no substantive disagreement

is possible on this question. Among other things, skepticism about the analytic/synthetic distinction weighs against this approach.⁶

2) '*F*' is defined by its theoretical role and refers to the kind, if there is one, that plays the role specified by theory *T*.⁷ But if the theory goes wrong in one small respect, this need not mean that there are no *F*s. If the theory includes various assumptions, what may be essential to kind *F* is just for most of them to be satisfied, in any one of a variety of possible ways. Or it may be that some of the theory's assumptions are agreed to be essential to the theoretical role in question, while others are inessential or are of unresolved essentiality.⁸ But if *T* is substantially and seriously wrong, then it could be discovered that there is nothing that occupies even the disjunctive or partial theoretical role that is essential to *F*.⁹ Since reference is on this view *theory-driven*, theoretically specified properties provide a referential grip that permits discoveries about nonexistence. A theory may give worldly contexts of use an important theoretical role, but in that case context is still responsible to theory.

3) '*F*''s reference and essence are determined by the worldly contexts of its original uses, or some other set of privileged uses, and its causal history.¹⁰ Such *context-driven* accounts make it harder for reference to fail than do theory-driven accounts. If there are appropriate explanatory connections between the privileged uses of the term and *something* in the contexts of those uses, and appropriate causal connections between those and other or subsequent uses, then it doesn't matter if extant theories about that something are wildly mistaken.¹¹ That *something* in contexts of use is not constrained to agree with anyone's preconceptions about it. We could think those sparkling things in the sky were the spirits of the great kings of the past. We could think we were talking about an intrinsic feature of some class of entities and discover we were in fact talking about a relational feature, or vice versa. Contexts of use plus the appropriate causal-historical relations between privileged and other uses provide the

referential grip that permits discoveries about theoretical properties. Such discoveries would reveal among others things aspects of the thing essential to it in virtue of their explanatoriness, in the way, for example, that H₂O-ishness is explanatory in relation to contexts in which 'water' is used. And here of course there is a role for theory, but it is theory responsible to worldly contexts. If nothing in contexts of use stands in appropriate explanatory relations to uses of the term, if users have just accidentally latched onto a coincidental conjunction of unrelated features, an arbitrary hodge-podge, then eliminativism may still win.¹²

4) Theorists of essence and reference disagree about the relative merits of theory-driven and context-driven approaches, both in general and for particular *F*s.¹³ This suggests a further possibility. Perhaps essential properties and referential grip emerge from reflective equilibrium between considerations based on original contexts of use and considerations based on theoretical role. When these conflict, perhaps neither trumps the other, but trade-offs must be made. There would then be scope for correction of both applications and theory as existence-preservation is traded-off against role-preservation. This reflective equilibrium approach is perhaps most familiar in discussions of normative terms such as 'justice', but it has wider application.

We can refer to these broad options for short as the meaning approach, the theory-driven approach, the context-driven approach, and the reflective equilibrium approach. This list of options is not intended to be exhaustive, and we shouldn't assume that one approach must be correct for all types of case. Many variants are possible within each approach, and there is middle ground: context-driven approaches give theory as role, and theory-driven approaches may give worldly context a role.

As indicated, the position we take on essence and reference for a given kind *F* may influence our view of the relative plausibility of revision and

elimination.¹⁴ We can call this the top-down approach to the general issue about eliminativism vs. revision. A worry about the top-down approach is that it gives us little real leverage on eliminativism because our position about essence and reference for a given *F* may reflect our prior intuitions about elimination vs. revision for *F* in particular cases. Could such intuitions indeed be part of the data to which theories of essence and reference are responsible?¹⁵

We could make a virtue of necessity and adopt a bottom-up approach to eliminativism. We have lots of examples to consider: thought, qualia, God, witches, phlogiston, triangles, SWFs, values, stars, persons. Can we sort these into three groups: uncontroversial cases of revision, uncontroversial cases of elimination, and controversial cases? Can we then analyze the groups and generalize inductively? Can we argue backwards for particular *F*s from *faits accomplis* about elimination or revision to the conditions that favor a particular account of reference and essence? Ideally, we'd like the top-down and bottom-up approaches to elimination vs. revision to mesh. If they don't, perhaps a bit of meta-reflective-equilibrating is needed. But it may also just be that the elimination vs. revision issue has no determinate resolution in many cases.

In the next section we will look a little further at the relation between the revise vs. eliminate issue and the context-driven vs. theory-driven issue, and consider how the two issues can cut across one another. Then we'll consider how impossible essences may fare. No pretense is made here at providing the missing general account of the conditions of elimination. There may not be one, or at least not one anything like what we are looking for (Stich's view in *Deconstructing the Mind* may be right). Our sights are on the question about responsibility in particular, to which we'll eventually turn. Even if the general issue about elimination is still wide open, the top-down and bottom-up approaches give it some structure. Given this structure, any features or constraints that apply in the case of responsibility in particular may yield some local progress. Moreover,

indeterminacy about eliminativism vs. revision itself undercuts claims of elimination.

3. Context , theory, and examples

Context-driven theories of reference and essence were developed in application to proper names and terms for observable natural kinds such as 'water', 'gold', 'tiger'. Because essence on such accounts is responsible to contexts of use rather than to the beliefs of users, users can make wildly false claims about, for example, stars, and still be talking about stars. Reference to an existent is more or less secured by context, and theory is up for grabs. Context-driven reference is very tolerant of descriptive and theoretical error--perhaps too tolerant, in some cases.

On context-driven accounts (among others), essences have explanatory depth. Such accounts require essential properties to do explanatory work and so give theory a role in discovering essences. But what needs explanation, what theory is responsible to, is determined by worldly contexts of use. Widely held descriptive beliefs about some kind may not reflect its essential properties because the theories these beliefs are embedded in fail to do the relevant explanatory work in relation to worldly contexts of use. Suppose we believe the Mufasa theory: that those twinkling things are the spirits of the great kings of the past. On a context-driven account, what needs to be explained by star-essence are those twinkling things, not just our beliefs. Our beliefs might be explained in some ways, for example, by a theory of religion or of social structure, that shed no explanatory light on those twinkling things themselves. So eliminativism about stars gets no support if it turns out those twinkling things aren't the spirits of the great kings of the past. Essential properties may be unknown for a long time and be empirically discovered, as it was discovered that having the composition H₂O

is explanatorily essential to that watery stuff. Essential properties may continue to be unknown, where theoretical dispute continues about what properties explain the behavior of the stuff referred to.

A context-driven account tends to protect application of a term at the expense of extant theory, and so to be attractive where revisionism seems intuitively correct, as for 'star'. However, because essences have explanatory depth on a context-driven, it is possible for some applications to be corrected. This could happen if an essential property that has the requisite explanatory role in relation to most contexts of use reveals certain applications to be erroneous: planets are not stars. But eliminativism postulates more than such limited error; it postulates global failure to refer of the term in question. Eliminativists may try to make their job easier by helping themselves to a theory-driven account instead.¹⁶

Indeed, by making it too easy to refer, a context-driven account threatens to deliver the wrong result where eliminativism seems intuitively correct, as for 'phlogiston', or 'witch'.¹⁷ But firm ground is hard to find here. Sometimes intuitions favoring elimination may be subject to correction, as has recently been argued for 'witch'.

We may assume that functional-kind terms, such as 'chair', demand a theory-driven rather than a context-driven treatment. But whether a kind is functional is itself something that people can disagree about¹⁸, and about which a theory might be mistaken. Context-driven accounts can even tolerate such categorial errors.¹⁹ Contexts of use might in principle be such that a term refers to a functional kind, even though users mistakenly think it does not. For example, suppose the theory that assigns to wizards the role of a kind of person with special magical powers, is a bad theory of the contexts in which 'wizard' has been used. It may eventually be realized, in part on the basis of empirical study, that a

better theory provides an explanation in terms of the functional role of wizards in society. On this better theory, wizards do exist but we have been wrong about their properties.²⁰

On the other hand, context-driven accounts don't make elimination impossible. Suppose the contexts in which 'wizard' has been applied are just an accidentally concatenated hodge-podge, a gruesome composite with no causal unity, functional or otherwise. No correction of local error on the planets-are-not-stars model, or relocation of the level of explanation, will do. Arbitrary hodge-podge could provide a case for deconstructing wizardkind and eliminating wizards even on a context-driven account of reference. If revision reflects discovered essences on such an account, then elimination of a purported kind would reflect discovery that nothing plays the explanatory role of an essence in relation to the relevant contexts. If the supposed kind has no essence, it is not a kind of any kind.²¹

Just as a context-driven account can support elimination, a theory-driven account can support revision. Consider triangles. People have been applying 'triangle' to objects in our world for a long time. Also, for a long time triangles were believed to satisfy two descriptions: to be bounded by three straight lines and to have interior angles that add up to 180 degrees. These two conditions are logically independent of one another; in some possible worlds one may be satisfied but not the other. Suppose it is discovered empirically (as it was in early 20th century physics) that in our world physical space-time is such that figures bounded by three straight lines have interior angles that add up to more than 180 degrees. Is this a discovery that physical triangles do not exist in our world, because having interior angles that sum to 180 degrees is essential to being a triangle? Or is it a discovery that people were wrong to suppose all triangles have this property, when in fact triangles in Euclidean worlds do and triangles in Riemannian worlds do not?

