



Why Richard Brandt Does Not Need Cognitive Psychotherapy, and Other Glad News about Idealized Preference Theories in Meta-Ethics

DAVID ZIMMERMAN

Department of Philosophy, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

1. Introduction

On the occasion of this centennial acknowledgement of G.E. Moore's *Principia Ethica* it is worth exploring one aspect of the plausibility of a kind of naturalistic meta-ethical theory that Moore briefly took notice of, only to dismiss.¹ The opportunity to do so is apt, because in the second half of the twentieth century, a family of theories of just that kind flourished, which may be called idealized preference theories. Their defenders analyze, reduce, or otherwise ground intrinsic value, basic rightness and, more generally, reason for action in a person's idealized, more particularly, his factually informed and otherwise rationally constrained, preferences and include Richard Brandt, Roderick Firth, David Lewis, John Rawls, Peter Railton, Michael Smith, and Bernard Williams.

Moore's famous open-question argument is, at best, inconclusive when directed against such accounts of the good and the right. George Nakhnikian puts the point nicely:

Is this [really] an "open question": After all, is a thing good which is so constituted that it would reinforce the desires, sustain the interest, and occasion the satisfactions and enjoyments of everyone who had a mature and comprehensive grasp of that thing's scientifically discoverable properties and relations? An appeal to the logical contours of ordinary discourse cannot settle the issue as to the openness of the question. If we are inclined, as we are, to saying that a thing of that description might, without contradiction, be said not to be good, *we are also inclined in the opposite direction of wondering what on earth more goodness could be.*²

Despite the good sense of this, idealized preference theories are vulnerable to a number of other objections, which must be taken seriously by anyone who is sympathetic with this meta-ethical program. A standard criticism since their emergence has been that they generate seriously counterintuitive implications, which undermine two important aspects of normativity, rational authority and recommendatory force. This problem can be solved, or at least blunted, with

the addition of a specifically historical dimension to idealized preference theories. In addition to being ideal observer, exemplar, deliberator, or advisor theories, as they have variously been conceived, idealized preference theories also need to be ideal development theories, with perhaps an ideal educator watching over the process.

There is also a less standard, but particularly daunting, cluster of objections, most fully developed by Connie Rosati, which can also be blunted by historicizing idealized preference theories. Rosati argues that the idealization at their core exhibits a serious incoherence. The bird's-eye-view set up by the idealization makes it impossible for these theories to accommodate the deeply constitutive fact that persons have particular evaluative perspectives or points of view, a feature inextricable from their very status as genuine agents.³ This objection can also be met or at least softened if defenders of idealized preference theories add a historical dimension to their accounts of what makes a preference rational, authentic or otherwise suitable for grounding the good and the right.

2. Idealized Preferences and Practical Reasons

Moore turned a critical eye upon a variety of theories of the good that, he claimed, run afoul of the naturalistic fallacy. On one sort of hedonistic theory that he dismisses in Chapter Three of *Principia Ethica*, intrinsic value is analyzed in terms of idealized subjective states of persons. Moore remarks:

Nothing is more natural than the vulgar mistake which we find expressed in a recent book on Ethics: "The primary ethical fact is, we have said, that something is approved or disapproved: that is, in other words, the ideal representation of certain events in the way of sensation, perception, or idea, is attended with a feeling of pleasure or of pain."⁴

In pursuing this theme Moore quotes the notable classicist A.E. Taylor, but he might just as well have cited central ideas in the theories of value to be found in the two celebrated moral philosophers he spends most of Chapter Three critically discussing. Alongside the more straightforwardly hedonistic accounts of intrinsic value that J.S. Mill and Henry Sidgwick defend, both articulate versions of idealized preference theories of the good. Mill's account of the qualitative dimension of utility in terms of the preferences of subjects of wide knowledge and experience is naturally construed in such a light.⁵ Sidgwick explicitly claims that "a man's future good on the whole" is constituted by "what he would now desire and seek on the whole *if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realized in imagination at the present point of time.*"⁶

On idealized preference meta-ethical theories, a state of affairs, an object, action, or institution has a particular evaluative or normative property such as goodness, rightness, or practical rationality if and only if a person suitably idealized in cognitive respects, specifiable independently of the property, would respond positively to the state of affairs, object, action, or institution by doing something. As Richard Brandt put it: "I shall call a person's desire, aversion or pleasure 'rational' if it would survive or be produced by careful 'cognitive psychotherapy' for that person. I shall call a desire 'irrational' if it cannot survive compatibly with clear and repeated judgments about established facts."⁷ Roderick Firth suggests that "it must be possible... to express the meaning of statements of the form 'x is right' in terms of other statements which have the form: 'Any ideal observer would react to x [positively] under conditions... in which she were omniscient with respect to non-ethical facts, omniscipient, disinterested, dispassionate, consistent, and in other respects normal."⁸ David Lewis states: "Something of the appropriate category is a value if and only if we would be disposed, under ideal conditions, to value it."⁹ Similar views have been held by John Rawls, Peter Railton, Michael Smith, and Bernard Williams.¹⁰