Intuition here favors revision rather than elimination for triangles. We have changed our view *about triangles*.²² Both a theory-driven account and a context-driven account can accommodate this intuition by allowing that one but not the other property is essential to being a triangle. Why is one property essential rather than the other? The two accounts ought to give different answers, if the distinction between them is not to be blurred. Both could begin by saying: because an improved theory revealed that one and not the other is true of those things we call 'triangles' in our world. After all, on context-driven accounts, essences are explanatory and theory plays a role in discovering them. But on such an account, in principle not just one but both theoretical properties could have been rejected.

By contrast, on a theory-driven account it is important to hold on to part of the specified theoretical role of the kind in question, even if another part is rejected on an empirical basis. Perhaps it is agreed that among all the properties a given theory assigns to kind *F*, some are essential to anything that would count as an *F* and others are inessential, even if the status of some intermediate group is unresolved or disagreed.²³ Or perhaps there is no independent agreement on some subset of properties as essential, and it is merely required that 'enough' of the theoretical role is retained to provide a referential grip. Any one of a variety of subsets could in principle provide the essence of the theoretical role in question. We can refer to these options for short as the *agreement account* and the *disjunctive account* of the essence of a theoretical role.

These accounts of the essence of a theoretical role don't seem to go to the heart of the matter. We want to know *why* people agree that some subset of properties specified by the theoretical role of kind *F* count as essential to *F*s. After all, when elimination is at issue, the putatively essential property has been revealed not to apply to anything, so the essential aspect of the theory is false.

Whence then the eliminativist's commitment to the essentiality of the discredited property? Similarly, we want to know *why*, among the various subset-disjuncts that could in principle have provided the essence of the theoretical role of kind *F*, one rather than another turns out to do so. Probing further, there are two possibilities (among others). The selection of subset could be determined by what anyone must understand who knows what '*F*' means. Or, the selection of subset could reflect the greater explanatory depth of certain elements of the theory in relation to others. We can refer to these deeper accounts of the essence of a theoretical role as the *meaning account* and the *explanatory depth account*.

Consider the last possibility further. Context-driven essences have explanatory depth, as we saw. There is a way in which theory-driven essences can also reflect explanatory depth, but the explanatory framework is somewhat different. Explanatory depth within a theory-driven account would relate to the theory itself. It would have a coherentist character. A subset of the properties the theoretical role assigns to the kind *F* may do better than any other subset at preserving the internal coherence and point of the theory. Such explanatory depth has a theory-internal normative and justificatory dimension.

For example, suppose we have a theory about just distributions that gives a role to both equality of welfare and to equality of resources.²⁴ Now suppose for the sake of argument that nothing could count as a discovery that justice does not require equality at all, because in some form or other equality is essential to the theoretical role that justice plays. Nevertheless, egalitarians disagree about what is essential to egalitarian justice. Such disagreement would be intelligible as a disagreement about explanatory depth and coherence internal to egalitarian theorizing. Suppose that a luck-neutralizing account of what it is we are concerned to equalize can deepen our understanding of the theoretical role of equality, and of the reasons that equality of welfare and equality of resources have the attractions they have. Such an explanatorily deep account could reveal,

for example, that equality of resources, properly understood, is essential to justice, if and when it comes apart from equality of welfare.²⁵ In this way a theory-driven account of justice could support the revision of certain welfare-theoretical egalitarian claims.

Consider a different example. A theory-driven account of 'qualia' might describe the qualitative properties of mental states as ineffable, intrinsic, private and directly accessible to consciousness.²⁶ If no mental states have *any* of these features, a theory-driven account supports eliminativism about qualia. But if these features are independent enough for it to make sense to suppose mental states might have some of them but not others, then what? Might only a subset of these features count as essential to qualia, so that we could revise our view of qualia rather than eliminate qualia? Suppose a new theory explains how relational rather than intrinsic properties can satisfy the other descriptions of qualia, and indeed explains how a relational view of qualia actually makes for a more coherent overall account of qualia. Then if the properties we take for qualia turn out to satisfy these other descriptions but are not intrinsic, we may just have been wrong to suppose qualia were intrinsic. In this case, the distilling of qualia-essence from the broader theoretical role reflects the explanatory depth of the essence from a perspective internal to the theoretical role.

What if meaning rather than explanatory depth determines which aspects of a theoretical role are essential? This option does seem to have one undesirable consequence. If some property *P* of kind *F* is essential in virtue of meaning, then someone who knows what '*F*' means cannot disagree about whether some or all *F*s have *P*. She can of course disagree about whether anything has *P* and about whether any *F*s exist at all. Meaning as a source of essence pushes dissent in the direction of elimination rather than revision. By contrast, if *P* is essential in virtue of its theory-internal explanatory depth, someone could disagree about whether

some or all *F*s have *P*. Within theory-driven accounts, explanatory depth as a source of essence is more accommodating to revision than meaning is.

So, theory-driven accounts can also go either way on revision vs. elimination.

Our intuitions about revision vs. elimination are often murky or conflicting: consider God, thought, persons, values. But even when intuitions about revision vs. elimination are clear, are they among the data to which theories of essence and reference for the corresponding term are responsible? Can we allocate terms to theories of reference on this basis, among others? We can now respond: things are not that simple; the connections are not that tight. Each account has a tendency but also has tolerances.

To sum up this section: First, we can see why context-driven accounts *tend* to support revision rather than elimination, and why eliminativism sits more easily with a theory-driven account. If contexts of application anchor reference, there is greater freedom of movement with respect to theory; if theoretical role anchors reference, there is greater freedom of movement with respect to applications. But these tendencies should not be overstated. In principle the distinction between context-driven and theory-driven accounts of reference and essence cuts across the distinction between revision and elimination as a response to theory change. Second, on some versions of a theory-driven account as well as on a context-driven account, essences are essential in virtue of explanatory depth of one sort or another.

4. Impossible essences?

Often eliminativism takes the form of claiming that some property P that is essential to kind F is uninstantiated, so that there are no F s. We've been considering how claims of that form look, and compare to their revisionist rivals, from the perspective of different approaches to essence and reference: context-driven accounts vs. theory-driven accounts. But eliminativism sometimes makes the stronger type of claim that some property P that is essential to kind F is not just uninstantiated as an empirical matter, but impossible, as in "...true self-determination is both necessary for freedom and logically impossible".²⁷ P may be a conjunction of properties which are internally inconsistent. Call this *impossible-essence eliminativism*. How does it look, and compare to its revisionist rival, from these different perspectives?

Notice that in asking this question we are approaching our target, which is impossible-essence eliminativism about responsibility. But before we get there, some general considerations are relevant. In this section the intuition is explored that there is something weird about the very idea of an impossible essence.

We've seen that on a context-driven view, essences have an explanatory role. They may be discovered a posteriori. Having the composition H₂O is essential to water because this property explains the behavior of the relevant stuff in the contexts in which 'water' is or was applied. In virtue of its explanatory depth, an essence can show up local errors in a term's application as well as globally mistaken theories. And in the arbitrary hodge-podge, gruesome concatenation scenario, where there is no unified explanation of the relevant contexts at any level, a kind may be eliminated. But this isn't because the purported kind does have an impossible essence. It simply has no essence.

Could impossible properties be essential to a kind on a context-driven account? No, because *properties in virtue of which things are impossible can't do the relevant explanatory work*. We could not come to realize that an impossible

property explains what is going on in worldly contexts in which '*F*' is used.²⁸ Perhaps one property explains the behavior of certain stuff in the context of *these* uses and another property explains the behavior of other stuff in the context of *those* uses, and it is impossible for anything to have both properties. This could be a case of local error in application: one set of uses is mistaken, these are really planets, not stars. Or, it could be a case like jade, in which the level of explanation is relocated: this is the jadeite kind of jade, and that is the nephrite kind of jade, and they have superficial characteristics and functions in common.²⁹ Or, it could be a case of arbitrary hodge-podge or misconceived composite: perhaps one property is relational and functional, and the other is intrinsic and microstructural, and it's just a confusion to suppose the different sorts of context have anything with even fairly shallow explanatory depth in common. But none of these are cases of impossible essence.