There are two importantly different versions of this view. On the first, the good and the right count as relational or response-dependent properties, on analogy with putatively secondary sensory properties like yellow and loud. Such idealized preference theories are offered as the metaphysical truth-makers of evaluative and normative statements. Brandt, Firth, Lewis, Rawls, and Williams offer accounts of this sort. On the second version, the position is offered merely as an epistemological or heuristic account of how to detect, track, or otherwise discover the nature of a good state of affairs or of a right action, the metaphysical status of which is somehow fixed independently of idealized responses. Railton defends this sort of view.

Moore does not take up the view explicitly, but there is no doubt that he would have found all idealized preference theories to be vulnerable to the open-question argument. The near-consensus today is that the emergence of causal-historical theories of reference and natural kinds in semantic theory blunt the force of the open question challenge to any proffered analysis of "goodness" or "rightness." On these semantic views meta-ethical theories of the good and the right are not to be construed as linguistic or conceptual analyses of evaluative or normative concepts, but as *a posteriori*, albeit metaphysically non-contingent, accounts of what the properties goodness, rightness and practical reason are.¹¹ Despite the enormous influence of Moore's general attack upon meta-ethical naturalism in the first half of the twentieth century, the heyday of the so-called linguistic turn in philosophy, idealized preference theories of the good and the right have endured since the days of A.E. Taylor, Moore's target.

3. The Dialectical Appeal of Idealized Preference Theories

Not all theorists who adopt a broadly naturalistic approach to meta-ethics advocate idealized preference cognitivism. Some prefer primary property realism, while others are attracted to non-cognitivism. Prominent proponents of primary property realism take the properties of goodness, rightness and practical reason to be non-analytically reducible to the non-moral properties that figure as good-making and right-making in an associated substantive evaluative or normative theory.¹² Such primary property realists hold that the metaphysical nature of goodness or rightness can be accounted for in terms of the non-moral, natural facts that figure as substantive reasons in defending basic moral claims. Even when all the semantic and metaphysical niceties about reference and truth conditions are in place, however, such theories strike some meta-ethical irrealists as attempting to pull off a kind of “fatuous validation by definition,” which is vulnerable to some kind of open question challenge, if not precisely Moore’s own.¹³

Primary property moral realists reply that such questions are to be closed by empirical discovery, guided by the appropriate method for settling all basic theoretical claims, the method of wide reflective equilibrium.¹⁴ The divisive question then becomes whether wide reflective equilibrium is more appropriately taken to be a realist method of empirical discovery or an irrealist method of moral construction. Proponents of idealized preference subjectivism will opt for the constructivist construal and will note that a more modestly formal or procedural cognitivist account of the truth conditions of moral statements, in terms of cognitively idealized preferences, upholds the worthy tradition of meta-ethical neutrality, by not begging any substantive moral questions in the very construction of its metaphysical account of what the properties of goodness, rightness and practical reason are. Such a meta-ethical view is attractive. It is a kind of cognitivism, which avoids the various semantic embedding problems that continue to plague even sophisticated versions of non-cognitivist norm-expressivism and quasi-realism. It also supports the cognitive status of substantive moral statements without employing a morally question-begging methodology. Moreover, it offers a plausible account of the two faces of normativity. Both the recommendatory illocutionary force insisted upon by non-cognitivists and the rational authority of moral sentences defended by moral realists are to be grounded in relatively unimpeachable constraints from the realm of theoretical reason.

4. Some Allegedly Counterintuitive Implications of Non-Historical Informed Preference Theories

All of the prominent idealized preference theories currently available are non-historicist, grounding the good or the right in the contemporary responses of

an idealized person. Richard Brandt advances his idealized preference theory in *A Theory of the Good and the Right*. There he says: “a person’s desire, aversion or pleasure [is] ‘rational’ if it would survive or be produced by careful ‘cognitive psychotherapy’ for that person. . . . a desire [is] ‘irrational’ if it cannot survive compatibly with clear and repeated judgments about established facts.”¹⁵ Brandt’s theory is non-historical. On his view, a preference counts as rational, and thus as constitutive of the good and a rational ground of the right, if and only if it would survive or be produced by contemporary “cognitive psychotherapy.” That process is an appropriate exposure at the time of assessment to “facts and logic,” alternatively, to “relevant information, represented in an ideally vivid way, and that the appropriate time. . . without the influence by prestige of someone, use of evaluative language, extrinsic reward or punishment, or use of artificially induced feeling states like relaxation.”¹⁶ Brandt resists more aggressive versions of attitudinal extinction, such as Freudian psychoanalysis, out of a concern to preserve the substantive evaluative and normative neutrality of his theory.¹⁷