Now essences discovered a posteriori from contexts of application are not going to be much use to the eliminativist even if they are not impossible. A property that could have applied but in fact turns out not apply to anything is hardly going to explain anything that happens in the relevant contexts. If wizards turn out not to have magical powers, then those powers can't be essential to wizard-kind in virtue of explaining what wizards get up to in the contexts in which 'wizard' is applied. Elimination on a context-driven approach turns on lack of essence--the arbitrary hodge-podge scenario.

Still, there is something further wrong with the idea of an impossible essence, if essences are supposed to have explanatory depth. When *F*s do not have *P*, then *P* is not the essence of *F*-kind. If wizards do not in fact have magical powers, then magical powers are not the essence of wizard-kind. But if it is not impossible for wizards to have had magical powers, then we can imagine a different world in which magical powers might have been explanatory. It would have to be different in lots of ways, a world of different kinds and different

essences. Perhaps there are no wizards, only sorcerers. But having magical powers might have been essential to sorcerer-kind, in virtue of the explanatoriness of such powers in another possible world. If things *were* different in lots of ways, including that stuff superficially similar to *F*--call it *G*--had *P*, then *G* *could* behave in various ways in various contexts that would be explained by its having *P*. In virtue of the intelligibility of this counterfactual, *P* is a possible essence.

But if *P* is an impossible property, no similar counterfactual is clearly intelligible. If things *were* different so that some stuff has an impossible property, *could* that impossible property have explained that stuff's behavior? There is a general difficulty about knowing what would be the case if something impossible were true. In particular, how could there be possible worlds in which impossible properties are explanatory? It seems not just false that impossible properties have explanatory depth, not just false that they are essences. They couldn't be. If it is essential to essences to have explanatory depth, we can eliminate impossible essences. When *P* is impossible, it is not even a possible essence. This point does not turn on whether we know that some property is impossible. We may be wrong about this, hence wrong about whether it can do explanatory work.

So, a context-driven approach does not bode well for impossible-essence eliminativism. How do impossible essences fare if we shift from a context-driven perspective to a theory-driven perspective?

They at least get a run for their money if we adopt a meaning account of how the essence of the theoretical role is determined. Since we may not realize that some combination of properties *P* and *Q* is impossible, we may believe that they are both essential to kind *F* on the basis of the meaning of '*F*'. Does it follow that they are? If someone comes to realize that they are not compossible

and denies that *F*s have *P*, does that show he doesn't know the meaning of '*F*'? If so, then impossible-essence eliminativism is vindicated.

However, this view shares the problems that afflict the simpler claim that it is necessary in virtue of meaning that *F*s have *P*. Stich treats similar moves on behalf of eliminativism about thought as unpromising: they are open to worries about the analytic/synthetic distinction and about how to allow for substantive disagreement over essential properties, such as disagreement over the properties needed for genuine thought.³⁰

Stich's worries about eliminativism based on properties essential in virtue of meaning apply whether or not the properties are impossible. There is a further worry when they are impossible. In the face of intuitive disagreement about whether a certain property is essential to *F*s or not, to resolve that disagreement in favor of the impossible essence is uncharitable. If intuitions diverge about what is necessary in virtue of meaning, it seems perverse to interpret what we mean to involve incoherence. Within the 'charity is not optional' methodology, we can trade off oddness of meaning or desire against truth of belief. Perhaps in some cases an interpretation that assigns an odd meaning but a true belief is more charitable overall than one that assigns a familiar meaning but a false belief. But the impossible-essence eliminativist is not in this position. He wants to pull off a kind of interpretative double-whammy: to claim it is both nonempirically impossible and necessary for *F*s to be *P*. The methodology of charity sees these claims as at least in tension with one another. That methodology is itself controversial and not neutral ground, but it provides a further worry about meaning-based impossible essences.

Consider an example. Arrow proved that it is impossible for a method of aggregating individual preferences, a social welfare function (SWF), to meet certain conditions.³¹ Considered individually, his four conditions have certain

attractions as conditions of rational social choice. But they are controversial, some more than others. And they are collectively inconsistent.

It is plausible that *social welfare function* is a theory-driven kind. Consider the claim that it is essential in virtue of meaning that a SWF has these collectively inconsistent properties, and therefore there is no such thing as a social welfare function. This version of impossible-essence eliminativism suffers from the problems noted above for the meaning approach to theory-driven essences.

What if we shift to an explanatory depth approach to theory-driven essences? On this approach, a subset of the properties the theoretical role assigns to the kind *F* may do better than any other subset at preserving the internal coherence and point of the theory. Such a subset has a normative status within the theory, as in the earlier supposition that a luck-neutralizing account provides the best account of the deep theoretical structure of distributive justice. Certain aspects of the theoretical role of SWFs might be essential in virtue of their explanatory depth in relation to the various *prima facie* attractive theoretical beliefs about SWFs.

But if specified aspects turn out to be inconsistent, as Arrow's four conditions do, they face the difficulties impossible properties have in playing explanatory roles, even coherentist, theory-internal explanatory roles. Each condition in an inconsistent set, considered by itself, may be attractive because it explains some of our beliefs about how preferences should be aggregated better than its rival conditions outside that set. But taken together they cannot be explanatory or add up to a coherent account of SWFs, since they are inconsistent. The explanatory depth account of essence is again in tension with impossible-essence eliminativism. This point applies whether the explanatory depth required is internal to a theoretical role or rather relates to worldly contexts of a term's use, as in a context-driven approach.

Still, a demonstration of impossibility internal to a theoretical role can play a part in eliminativism, not via the claim that the impossible property is essential to the theoretical role but rather via the unsalvageable shambles it makes of the theoretical role. There is a large literature debating whether one or another of Arrow's conditions should be rejected or refined, and vigorous disagreement. Compare two eventualities (there may be others).

On one scenario, a coherent account of the deep point and structure of social choice theory bypasses Arrow's result. Perhaps one of Arrow's conditions can be decisively criticized as less attractive than the others, or perhaps its apparent attractions can be approximated by other conditions that don't yield the impossibility. That would suggest that the other conditions have greater explanatory depth and would support revision of our descriptions of SWFs. This suggestion would be even stronger if, as is not the case in Arrow's result, the impossibility were internal to one or two assumptions, so that blame could be pinned down and the rest of the theoretical role in question exonerated.

On the other scenario, there is no good account of the deep structure of social choice theory that preserves the point and coherence of SWFs. No part of the theoretical role of SWFs has explanatory depth, and the right response to the impossibility is that whole idea of aggregating individual preferences subject to certain conditions is misconceived, and should be replaced by, say, a theory of deliberative democracy that gives no role to SWFs. Then the impossibility would contribute to eliminativism about SWFs. But it would be misleading to express this contribution by claiming that it is both essential to SWFs and impossible that Arrow's conditions be met. Rather, on this scenario, SWFs have no deep explanatory role and in this sense no essence. The eliminativism that emerges here is based on paradigm change, not on an impossible essence.³²

Neither the meaning account nor the explanatory depth account of what is essential to a theoretical role support impossible-essence eliminativism. We saw earlier that the context-driven approach was not hospitable to impossible essences either. An exhaustive, knock-down case against impossible essences has not been made. The point has just been to engender some skepticism about impossible essences, by reference to various familiar positions. It should not simply be taken for granted, assumed without argument, that an impossible property can be essential. Versions of eliminativism that depend on such an assumption are problematic.

Part 2

5. Elimination, revision, and the essence of responsibility

In light of these general considerations, let's now consider eliminativism about responsibility. To fix ideas, we will concentrate on an argument to the effect that no one is ever responsible for anything because responsibility essentially requires regressive choice or control yet regressive choice or control is impossible.³³ A regression requirement is a formal or structural condition for responsibility, which can be combined with various substantive conditions, such as requirements of choice, control, hypothetical choice, or reason-responsiveness. A *regression requirement* says that to be responsible for something you must be responsible for its causes, and it applies recursively. So, for example, combined with a requirement of choice or control, regression requires that you choose or control not just what you are responsible for, but its causes, and their causes, and so on, all the way back. Eliminativism based on the impossibility of regressive choice or control will be our target. The counterargument concedes that regressive choice or control is impossible, but disputes that responsibility is essentially regressive.