The idea of cognitive psychotherapy has been one of the most criticized aspects of Brandt’s theory. Yet it has also been characterized as “arguably the lynchpin” of the theory.¹⁸ A standard objection is that grounding the rationality and irrationality of preferences in such a therapeutic process yields seriously counterintuitive implications and has two forms. One concern is that the mere fact that a preference is not eliminated by current cognitive psychotherapy cannot be constitutively sufficient for its rationality or authenticity. Another is that the mere fact a preference is eliminated in current cognitive psychotherapy cannot be constitutively sufficient for its irrationality or artificiality. The first concern yields a problem of recalcitrant but obviously irrational preferences, the second yields a problem of informed but intrinsically objectionable preferences.¹⁹

The first concern is the more straightforward. Allan Gibbard offers the apt example of the persistent hand washer, subjected to cognitive psychotherapy, whose obsessive-compulsive impulse nonetheless remains. His deeply engrained but clearly unreasonable and even self-destructive compulsion is not extinguishable by cognitive psychotherapy, even though the process is administered with all due persistence and care to highlight accurately and vividly the seriously harmful consequences of repetitive hand washing. The recalcitrant compulsive hand washer might nonetheless respond: “I realize all that. But I just do not want those creepy-crawly things on my hands – and least of all do I want to be a person who would be willing to tolerate them on his hands.”²⁰ However, on Brandt’s view, a preference that is currently non-extinguishable by cognitive means counts as rational or authentic, perhaps even as a constituent of the person’s current good, solely for the reason that it is cognitively non-extinguishable.²¹ This is clearly wrong. It is a serious problem for Brandt’s theory, as it stands, that such an overwhelming, indeed crazy,

desire is not eliminated by unimpeachable epistemic purification. However, we may ask: What explains this unfortunate state of affairs, and can a good explanation reveal how the theory might be revised to solve the problem of recalcitrant but obviously irrational preferences?

The second concern raises subtle issues. Gibbard offers the examples of an honest civil servant and a committed pure egoist, who refuse to be subjected to cognitive psychotherapy, because they are afraid that vivid exposure to certain facts will extinguish preferences they currently endorse as signally important to their identities. He observes: "These examples have a common structure. . . . [They] are cases in which a person thinks himself an unreliable transformer of vivid realizations [of the facts] into rational desires."²² He adds: "On a full-information account like Brandt's, this talk of reliability has no substance: that we are reliable transformers of vivid realizations into rational desires is analytic – true just in virtue of what 'rational' means. . . . In all the examples, though, there is a common element that an account like Brandt's misses: the protagonist endorses a system of ends he thinks would not survive a vivid, repeated confrontation with the facts."²³ These examples raise a question about whether Brandt's meta-ethical theory can accommodate the crucial element of normative endorsement. However, Gibbard's characterization of the cases, and thus of the nature of the allegedly counterintuitive implications of Brandt's theory, are liable to be misleading in two respects. He takes all three cases to exemplify a common structure. But there are two distinct putative problems here, the problem of recalcitrant, irrational preferences, and the problem of informed, intrinsically objectionable preferences. As well, it is not at all clear that what the civil servant and the pure egoist refuse to undergo is indeed genuine cognitive psychotherapy as Brandt would require.

The honest civil servant has a deep moral aversion to the idea of betraying his public trust by taking a bribe. At the same time, he fears that his moral resolve would weaken to the breaking point if he were to undergo cognitive psychotherapy and thus vivid exposure to all of the available facts, including those that concern the wonderful goods he could buy with the accepted bribes. On Brandt's account, it seems to follow that the civil servant's intrinsic aversion to taking bribes is irrational or inauthentic. On some reasonable assumptions, this too is clearly the wrong result, but it is a different wrong result. Whereas the persistent hand washer has a current preference that is clearly irrational, however deeply he may identify with it, the honest civil servant has a current preference that is *prima facie* legitimate, however vulnerable it may be to vivid exposure to certain facts.

It is not clear from Gibbard's characterization that what the honest civil servant and the committed egoist fear is vivid exposure to all the available facts, as opposed to exposure to a selected sub-set of them, the fruits of crime in the one case and of selfless altruism in the other. Brandt's idea is that a

preference is rational or authentic if it would continue or be produced by vivid exposure to all the available facts. This opens up the possibility that the honest civil servant and the pure egoist are simply wrong about how their respective preference structures would end up after thoroughgoing cognitive psychotherapy. They exhibit an alarming lack of confidence in the ability of their cherished values to stand up to the facts. If their lack of confidence is misplaced, then there is no problem of informed but intrinsically objectionable preferences, since everything will come out all right at the end of the day.