To set the stage for this dispute, we will also consider why a regression condition is compatible with a hypothetical choice condition for responsibility, but incompatible with animism (to be explained shortly) and with a requirement of choice or control. And to fix ideas further, it will be assumed that both sides to the dispute reject animism and agree that responsibility essentially requires choice or control. Thus, they agree on the source of the impossibility in question: responsibility cannot satisfy both a condition of choice or control and also a regression condition. They disagree on what this shows: that no one is ever responsible for anything, or that responsibility is not regressive. We'll refer to these as the eliminativist and revisionist positions respectively (though since not everyone supposes responsibility must be regressive to begin with, the revision in question would not be universal).

The loyalty of eliminativists to a regression condition may be unshaken by the incoherence they take it to wreak upon the notion of responsibility.³⁴ But given that intuitions conflict about whether responsibility is regressive, why doesn't the impossibility of regressive responsibility count in favor of a nonregressive view of responsibility rather than in favor of eliminating responsibility?³⁵ This question has no *obvious* answer (if it has a determinate answer at all). Neither position should simply be taken for granted. The belief that responsibility is essentially regressive requires defense; it is not something we are simply stuck with, like it or not. On the other hand, we need to know what could make it the case that someone who denies that responsibility is regressive is talking *about responsibility* rather than just changing the subject.

What could give us enough of a grip on responsibility to enable us to deny that it is regressive, or to insist that anything that would count as responsibility must be regressive and hence impossible? We've noted above several general lines of reply, which appeal to meaning, to worldly contexts of use and causal history, to theoretical role, to reflective equilibrium. We need next to consider the

particular features of responsibility that might be relevant to these various approaches. Then we can put the two together, and see how the particular features of responsibility bear on eliminativism under these various approaches. Whether there is a determinate solution to the disagreement between the eliminativist and the revisionist about responsibility remains an open question.

6. Responsibility: context and theory

Here is a rough sketch of some features of the term 'responsible' that we may be able to relate to the general approaches canvassed above.

A) Contexts of use. For a long time now, people have been making applications of 'responsible' and holding one another responsible for their actions, deserving of praise or blame, reward or punishment, gratitude or resentment. Nowadays people are not generally regarded as responsible for what they do when it results from brain damage, brainwashing or hypnotism, or mental illness. There is disagreement over whether people are responsible for what they do when it results in part from serious deprivation. People generally are regarded as responsible for what they do when it results from personality traits or tendencies within a normal range and with normal causes, without attention to whether these traits are genetically influenced or otherwise unchosen or outside the person's control.

B) Theory. Many people believe that responsibility satisfies various general conditions or descriptions, some of which are more controversial than others. A set of such conditions could be regarded as collectively defining a theoretical role. For example, it is widely held that responsibility for something requires choice or control of it. It is also widely held, though more controversial, that to be responsible for something you must be responsible for its causes.

Together, these conditions define the role of regressive choice or control. Various other conditions are discussed in the literature, such as the could-have-done-otherwise condition, and various versions of a reason-responsiveness condition.³⁶ But the conditions of choice or control and of regression will be used here to raise issues about elimination vs. revision.

C) Changes in contexts of use and theory. Over time, both tendencies to apply 'responsibility', and beliefs about the general conditions of responsibility, have changed. Older applications of 'responsible' and 'deserves' in, say, feudal contexts, were probably less restrictive than contemporary applications, and corresponding beliefs about responsibility were less prone to recognize a regression requirement.

D) Intuitive conflict and disagreement. Even now, intuitions conflict and people disagree about whether certain general conditions hold or not. For example, intuitions today conflict about whether a regression requirement holds. Intuitions continuous with an older tendency may not recognize this requirement.

E) Ethical character of disagreement. Disagreement about conditions of responsibility is ethical disagreement. For example, people disagree ethically when they disagree about whether responsibility has to be regressive, and hence about whether people can genuinely deserve reward or punishment if they are not regressively responsible for the causes of what they do.

F) Impossibility internal to theory. Certain conditions of responsibility that have been believed to hold do not just fail to be satisfied jointly as an empirical matter. Rather, they are impossible to satisfy jointly. For example, it has been argued that regressive choice or control is logically impossible.³⁷

We'll now examine these features more closely. The aim is not to provide an account of responsibility, but to consider how these features might constrain eliminativism about responsibility based on the impossibility of regressive choice or control, in the light of the general issues reviewed in Part 1. Three features in particular will turn out to bear on the issue of elimination vs. revision for responsibility. They are: (D) intuitive conflict and disagreement about some of the conditions of responsibility; (E) the ethical character of this conflict and disagreement; and (F) the impossibility of jointly satisfying certain conditions. Under various of the semantic options considered above, these three features tend to support revision of certain beliefs about responsibility rather than the conclusion that responsibility is impossible.

What can we say about the theoretical role of responsibility, under heading (B)? First, what is the relationship of causal responsibility to responsibility in our desert-involving sense?³⁸ Causal responsibility might be held to be necessary for responsibility, though not sufficient. Some relationships between responsibility and causal responsibility may be compatible with a regression condition for responsibility, while others may not. Control or choice conditions of responsibility also have implications for the relationship between responsibility and causal responsibility.

Consider the claim that some form of causal responsibility is necessary for responsibility. Actual choice or control of some outcome may be ways in which people can be causally responsible for it, so may be ways of meeting this necessary condition. But *is* choice or control necessary for responsibility? In Scanlon's example of the guilt-ridden believer, the agent did not choose and does not control the religious beliefs he was brought up with, and their associated burden of guilt. Yet they are not plausibly regarded as matters of luck for him, because he would have chosen them if he had been able to and would not have chosen to be without them. Hypothetical or counterfactual choice of something

is not a way of being causally responsible for it.³⁹ The fact that you would choose something if you could, or would not choose to avoid it if you could, does not put you in an actual causal relationship with it. If people can be responsible because of their hypothetical choices, in the absence of actual choice or control, then causal responsibility may not be necessary for responsibility.

Holding that causal responsibility is necessary for responsibility of course does not entail that causal responsibility is sufficient for responsibility. The latter assumption could be called *animism*. It may be part of what has at times motivated people to punish the bearers of bad news, or part of what motivates children to fear and personify inanimate objects that hurt them. Animism is far less plausible than the claim that causal responsibility in some form is necessary for responsibility. A thermostat, or an infant with well-trained parents, might control effectively, but is not responsible for what it/she controls. An animal or a child or someone suffering from a mental illness might choose without being responsible for what he chooses. Attributions of responsibility should be restricted to persons who are mature and competent in ways the theory of responsibility tries to spell out.

Animism is also relevant to headings (A) and (C). Despite the implausibility of animism to most modern people, animistic attributions of responsibility and desert to inanimate objects and animals that are not persons are still made, accompanied by blame and resentment. This may be a natural or primitive tendency. It is very common in young children, but some adults seem also to be intuitive animists. It seems probable that more animistic attributions of responsibility by adults were made in the past than are now. But the tendency to animism is still natural enough that it needs to be guarded against. Children are encouraged to grow out of it, and it is regarded as a regrettably childish trait in adults.

Does responsibility for something require responsibility for its causes? We can consider the issue of whether responsibility is regressive under headings (D), (E), and (F).

First, heading (D). Among contemporary adults there is widespread (though not universal) agreement on some version of a choice or control condition as a necessary condition of responsibility, and also on the rejection of animism. But the regression condition is highly controversial and the object of intuitive conflict.⁴⁰

According to one natural set of intuitions, a person need not be responsible for being what he is in order to be responsible for choices that are determined by what he is. If we can understand “foundations” in causal terms, then Robert Nozick seems to reject the regression condition when he says that the foundations underlying desert don't themselves need to be deserved, all the way down.⁴¹ And Galen Strawson describes common intuitions about responsibility that are not committed to the regression condition when he writes:

Many people accept that they are, ultimately, entirely determined in all aspects of their character by their heredity and environment. But it follows from this that, whether the heredity-and-environment process that has shaped them is deterministic or not, they cannot themselves be truly or ultimately self-determining in any way. And yet they do not feel that their freedom is put in question by this--even though they naturally conceive of themselves as free in the ordinary, strong, true-responsibility-involving sense. ... This is a very common position.⁴²

But intuitions here conflict. For example, despite his recognition of the “very common position” just described, Strawson also views us as deeply committed to a regressive choice conception of responsibility, even though it

makes responsibility impossible.⁴³ On this view, a person's responsibility requires that she be self-determining, in a sense that makes her responsible for how and what she is. But this is impossible, because it requires the actual completion of an infinite regress of choices of principles of choice.⁴⁴

It may be objected that the kind of freedom this argument shows to be impossible is so obviously impossible that it is not even worth considering. To this the reply is simple: the kind of freedom that it is an argument against is just the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess. The idea that we possess such freedom is central to our lives.⁴⁵

Strawson also regards us as "stuck with" a natural sense of self that is "irremediably incompatible" with any deep acceptance of the idea that all we are and do is determined.⁴⁶ "...[W]hat one naturally takes oneself to be...is a truly self-determining agent of the impossible kind".⁴⁷

On to heading (E). The intuitive conflict over the regression condition just canvassed constitutes a deep normative disagreement, ethical and political in character. If Adam claims that people deserve the fruits of their talents even though they don't deserve their talents, and Karl denies this, Karl has not simply changed the subject. If this doesn't count as a substantive ethical and political disagreement, what does?

Finally, heading (F). Consider the pairwise consistency of three candidate characterizations of the theoretical role of responsibility: animism, a regression requirement, and a choice or control condition.

Animism is inconsistent with a regression requirement. If an object or animal is causally responsible for a significant effect, someone with animist

tendencies will be inclined to treat it as responsible, deserving of reward or punishment, gratitude or resentment. But an animist does not suppose such objects or animals are responsible for the causes of what he takes them to be responsible for: to be either responsible for themselves or for the stream of further causes leading from the significant effect beyond them back into the past. Causation itself is not regressive: to cause something does not require causing its causes in turn. For *X* to cause *Z* may involve *X*'s causing *Y*, where *Y* is among *Z*'s causes, but it cannot require *X* to cause *Z*'s causes all the way back to *D*, *C*, *B*, etc., which are among *X*'s own causes.⁴⁸ So if causal responsibility is sufficient for responsibility, then responsibility is not regressive either. Of course, since animism is false and causal responsibility is not sufficient for responsibility, this implication does little to resolve conflict about the regression condition.

Note that while animism is inconsistent with a regression condition, denying the regression condition does not support animism. This is important, because there may be a subliminal tendency to view the denial of the regression condition as a kind of remnant of animism: as something we moderns can no longer be comfortable with for the same reasons, but must put behind us, into our collective childhood. But there is plenty of scope for improving on animism, for restricting responsibility to persons and in other ways, without accepting a regression requirement.⁴⁹

A regression requirement is inconsistent with holding that *A*'s responsibility for *X* requires *A* to be causally responsible for *X* in certain ways, such as by actually controlling or choosing *X*. It does not make sense to suppose that *A* actually controls or chooses not just *X* but also to *X*'s causes, all the way back. It is not possible for causes of *X* that occurred before *A* existed to be the objects of *A*'s actual choice or control.⁵⁰ So if responsibility does require causal responsibility in the form of actual choice or control, responsibility cannot consistently be supposed to be regressive. Since it is far more plausible that

causal responsibility is necessary than that it is sufficient for responsibility, this implication may do somewhat more to resolve conflict about the regression condition.

By contrast, regressive hypothetical choice is not incoherent. For *A* hypothetically to choose the causes of *X*, all the way back, is not for *A* to stand in an actual relation of causal responsibility to them. It is merely for it to be true that *A* would have chosen them if he could have, or would not have chosen to avoid them. This per se is not problematic, so long as in some (nonactual) world *A* can choose the relevant causes (i.e. the antecedents of the relevant counterfactuals are not impossible). Hypothetical choice can in principle apply to the stream of events leading up to an action or to whatever it is (his character?) in virtue of which it is true that the agent would make certain choices if he could. If *A* can be responsible in virtue of hypothetical choice, then causal responsibility is not required for responsibility, and regressive responsibility is not incoherent. Notice that if hypothetical choice is the substantive condition, then no actual choice is required of the agent's dispositions to hypothetical choice. There is no inconsistency in a self-referential hypothetical choice of not only the primary item responsibility for which is at issue but also the disposition to make just this hypothetical choice plus all its causes.

However, there are at least two problems here. On the one hand, the very causal costlessness and indefinite extendibility of mere counterfactual choice may count against its being sufficient for responsibility. There are too many things that people would choose, or would not choose to avoid, if they could. Further constraint is needed to narrow the focus down to those things people are actually responsible for, and the most plausible candidates to provide such constraint involve causal relations such as control or choice. On the other hand, under certain assumptions the counterfactual supposition that *A* can make the relevant choices may be impossible, so that the truth of hypothetical choice claims cannot

be evaluated. These points may make the view that hypothetical choice is sufficient for responsibility less attractive than the view that some causal relation such as control or choice is necessary, so that regressive responsibility is impossible after all.

From here on we will concede for the sake of argument that regressive responsibility is impossible because responsibility requires choice or control and regressive choice or control is impossible. Our interest is in the implications of this impossibility for the issue about elimination vs. revision. How might the eliminativist defend the view that responsibility is essentially regressive?

7. Meaning and disagreement about responsibility

First, he might claim that it is necessary in virtue of meaning that responsibility be regressive. Stich's reasons for regarding a similar move for eliminativism about thought as unpromising apply equally here.⁵¹ This claim is open to worries about the analytic/synthetic distinction, and about how to accommodate substantive ethical disagreement over a regression requirement (feature D).

A further objection applies to the claim that it is essential to responsibility in virtue of meaning that it has a nonempirically impossible property. To resolve disagreement about the meaning of 'responsibility' in favor of incoherence would be less than charitable. In this respect eliminativism for responsibility has a harder job than eliminativism for thought. It would be less difficult to justify attributing a conceptual commitment to an internal causal structure such as a language of thought, since if there is no language of thought that would be an empirical truth. It is not incoherent to suppose that thought depends on an internal language of thought. But it is incoherent to suppose that a responsible

person actually controls or chooses causes of what he does that occur before he exists. If charity in interpretation is not optional, this counts against interpreting the concept of responsibility as essentially incoherent in virtue of meaning.

If regression is inconsistent with choice or control, why suppress a requirement of regression for the sake of charity rather than one of choice or control? Someone might argue that the best interpretation of responsibility makes regression, but not control or choice, essential in virtue of meaning. By switching to a hypothetical choice requirement, incoherence may be avoided, as we've seen. Charity in interpretation is not hostile to regression per se, only to incoherence. But this move is of no help to the eliminativist, since if responsibility essentially involves regressive hypothetical choice, it is not impossible.

The eliminativist may be tempted to immunize his position by arguing that someone who denies that responsibility must be regressive is simply using 'responsibility' in a different sense. But we don't need to depend on controversial arguments from charity or general worries about the analytic/synthetic distinction to see that this cannot be right. Here feature (E) is relevant. The disagreement between Adam and Karl over whether people can deserve the fruits of their talents, even if they don't deserve their talents, is an ethical disagreement. Even if Adam is wrong, he is disagreeing with Karl, not talking past him. If regressivity is essential to responsibility, this is a deep truth that is it possible to get wrong, not a truth that must be understood by anyone who understands the meaning of 'responsible'. The point applies in reverse: the revisionist claim should not claim that it is essential in virtue of meaning that responsibility not be regressive. We can talk about responsibility in some thin or flexible way that does not respect such conceptual limits. Ethical disagreement about whether responsibility must be regressive cannot be pre-empted by appeal to meaning.

8. Can the essence of responsibility be discovered from contexts of application and causal history?

Necessity in virtue of meaning does not hold out strong hope for showing that responsibility is essentially regressive. Could a context-driven approach go further to resolve the issue between the eliminativist and the revisionist: either to show that responsibility is indeed essentially regressive, or to show how the revisionist gets a referential grip on responsibility that permits her to deny that it is regressive?

Some elements in context-driven accounts of reference that may tempt the eliminativist. A posteriori necessities allow the application of even essential properties to be controversial. So the eliminativist may be tempted to claim that responsibility essentially requires regressive choice or control, even though there is substantive disagreement over regressivity (disagreement that, as we've seen, makes it difficult to defend regressivity as necessary in virtue of meaning). That is, the eliminativist may be tempted by a context-driven approach in order to respect feature (D).

Brief reflection should dispel this temptation. On a context-driven account, the essential properties of kinds have explanatory depth in relation to worldly contexts of a kind term's use. Now it may or may not be plausible to apply such a context-driven account to 'responsible', but it's worth spelling out why this approach won't help the eliminativist in any case. How could conditions of regressive choice or regressive control describe properties with explanatory depth in relation to contexts of the use of 'responsible'? The very impossibility of satisfying these conditions, feature (F), cuts against their title to context-driven essentiality. We couldn't come to realize, despite previous disagreement, that they do the relevant explanatory work. Impossible properties are not instantiated in any contexts of use and cannot do explanatory work, so cannot have the

required explanatory depth. It's not just that regressive control or choice *turn out* not to be involved and hence not essential; they're not even *possible* essences (see sect. 4 above). So the context-driven approach to reference does not give the eliminativist a way to defend the claim that regressivity is essential to responsibility despite substantive disagreement about it.