Gibbard takes these to be cases in which “a person thinks himself an unreliable transformer of vivid realizations [of the facts] into rational desires. . . .”²⁴ However, this prompts the question of what notion of reliability the civil servant and the egoist might have in mind in entertaining such deep self-doubt. What reason might they have for fearing that their very selves are filters that would distort an otherwise epistemically impeccable process like cognitive psychotherapy? Gibbard attempts to prise apart conceptually the ideas of reliable attitude formation and contemporary cognitive psychotherapy, evidently because he believes that this is necessary to preserve a sufficiently robust element of normative endorsement in an adequate meta-ethical account of practical judgment. That is an important aim but it is unclear why Gibbard thinks that an agent’s good reasons for such normative endorsement are so independent of the sorts of factual considerations that would figure in cognitive psychotherapy that Brandt would require, such as instrumental facts about which relatively stable self-regarding and other-regarding preferences would enable human beings to solve certain coordination problems. It is notable that Gibbard’s own account of the internalization of norms and the “biology of coordination” is replete with such considerations.²⁵

5. Historicized Ideal Preference Theory

Brandt would have been better served if he had added an explicit historical dimension to his account of the difference between authentic and inauthentic preferences. We can be confident that most, perhaps all, victims of obsessive-compulsive, phobic, and similarly defective preferences have acquired their preferences during the course of a history of attitude formation seriously tainted with the kind of errors of fact and logic that Brandt hoped cognitive psychotherapy in adulthood would redress. Nonetheless, it should come as no surprise that some personal histories are so burdened with cognitive error and the resulting preferences so deeply engrained, say via vigorous classical and operant conditioning, that little can be done to extinguish them by means of exclusively cognitive techniques. The core idea of idealized preference theories is that the good and the right are grounded in preferences calibrated to

the way the world actually is. The obsessive-compulsive hand washer has acquired his preference, his impulse, in a fashion that clearly fails this reasonable constraint, since the creepy-crawly things on his hands, though present, are harmless. Moreover, being the kind of person who would tolerate them on his hands is no shameful thing. In the face of such facts, and at great cost to himself, he cannot respond appropriately to such facts, given his history.

Even if historical idealized preference theorists do succeed in heading off the charge of counterintuitive implications, there are bound to remain serious epistemic difficulties in determining whether a plausible historicized condition is satisfied in any given case. However, this is true of nearly all historicist conditions in philosophy, as can be seen in the difficulties to be encountered in a serious practical application of causal-historical conditions of reference fixing or of historical-entitlement conditions of just distribution. Even proponents of historicist conditions acknowledge that there may be epistemic difficulties applying the condition. This sort of difficulty is blunted, however, if idealized preference theories are offered as metaphysical accounts of the truth-makers for statements about the good and the right, not discovery-procedures.

6. Richard Brandt's Latent Historicism

The standard objections to Brandt's conceptual deployment of current cognitive psychotherapy suggest that an adequate idealized preference theory of the good and the right cannot focus exclusively on a subject's present psychological openness to the actual nature of the object of his preferences, but must add a diachronic dimension. On Brandt's view, "a person's desire, aversion or pleasure [is] 'rational' if it would survive or be produced by careful 'cognitive psychotherapy' for that person. . . 'irrational' if it cannot survive compatibly with clear and repeated judgments about established facts."²⁶ However, Brandt's own prior and more basic theory of rational preference in *A Theory of the Good and the Right* is historicist through and through, in being "based on the theory of the genesis of pleasures and desires."²⁷ Thus he devotes a key portion of his theory to an account of the various ways in which the process by means of which a preference is acquired can render it mistaken. The distinctly historicist aspect of his theory comes out in his explanation of why certain preferences are extinguishable by contemporary cognitive psychotherapy and thus count as irrational or artificial. He remarks: "The production of . . . intrinsic desires and aversions is artificial if they could not have been brought about by experience with actual situations which the desires are for or the aversions against"²⁸ This is an unequivocally historical condition.

Brandt cites four kinds of mistakes that can generate inauthentic preferences: having false beliefs about instrumental consequences, misgeneralizing

from untypical examples via classical conditioning, acquiring preferences by means of cultural reinforcers involving social status, and acquiring desires with an overweening strength that is traceable to early deprivation of the object of those desires.²⁹ In each instance the explanation of artificiality or inauthenticity is more or less the same: the person would not have acquired the intrinsic preference but for having undergone a historical process of attitude acquisition that involved some kind of epistemic mistake about the object of the preference. The mistake in each instance is also more or less the same: during a learning process that involves conditioning the person confuses the actual content of the acquired preference with the content of an extraneous reinforcer.