A context-driven approach does allow that some of our applications may be in error: planets are not stars; pyrite is not gold; falling stones, ravaging wolves, and some people with brain damage, are not responsible. But this is not enough for the eliminativist, who claims that all our positive applications of 'responsibility' are in error. What might do the work of elimination would be an argument not that something impossible is essential to responsibility, but rather that the contexts in which 'responsibility' is applied are an accidental, arbitrary hodge-podge, a gruesome concatenation. Such a deconstruction of responsibility would reveal that nothing has the right explanatory relationship to the various contexts in which the term is applied, not even at the cost of revealing some of those applications to be in error, or of relocating the level of explanation. Uses of the kind term have no unifying point, cotton on to no kind at all. Certain applications in certain limited contexts may have a point or an explanation, but it is quite unrelated to, or at a different level from, or utterly at odds with, the point or explanation of other applications.

Context-driven theories of reference are more obviously tempting to the revisionist. She may be tempted to secure her claim to be talking about responsibility rather than something else (when she disagrees about its regressivity) in terms of the immunity of context-driven reference to bad theory. The theory of regressive choice or control is a bad theory of responsibility, just as animism was a bad theory. Neither theory has explanatory depth, neither has an explanatory relationship to contexts of use that reveals what is essential to responsible acts. Just as animism fails to capture the point of responsibility by

licensing its attribution too widely, the regressive choice or control theory fails to capture the point of responsibility by restricting its attribution to the vanishing point.

A better theory would strike a middle ground. It would explain the point of our positive attributions of responsibility in a variety of contexts, in a way consistent with our exempting from responsibility not just the inanimate realm, but also people suffering from brain damage, acting under hypnotic suggestion, and so on. Perhaps some version of a reason-responsiveness condition combined with some version of a choice or control condition could do this work. Such a theory may not be easy to find, and we may not have found it yet. On a context-driven approach (by contrast with a theory-driven approach), the revisionist is not constrained to preserve elements of existing theories; the successful theory may be quite different. Nor is she constrained to respect all our applications of the term in question; we may be mistaken in a variety of cases, just as original users were mistaken if they made animistic applications. But the revisionist who invokes a context-driven account cannot allow that none of our positive applications of 'responsibility' hit their mark, since some such must anchor her claim to be talking about responsibility.

So, on a context-driven approach, the dispute between the revisionist and the eliminativist could take the following shape. The revisionist tries to find a theory, however novel, that accounts for the contexts in which 'responsible' is applied, even if not all of them, and that supports her claim that responsibility is not essentially regressive. The eliminativist tries to deconstruct responsibility by arguing that there is no such account. On a context-driven approach, the prospects are dim for arguing that responsibility essentially requires regressive choice or control and hence is impossible.

Eliminativists sometimes suggest the following argument. Only a regressive control or choice requirement can explain why people are not responsible in cases of brain damage, hypnotism, etc., But this requirement cannot be reconciled with recognizing responsibility in more ordinary cases, since it is impossible to satisfy. So no one is ever responsible.⁵²

This is not the deconstruction argument the eliminativist needs. It may not be intended as a context-driven argument. Nevertheless, from a context-driven perspective, there are several related things wrong with such an argument.

First, it gives explanatory priority to contexts in which the term in question does not apply, and tries to use those 'negative contexts' to correct all our positive applications. But this gets things backward. We don't start with an explanation of things not being stars, and use that to correct all our applications of 'star'. Rather, we first explain what stars are, and then use that explanation to explain further why some things are not stars. A context-driven account gives explanatory priority to contexts of positive use, though it can admit that some of these are mistaken.

Second, there is the problem with attributing explanatory depth to impossible properties. Now it may be possible to explain why some things are not *F* by reference to their failure to meet a requirement on *F*s, namely, that *F*s have property *P*. Pyrite isn't gold, and behaves differently from gold, because it turns out not to have the essential properties of gold. But such negative explanation, of non-*F*-hood by non-*P*-hood, is parasitic on the explanation of *F*-hood by *P*-hood in other cases. The essential properties of gold do have explanatory depth in relation to instances of gold. Where *P* is impossible, *F*-hood cannot ever be explained by *P*-hood. So parasitic negative explanation is not available. On a context-driven approach, it is an illusion to suppose that a condition that is impossible to satisfy can explain why people are not responsible.

Third, even if we don't assume that contexts of positive use have priority, the eliminativist seems to be assuming the opposite. Why should we correct our ordinary positive applications to bring them into line with our exemptions of people from responsibility, rather than the other way round? The regression condition doesn't provide any independent leverage here, since it is at least as controversial as our normal positive applications of 'responsible'.

Fourth, it is not at all clear that no explanation can be given of both our ordinary positive applications of 'responsibility' and our exemptions of the brain damaged, the hypnotized, and so on, from responsibility.⁵³ For example, it is not clear that this could not be done in terms of nonregressive choice, control, and/or reason-responsiveness.

9. Is responsibility essentially whatever (if anything) occupies a theoretical role?

Suppose we shift from a context-driven to a theory-driven approach. Let us set aside versions of this approach on which meaning determines the essence of a theoretical role (see sects. 4 and 7). Consider an explanatory depth view of what is essential to a theoretical role (rather than to contexts of use). On context-driven accounts, essential properties count as essential in virtue of explaining something in actual worldly contexts of a kind-term's application. On theory-driven accounts, by contrast, essential properties count as essential in virtue of their revealing the deep structure and coherence of the theoretical role assigned to the kind in question by users of the kind term. Theory-driven accounts of reference make life easier for the eliminativist. The essence of the theoretical role assigned to triangles might be such that no existing thing has it, but contexts of application of the term do exist, no matter how misguided our theory is.

On the view in question, the claim that responsibility has an impossible essential property requires this property to have theory-internal, coherentist explanatory depth. This is hard to credit, for reasons already reviewed in other examples. Let's spell them out for responsibility.

We can begin, as heading (B) above, by collecting conditions for responsibility proposed by various theories of responsibility. There is the alternate-sequence *could-have-done-otherwise* requirement. There are also a variety of actual-sequence conditions, conditions on the character of the actual causes that lead to an act.⁵⁴ Among these are various substantive conditions: a condition of mere causal responsibility (*animism*); a requirement that the act or its consequences be *chosen* by the agent; various versions of a requirement of *control* by the agent; various versions of a requirement that the agent acts on a *reason-responsive mechanism*⁵⁵. Also among actual-sequence conditions is the structural *regression* requirement, that the agent be responsible for the causes of whatever he is responsible for. Then there is a *hypothetical or counterfactual choice* condition.⁵⁶

Some of these descriptions of responsibility are more controversial than others: conditions of control and choice are less controversial than a regression requirement. Some subsets of these descriptions are consistent with one another and others are not. To support a claim about the deep structure of the theory of responsibility, we should look for a subset that is coherent and that, other things equal, excludes more controversial conditions rather than less controversial ones. Let's make the following suppositions for the sake of argument⁵⁷ and consider where they leave the issue about eliminativism vs. revision.

The *could-have-done-otherwise* condition requires indeterminism: that causal laws may take essentially statistical form. Animism, and requirements of choice, control, reason-responsiveness, and regression are all compatible with both deterministic and indeterministic causation.⁵⁸ The regression requirement

and the could-have-done otherwise requirement are compatible; indeed, they have often not been distinguished, though neither entails the other. They could both be satisfied, for example, by regressive hypothetical choice in an indeterministic world. Though regressive hypothetical choice is not impossible, the regression requirement is not compatible with animism, or with the requirements of choice, control, or reason-responsiveness. As has been argued above, regressive choice or control is impossible.

On these assumptions, various trade-offs to achieve coherence are possible. Skipping the subtleties⁵⁹, we can hold onto choice or control at the expense of regression, or vice versa. The issue is neutral with respect to indeterminism. If we preserve regression, we can combine it with a hypothetical choice condition. So a revisionist could claim that choice or control but not regression is essential to the theoretical role of responsibility, or that regressive hypothetical choice but not actual choice or control is essential. So far, neither option eliminates responsibility on the grounds that it is essentially impossible. However, it might turn out as a matter of fact that nothing occupies either of these deep theoretical roles. Maybe the needed kind of choice or control by agents, or regressive hypothetical choice, doesn't in fact exist. Eliminativism about thought might play a role in showing this. That would yield eliminativism about responsibility, but it would not be impossible-essence eliminativism. The incoherence internal to the unpruned theoretical role would first prompt us to give a coherent account of the deep structure of the theoretical role, and then the world would turn out not to supply anything with the essential property identified.