The explanation of how someone might become an obsessive-compulsive hand washer can easily involve mistakes of the first two types Brandt identifies, and perhaps the third. A compulsive hand washer clearly has some empirically incorrect beliefs about the consequences of allowing the creepy crawly things to remain on his hands. Conditioning models can be employed to explain how he acquired them. We can see why the honest civil servant and the committed egoist should be regarded differently from the persistent hand washer with respect to the authenticity and inauthenticity of their preferences. Unlike the persistent hand washer, the honest civil servant and the committed egoist did not acquire their commitments to honesty and egoism, respectively, in histories of attitude acquisition that involved epistemic errors. This is, perhaps, clearer with the civil servant, given that we can expect that norms of honesty have the sort of empirically grounded and instrumentally rational role in solving coordination problems that Gibbard suggests.

7. Why Richard Brandt's Official View is Non-Historicist

Given that the idea of cognitive psychotherapy has come under such critical fire, we would expect Brandt to have done more to vindicate the alternative historicist version of his own idealized preference explanation of the difference between authentic and inauthentic desires. However, Brandt shies away from this option because he places a great emphasis upon the acquisition of preferences by classical conditioning. Thus, he is afraid that a historicist version, applied strictly, has the counterintuitive implication that all acquired preferences, including motivations as basic to the grounding of the good and the right as benevolence itself, will count as irrational, inauthentic or artificial. He puts the point this way: "Conditioning is in a sense unfair; for it produces a positive or negative response to something *not guilty except by association*."³⁰ He then applies the point to our area of concern:

For this reason I explained "cognitive psychotherapy" in such a way as to keep some touch with reality; so that a desire or liking ends up as irrational

only to the extent that repeated self-stimulations would actually diminish it. Some likes and desires extinguish in this way; but many (partly because of early conditioning) will not: among them, presumably, my liking for the company of my mother, or my horror at seeing someone tortured.³¹

Brandt's concerns about the counterintuitive implications of the historicist version are exaggerated. He places too much emphasis on the role of classical conditioning in the acquisition of benevolence, and too little emphasis on how thoroughly a child's internalization of this motivation is implicated with her distinctly cognitive advancement as an epistemic agent. This is not to suggest that classical conditioning plays no crucial role in the genesis of preferences like benevolence. Any plausible story of preference acquisition will include some reliance on conditioning mechanisms.³² The question is whether their role automatically impugns the rationality or authenticity of the acquired preferences. Brandt to the contrary, in the crucial case of infant benevolence, it does not. Under certain conditions, preferences acquired through classical conditioning can be appropriately sensitive to the facts and logic and serve as suitable material for grounding the good and the right.

8. The Genesis of Infant Benevolence

The conditioning challenge to the epistemic integrity of infant benevolence has two aspects. Following the lead of the psychologist Martin Hoffman, Brandt distinguishes between benevolence or sympathy, the belief mediated aversive response to another's expression of pain and empathy, the primitive, belief unmediated, response to the same kind of stimulus.³³ Brandt maintains that benevolence or sympathy is an epistemically compromised motivation for two reasons. It is compromised because the best explanation builds upon the earlier acquisition of empathy through classical conditioning, a process that involves epistemic error and because the best explanation of how a baby acquires sympathy upon that foundation also involves putatively error-laden classical conditioning.

On Brandt's view, the unconditioned stimulus of a baby's own pain natively elicits in her the unconditioned responses of behavioral withdrawal, certain autonomic events and crying. Since the baby's own pain, an unconditioned stimulus, is typically accompanied by her own crying, an unconditioned response, her crying becomes, through classical conditioning, a conditioned stimulus that comes to elicit the other response types, which are then conditioned responses to crying in general. When the baby hears a second baby crying, she is conditioned to respond with crying. One epistemic error enters the picture with the first infant's stimulus misgeneralization from its own pain to its own crying. Another error enters with the first baby's further stimulus misgeneralization from the second baby's crying to its own crying.