Another possibility is that there is no good account of the deep theoretical structure of responsibility that preserves its coherence and point; it's simply an unsalvageable shambles and we should abandon it. That would eliminate responsibility rather than merely revise our view of it. But it would not do so on the grounds that an impossible property has explanatory depth internal to the

theoretical role of responsibility, and is hence essential to responsibility. A theoretical role that is an incoherent shambles with no deep structure does not have an impossible essence, something we could mourn or regret the impossibility of. It is not poignant, just confused.

It may be indeterminate whether a theoretical role is just confused, or has a coherent deep structure, if only we could figure it out. In that case it may be indeterminate whether or not the theoretical role has an occupant and whether or not eliminativism is warranted. Perhaps responsibility provides an example of such indeterminacy. But that would not show that responsibility is essentially impossible either.

10. Context and Theory in Reflective Equilibrium

I'll close with two morals. First, we're not stuck believing responsibility is impossible. If the contexts in which we apply 'responsible' are an arbitrary hodge-podge or if the theoretical role of responsibility is unsalvageably confused, then it may be that no one is ever responsible for anything, and eliminativism about responsibility is correct. But this would not be because responsibility is essentially impossible. And these are big 'ifs', far from settled.

Second, rather than rely on either a context-driven or a theory-driven approach to adjudicate these matters, we might do better to combine them into a reflective-equilibrium approach. The essence of *F*-hood would, on such a view, have explanatory depth in relation *both* to the contexts of application of '*F*' and to theoretical beliefs about *F*s. Rather than either context or theory having priority in determining essence, perhaps they work together. This is a familiar idea about 'justice' from Rawls' work and is natural to extend to 'responsibility', though it would also appear to have wider application, outside ethics. This view would not

presuppose that either the contexts of our uses of ‘responsible’ or our theoretical beliefs about responsibility have priority as explananda. Rather, it would require us to bring applications into systematic contact with theoretical role, revising and correcting each in the light of the other, and to seek a coherent deep structure or essence that sheds normative light on both contexts of use and theoretical role. This balancing approach would make it unlikely either that a wholly new theory of responsibility is needed or that all our positive applications of responsibility are mistaken.

NOTES

For helpful comments on earlier drafts and discussion of related ideas, I am grateful to: Karin Boxer, Ruth Chang, Gerald Cohen, Martin Davies, Mark Greenberg, Michael Otsuka, David Papineau, Derek Parfit, Steven Stich, Galen Strawson, and an anonymous referee for *Philosophical Studies*. I am also grateful to the Nuffield Foundation for their support of this work in the form of a Social Science Senior Research Fellowship.

¹ Stephen P. Stich, “Radical Ascent: Do True Believers Exist?”, *Proceedings of the Aristotelian Society*, Supp. LXV (1991), pp. 229-244, at p. 235; cf. Barbara Hannan, “Don’t Stop Believing: The Case Against Eliminativist Materialism”, *Mind and Language* 8(2), 1993, pp. 165-179, at p. 176n11.

² William Ramsey, Stephen Stich and Joseph Garon, “Connectionism, Eliminativism, and the Future of Folk Psychology”, in John D. Greenwood, ed, *The Future of Folk Psychology: Intentionality and Cognitive Science* (Cambridge: Cambridge University Press, 1991), pp. 93-119, at pp. 95-96.

³ Stich, “Radical Ascent”, op cit.; Steven P. Stich, *Deconstructing the Mind* (New York: Oxford University Press, 1996), ch. 1, sect. 12.

⁴ See Stich’s discussion of Lohmann’s claims in *Deconstructing the Mind*, op cit., p. 69.

⁵ Daniel C. Dennett, "Quining qualia," in *Consciousness in Contemporary Science*, A. J. Marcel and E. Bisiach, eds. (Oxford: Clarendon Press, 1988), pp. 42-77; see pp. 44, 47.

⁶ See and cf. G.E. Moore, "Reply to my Critics", in P. Schilpp, ed., *The Philosophy of G.E. Moore* (Evanston, Ill., 1942); Stich, *Deconstructing the Mind*, op cit., pp. 60ff, 171ff. Moore admitted that his account of analysis generated and did not resolve the "paradox of analysis": that some supposedly analytic truths are unobvious and open to disagreement.

⁷ David Lewis, "How to Define Theoretical Terms", *Journal of Philosophy* 67 (1970), pp. 427-446. A theoretical role may itself direct us to worldly context, such as the causes of our sayings, or direct us to defer to expert reference.

⁸ Another question is whether the term is defined functionally by its theoretical role, so that it refers to whatever might occupy that role in other worlds, however realized, or instead refers rigidly, to whatever does occupy it in this world.

⁹ For discussion, see Stich, *Deconstructing the Mind*, op cit., pp. 32-33.

¹⁰ Hilary Putnam, "The Meaning of 'Meaning'", in *Mind, Language, and Reality* (Cambridge: Cambridge University Press, 1975), pp. 215-271.

¹¹ Context-driven theories of reference need not require that the term whose reference is being theorized must figure in a scientific causal explanation of contexts of its use, as opposed, say, to normative explanations. But there is not space to develop this point here.

¹² Cf. Adrian Cussins, "Nonconceptual Content and the Elimination of Misconceived Composites!", *Mind and Language* 8(2) (1993), pp. 234-252.

¹³ See Stich, "Radical Ascent", op cit., on what this theoretical disagreement is responsible to, and to observe brisk breezes blowing through the temples of reference. See also n.7 above.

¹⁴ Stich, *Deconstructing the Mind*, op cit., ch. 1, spells out how various of these options bear on eliminativism about thought in particular. David Papineau spells out how the theoretical role approach bears on eliminativism in general, in "Theory-Dependent Terms", *Philosophy of Science* 63 (1996).

¹⁵ See Stich, *Deconstructing the Mind*, op cit., ch. 1, on the defects of the top-down ‘semantic ascent’ strategy, lack of clarity about what a theory of reference is supposed to do, and why what I have called the bottom-up, ‘normative naturalism’ strategy doesn’t resolve issues about eliminativism either. Cf. Nathan Salmon, who argues that the theory of direct reference associated with Putnam and Kripke needs supplementation with nontrivial essentialist premises to yield nontrivial essentialist conclusions, in *Reference and Essence* (Oxford: Blackwell, 1982), especially ch. 6. The issue in the text is not so much about the relation between reference and essence as the relation between positions about both reference and essence, on the one hand, and about elimination vs. revision, on the other.

¹⁶ Again, see Stich, *Deconstructing the Mind*, op. cit., ch. 1.

¹⁷ See and cf. Terence Horgan, “The Austere Ideology of Folk Psychology”, *Mind and Language* 8 (1993), pp. 282-297; Stich, *Deconstructing the Mind*, op cit.; Papineau, “Theory-Dependent Terms”, op cit.

¹⁸ At a certain level of abstraction, natural kinds might be regarded as functional kinds, where the function is given not by theory but by the world; see also notes 7 and 8 above. And see Paul M. Churchland, “Evaluating our Self-Conception”, *Mind and Language* 8 (1993), pp. 211-222. Cf. eliminativism about persons and disagreement about whether persons must be substances if they exist at all.

¹⁹ See and cf. Stich, *Deconstructing the Mind*, op cit., p. 178.

²⁰ Cf. Stich on Luhrmann on witches, *Deconstructing the Mind*, op cit., p. 68ff. Could thought be like this? Cf. Churchland, “Evaluating our Self-Conception”, op cit. Values? Cf. J. L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth, Middlesex: Penguin, 1977), ch. 1, 2.

²¹ Cf. Putnam’s denial that the members of the extension of a natural kind term necessarily have a common hidden structure. “It could have turned out that the bits of liquid we call ‘water’ had no important common physical characteristics except the superficial ones. In that case the necessary and sufficient condition for

being ‘water’ would have been possession of sufficiently many of the superficial characteristics.” But this doesn’t mean that water might not have had a hidden structure, but rather that various bits of liquid with no common hidden structure, with only superficial characteristics in common, might have looked and tasted like water, filled lakes, etc. (“The Meaning of ‘Meaning’”, op cit., pp. 240-41) . This suggests that we cannot infer from lack of any common hidden structure to elimination. Enough commonality to support reference to a kind may be found at a different, perhaps more superficial level. But if elimination is not to be impossible on context-driven accounts, there should be a distinction between cases where the explanation of the kind is relocated (eg from hidden essences to superficial characteristics, or from a physical to a functional essence) and cases where there is no explanation and the unity of the kind is illusory. This distinction requires there to be some constraint on arbitrary or accidental concatenations of properties counting as the referents of kind terms. The text appeals to explanatory depth to make this distinction. In the hodge-podge scenario, nothing has explanatory depth, even relatively shallow explanatory depth, in relation to the relevant contexts, so deconstruction/elimination rather than revision is appropriate.