The pain behavior of the second child is already a conditioned stimulus that elicits the first infant's empathetic conditioned response of crying. By this stage of the first baby's cognitive development she has begun to acquire the conceptual capacity to make an elementary kind of distinction between herself and others. The second baby's crying thus comes to be "linked cognitively with the first baby's representation of correlated internal states of the second baby."³⁴ The first infant is able to entertain proto-thoughts with the content "I infer from his crying that the other baby is the one who is in pain." Why, then, does the first baby cry upon hearing the second cry? How does the first child come to entertain proto-aversions with the content "I do not like it, as an end-in-itself, that the other is in pain"? Brandt's answer is, yet again, classical conditioning. The first baby's representation of the second baby's inner state of pain comes to be associated through the conditioning mechanism with her own innately disliked pain, an unconditioned stimulus, so that the representation becomes a new conditioned stimulus, which in turn elicits her further crying as yet another conditioned response. Her further crying manifests sympathy, or benevolence properly so-called, because its content makes essential reference to the pain of the other *qua* other. Nonetheless, Brandt claims that epistemic error is still in the picture by virtue of the conditioning mechanism that operates.

In the face of developmental accounts that seem to be so thoroughly tainted with epistemic error, what can we say in vindication of a motivation so basic to moral agency as infant benevolence? On Brandt's view, it is as though the first baby were like human subjects of a Pavlovian experiment who end up with an intrinsic aesthetic taste for the sound of buzzers, acquired through classical conditioning of the buzzer sound to the taste of meat powder. Moreover, on Brandt's view, it is as though the baby were like any of the victims of the four kinds of mistaken desires he cites as examples of epistemically compromised preference acquisition.

Cases of conditioned preference and aversion do not provide accurate parallels with what is going on with the first baby at either stage of her development of benevolence. Hoffman, in effect, supports the idea that the law of effect does not epistemically compromise this first stage in the infant's acquisition of benevolence when he stresses the innate aspects of infant empathy. He remarks: "'experiencing distress when another is in distress seems primitive, naïve, reasonably universal' – as natural a response as anger is to threats to the self (and, as with anger, *only the specific form of empathy is due to learning*."³⁵ Evidently, a case can be made that the disposition to react empathetically is itself innate; only the manner in which it comes to manifest itself is acquired. The idea that infant empathy is not acquired in fact-insensitive fashion is reinforced by Brandt's own sensible acknowledgement that innate or native preferences count as rational or authentic.³⁶ Therefore, it is reasonable to view a child's early conditioning as inducing determinate preferences upon the foundation of an innate space of determinable preferences.

A prevailing idea in contemporary theories of mental representation is that cognitive systems cannot entertain genuine representational states unless they can also entertain misrepresentations. This suggests that conversely there can be no misrepresentation without the possibility of accurate representation. However, at the first stage of her story the baby lacks any sort of conceptual grasp, however primitive, of the distinction between her mind and its states and other minds and their states. Therefore, if the prevailing idea is correct, it is conceptually impossible for the baby to represent her conditioned response accurately as directed upon the wrong self, or upon the right self, for that matter. Therefore, if the converse slogan is correct, she cannot misrepresent anything about the real locus of the pain she responds to by crying. Since there is no misrepresentation, there is no cognitive error, properly so-called.

A further reason for concluding that the infant's acquisition of sympathy through classical conditioning does not involve epistemic error in the form of misrepresentation is that the first baby acquires the capacity to feel sympathy, a cognitively mediated state, only as she develops the capacity to differentiate herself from others. This in itself happily implicates the emergence of sympathy with a distinctly cognitive achievement, a fact that ought to enhance the epistemic credentials of benevolence, not impugn them. The cognitive advance takes place gradually. As Hoffman notes:

The gradual nature of self-other differentiation is . . . important, because it gives the child the experience of simultaneously wanting to terminate the emerging other's distress as well as its own – thus providing a link between the initially hedonistic empathetic distress response and the earliest trace of sympathetic distress. If the sense of the other were attained suddenly, the child would lack this experience; when he discovers that the pain is someone else's, he might simply react with relief (or even blame the other for his empathic distress).³⁷

Thus, the child's first real differentiation of herself from others occurs roughly simultaneously with the emergence of her first real capacity to feel sympathy. If there is indeed no genuine misrepresentation without representation, then the sort of epistemic error that Brandt fears would impugn sympathy becomes possible only after the child is possessed of this motivation.

9. Diachronic Persons and Particular Perspectives

We may put the historicist idea to use in trying to meet a more daunting kind of objection to idealized preference theories concerning their coherence. Connie Rosati puts the main point this way:

The fact that our motivational and cognitive features affect how we experience things poses a deeper problem than it may initially seem. For it suggests a tension between what it is like to be a particular person and what is required for a person to be fully informed. Part of being a particular person with particular traits is occupying a point of view – one that involves a certain way of seeing, feeling and evaluating and which gives access to certain information while making other information inaccessible. If a person is to be fully informed, however, she must be able to enter into all possible points of view. She must be capable of appreciating all her lives as the persons she would be if living them – side by side, so to speak. The problem concerns how she can occupy a point of view that gives her equal access to viewpoints that may be in direct conflict, each excluding information accessible from the other.³⁸

The charge of incoherence is deep and subtle. The idea is not simply that there is a vast difference between the cognitive capacities of ideal advisors and the capacities of ordinary fallible people. That much is obvious already. Nor is it simply that the difference is enough in itself to jeopardize the normativity of judgments about the good and the right when understood in terms of idealized preferences. Normativity has action-guiding motivational power and justificatory authority. Rosati believes that advocates of idealized preference theories have trouble accommodating both aspects of evaluative judgments.³⁹ These problems are derivative from the putative incoherence in the idealization. The incoherence has two aspects, one internal to the ideal advisor herself, the other involving her relationship with her non-idealized self.

The first aspect involves the difficulty of an ideal advisor meeting Sidgwick's standard of accurately foreseeing and adequately realizing in imagination, either at a single time or manageably *seriatim*, all the different lines of conduct or ways of life open to the ordinary person she is advising. Such a task requires the ideal advisor to represent to herself in usable, commensurable form many radically different particular kinds of lives, some of which are mutually incomprehensible from the point of view of someone living some of the other particular lives represented. Rosati considers various ways in which such representation and summing problems might be solved. They involve different methods of entertaining and ordering the representations, and different uses of memory and empathy in measuring them. However, she is pessimistic about all of them on the grounds that no idealized observer could "occupy a perspective that gives her equal access to lives in which she would have conflicting traits."⁴⁰

J. David Sobel helps us to get a fix on the second aspect of putative incoherence when he says:

some of the limitations which are idealized away by the full information account play a fundamental role in shaping our capacity to value in the ways

we do. In order to have many experiences one must be a particular kind of person. The idealized self which the full information theorist recommends is not the kind of person who could have some of the experiences which could be ours.⁴¹

The problem here is the ideal advisor's inability *qua* cognitively idealized subject to understand crucial elements in the lives open to her non-idealized advisee. The idea is that cognitive idealization can itself be a limitation on comprehension, because occupying a particular non-idealized way of seeing, feeling, and evaluating is necessary for the availability of some information, which is in turn necessary for understanding that particular perspective.

10. There is Nothing Like a Particular History to Concentrate a Person's Evaluative Perspective

The problem Rosati identifies is deep but may not be irremediable. Perhaps it can be solved by adding a historical dimension to the idealization at the core of the position. The key thought is that a person's history of psychological development, from infancy through childhood and adolescence into adulthood, is a major determinant of the values that constitute her eventual particular personal evaluative perspectives or points of view. Her history crucially includes her own choices, some of which determine the range of substantive values that are accessible, or inaccessible, to her later in life. At crucial moments she herself initiates courses of motivational development, which open up some ways of life for her and close off others. For example, if she embarks far enough along the path of a sybarite, it is unlikely that she can easily redirect her steps on the path of an ascetic. These are the sorts of incommensurabilities of evaluative perspective that Rosati and Sobel argue cannot be encompassed within the synchronic perspective of Sidgwick's idealized advisor.

Human beings have final and instrumental ends, but crucially they also have determinable or maieutic ends.⁴² Infants are born with a set of innate concepts, beliefs, cognitive capacities and preferences, as well as with a set of innate dispositions to develop more. They gradually become young children, who somehow manage to develop and acquire new concepts, beliefs, cognitive capacities and preferences as they interact with their early environment. As they move into adolescence and early adulthood, they continuously articulate an initially simple framework of final and instrumental ends with more or less determinate content. Eventually they come to have another kind of end, which is determinable or maieutic, in that they can satisfy it only through a process of bringing themselves to have further determinate final ends. Such processes create the very conditions for some paths to evaluative fulfillment and more or less irrevocably close off others.

For example, a young adult starts out with the maieutic end of having a fulfilling career. Her pursuit of this end requires her to choose more determinate ends, a certain program at school, a certain kind of apprenticeship. Eventually, she settles on a plan to enter a profession and become a politician rather than an explorer, an artist, or an academic. Once she has made such a choice, she has a new set of determinate final ends, the pursuit of which becomes an end in itself for her. Satisfying her maieutic end of having a fulfilling career gives her life meaning through her own specification and articulation of final ends in the sphere of work. Later, having gone far toward cultivating the skills and values of a just compromising enough legislator, she finds it difficult, perhaps psychologically impossible, to go back and to cultivate in herself the values of the uncompromising, finely perceptive painter, composer, or novelist. To use terms that Rosati favors in this context, the young politician simply cannot “appreciate” a way of life at the end of the path not taken and she finds it nearly impossible to be genuinely “informed” about what such a life would be like from the inside.⁴³ Parallel processes occur in the other spheres that really matter in life, especially in the emotional realms of love and friendship. Again, this sort of process can create the sort of evaluative incommensurabilities that Rosati and Sobel find in synchronic idealized preference theories.