²² These remarks are indebted to David Owen’s discussion in *Causes and Coincidences* (Cambridge, England: Cambridge University Press, 1992), pp. 64-65.

²³ In “Theory-Dependent Terms”, op cit., Papineau develops a theoretical-role account in terms of T-yes, T-perhaps, and T-no assumptions: yes, that assumption is criterial, perhaps that assumption is criterial, and no that assumption is not criterial. See also Frank Jackson and Philip Pettit, “Folk Belief and Commonplace Belief”, *Mind and Language* 8 (1983), pp. 298-305, at p. 302, on “commonplace psychology” and “cautious Ramsey sentences”.

²⁴ S. L. Hurley, *Natural Reasons* (New York: Oxford University Press, 1989) develops an account of ethical terms that in effect allows for both theory-internal

and context-related explanatory depth by calling for reflective equilibrium between theory and context.

²⁵ See and cf. John Roemer, *Theories of Distributive Justice* (Cambridge: Harvard University Press, 1996).

²⁶ See Dennett, “Quining Qualia”, op cit.

²⁷ Galen Strawson, *Freedom and Belief* (Oxford: Clarendon Press, 1986), p. 56.

²⁸ This is true whether the explanation in question is causal or not; impossible properties don’t do normative explanatory work either.

²⁹ Putnam, “The Meaning of ‘Meaning’”, op cit., p. 241.

³⁰ *Deconstructing the Mind*, op cit., pp. 60-63.

³¹ Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd edition (New Haven and London, Yale University Press, 1963).

³² For a materialist, are zombies like SWFs?

³³ See and cf. Galen Strawson, "The Impossibility of Moral Responsibility," *Philosophical Studies* 75 (1994), 5-24; and his *Freedom and Belief*, p. 36 of the 1986 edition, on how true self-determination is “both necessary for freedom and logically impossible”.

³⁴ E.g. Strawson, *Freedom and Belief*, 1986 edition, op cit., p. 28ff; “The Impossibility of Moral Responsibility”, op cit..

³⁵ Pace Thomas Nagel, “Moral Luck”, *Mortal Questions* (Cambridge: Cambridge University Press, 1979), pp. 26-27.

³⁶ See especially Susan Wolf, *Freedom Within Reason*. New York: Oxford University Press, 1990; Martha Klein, *Determinism, Blameworthiness and Deprivation*. Oxford: Clarendon Press, 1990; John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell, 1994. These conditions are also discussed in my work in progress, *Justice, Luck, and Knowledge*.

³⁷ See Strawson, *Freedom and Belief*, op cit.1991 edition, eg. p. 29, and “The Impossibility of Moral Responsibility”, op cit.; cf. Nagel, “Moral Luck”, op cit., pp. 26-27.

³⁸ When used without the qualification 'causal', 'responsibility' is here intended in a desert-involving sense.

³⁹ For discussion, see G. A. Cohen, "On the Currency of Egalitarian Justice," *Ethics* 99 (4) (1989), 906-44, at 935ff; Thomas Scanlon, "Equality of Resources and Equality of Welfare: A Forced Marriage?," *Ethics* 97 (1986), 111-8.

⁴⁰ Note that neither choice nor control is intrinsically regressive. This is argued more fully in *Justice, Luck, and Knowledge*, in progress. Briefly: It is not hard to see that to choose something does not require choice of its causes. Control involves maintenance of a variable at a target value in the face of exogenous disturbance, where the variable is caused to take values caused jointly by factors endogenous and factors exogenous to a control system. To control something not only does not require control of its causes, but in fact presupposes causes exogenous to the control system. Control occurs in nature as well as in human affairs.

⁴¹ *Anarchy, State, and Utopia* (New York: Basic Books, 1974), p. 225.

⁴² Strawson, *Freedom and Belief*, op cit., 1986 edition, p. 106.

⁴³ Such deep commitment to a regressive conception of responsibility can take a form such that when we press the question what such responsibility could be, or what it would require, we are led into the regressive choice story, even though we never ordinarily think of such regresses. This is what Strawson has in mind (personal communication). Deep conceptual commitments may not be superficially accessible, may require such reflective questioning to reveal. Suppose this is what our commitment to a regressive choice conception is like. Such a commitment still constitutes us as having conflicting intuitions if at the same time we do not feel that our freedom is put in question by the fact that we cannot be truly or ultimately self-determining in any way. Thanks to Galen Strawson for prompting clarification here.

⁴⁴ *Ibid*, pp. 8, 26-30, 49-50.

⁴⁵ Strawson, *Freedom and Belief*, op cit., 1986 edition, p. 30. In the 1991 edition Strawson revises this claim as follows: "But the freedom that is shown to be

impossible by this sort of argument against self-determination is just the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess, even though the idea that some sort of ultimate self-determination is presupposed by their notion of freedom has never occurred to them. It is therefore worth examining the argument in detail. For the idea that we possess such freedom is central to our lives” (p. 30). Two points in response to this revision: First, my points are not ad hominem but directed to a position, so Strawson’s earlier statement can still serve to put the position in play. Second, the revision does not in fact affect the points to be made. For present purposes, it doesn’t matter whether what people presuppose has occurred to them or not. Suppose people do not feel that their freedom is put into question by the influence of heredity and environment, which means that they cannot be truly or ultimately self-determining. Suppose also they presuppose that their freedom has a feature that is incompatible with such influence. Then they are conflicted. This is true whether or not the presupposition they make has occurred to them, and whether or not they realize that they are conflicted. Thanks again to Galen Strawson for prompting clarification.

⁴⁶ Ibid., p. 101.

⁴⁷ Ibid., p. 96; see also p. 88. Nagel also suggests (in “Moral Luck”, op cit.) that we are conflicted, tending sometimes to more restrictive, sometimes to less restrictive views of responsibility. He regards the more restrictive views as most in accord with intuitive ethics. See also Bernard Williams, “Moral Luck”, in his *Moral Luck* (Cambridge, England: Cambridge University Press, 1981).

⁴⁸ So causation can be transitive even if it is not regressive.

⁴⁹ Consider various versions of a reason-responsiveness requirement, perhaps coupled to requirements of choice and/or control; see, for example, John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control*, op. cit.

⁵⁰ Cf. Galen Strawson: "True self-determination is logically impossible because it requires the actual completion of an infinite regress of choices or principles of choice" (*Freedom and Belief*, op. cit., 1986 edition, p. 29).

⁵¹ *Deconstructing the Mind*, op. cit., pp. 60-63, 171ff.

⁵² Cf. Klein, *Determinism, Blameworthiness, and Deprivation*, op. cit., ch 4.

⁵³ Cf. Klein, *ibid.*, ch. 4

⁵⁴ For the alternate-sequence/actual-sequence distinction, see John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control*, op. cit.

⁵⁵ See Fischer, *Ibid.*, on the idea of reason-responsiveness.

⁵⁶ See Cohen, "On the Currency of Egalitarian Justice," op. cit., pp. 935ff; Scanlon, "Equality of Resources and Equality of Welfare: A Forced Marriage?," op. cit.

⁵⁷ Which are argued for in *Justice, Luck, and Knowledge*, in progress.

⁵⁸ Animism need not differentiate between deterministic and indeterministic causation. Deterministic causation is no guarantee against minor malfunctions; an indeterministic mechanism can in principle be just as reliable as many deterministic mechanisms. Just as a thermostat that includes an indeterministic process can be just as good at controlling the temperature as one designed differently that includes no indeterministic process (Randolph Clarke, "Indeterminism and Control," *American Philosophical Quarterly* 32 (2) (1995), 125-37 at 129), so a person who is realized in part by indeterministic processes can be just as good a chooser, a controller, or a responder to reasons as a person who is deterministically realized. Moreover, a regression requirement could in principle be applied to indeterministic causes as well as deterministic causes.

⁵⁹ --In particular, setting aside complications introduced by important issues about reason-responsiveness. Various reason-responsiveness conditions for responsibility are investigated further in my *Justice, Luck, and Knowledge* (in progress), but they do not support a regression requirement.