But the fact that people transform maieutic into final ends need not be viewed in this light. The capacity to recognize determinable ends and then to articulate them as more and more determinate ends is a signally important aspect of human motivational development. It may be hard, perhaps impossible, to imagine what it would be like synchronically to occupy the bird’s-eye view of Sidgwick’s ideal advisor, capable of pulling into vivid, informed focus, simultaneously or manageably *seriatim*, all the lives that might be open to a particular young person, by virtue of her actual aptitudes and resources. But, despite the dominance of such bird’s-eye synchronic models, that is something no advocate of an idealized preference theory should aim for. The appropriate degree of cognitive idealization can be captured diachronically. Young people do make value-shaping, life-guiding decisions all the time, sometimes wisely, sometimes unwisely. When they make them wisely, they manage to take into accurate account important facts about their own personalities and characters, their own capacities, aptitudes and resources. The sensible potential sybarite does not aim for the monastic life; the prudent potential ascetic does not aim for the life of sensuous and sensual abandon.

There is room in an idealized preference theory of a person’s good for still another sort of historicist condition along self-transformative causal paths, which bears on a person’s transformation of determinable maieutic ends into her own determinate final ends. A person’s final end is rational or authentic for her only if it developed, or could have developed, during the course of a process that took her from the possession of some maieutic end to her own

selection and articulation of her final end without her having made any errors of empirical fact about the crucial properties of her nature or circumstances, such as her own capacities, aptitudes, and resources, that bear on a person's ability to transform maieutic ends into final ends.

In focusing solely on epistemic factors, this developmental historicist theory sets only a necessary condition for the rationality or authenticity of a person's final ends. Further conditions will be needed to specify, say, control factors. Also, the scope of the condition is limited to the self-transformation involved in a maturing person's selection and articulation of maieutic ends into final ends. A more complete historicist theory of authentic preference formation would have to broaden the focus to include other processes of development.

Even in this incomplete formulation, the historicist condition helps to meet the deep incoherence objection Rosati and Sobel raise concerning the synchronic mutual incomprehensibility of radically differing ways of life, which allegedly prevents a single idealized advisor from encompassing them all as he plays his appointed role. Allowing the particular person's own epistemically idealized history to stand in for the synchronic ideal advisor of Sidgwick goes a long way toward solving an otherwise intractable problem. Such a history provides a historically particularized cognitive filter that fixes the rational status of the choices that transform a particular person's maieutic ends into her final ends. Thus, there is no need to posit a capaciously idealized advisor who considers all of the possible lives more or less open to his advisee.⁴⁴ Rosati and Sobel do take notice of the historical dimension of human lives, but they consider this to be a problem for idealized preference theories, rather than an opportunity for such theories to add a diachronic perspective, which casts particular historical paths in the role of idealized filters for the preferences of a developing person. Thus neither critic of idealized preference theories takes the historical dimension of human development seriously enough.

The historical condition sets a more modest, though still formidable, task for the advisee herself. She is to set about in young adulthood to render more determinate an initially merely determinable set of maieutic ends. This involves selecting possible careers, say, or ways of being a friend or a lover, from a range of determinate evaluative perspectives that are in fact suitable for her, given her aptitudes and circumstances. Once the process is well underway, along a path she chooses, certain other paths initially open to her will become closed to her later and thereafter, because she, as one factor among others, has made herself into a certain kind of person.

11. The Open Question Redux?

An especially appealing aspect of idealized preference theories is that, unlike certain other cognitivist accounts of practical reason statements, they are neu-

tral on substantive evaluative and normative issues, since they are limited to constraints of purely epistemic and instrumental rationality. Unlike some of their realist competitors, do not run afoul of the important core of truth that endures as the legacy of Moore's open-question argument, even when all the semantic niceties have been observed. In Nakhnikian's apt phrase, they avoid "fatuous validation by definition," even non-analytic definition of the sort associated with "new wave moral semantics."⁴⁵ The reason, again, is that the accounts by idealized preference theorists of the good and the right do not make essential reference to non-moral good-making and right-making properties, as primary property moral realists do.

However, Rosati disputes the substantive neutrality of idealized preference theories in a way that links the challenge to the problem of normative authority:

In order for us to be sure that we can regard the fully informed individual as authoritative, we must have a conception of what it would be for an individual's motivational system to change for the better, and thereby a more substantive conception of an ideal advisor – one that incorporates an ideal of the person. . . . There is no neutrality here.⁴⁶

Rosati concedes that idealized preference theorists do have some room for maneuver. They can limit themselves to providing an account of a person's good, given certain contingent fixed facts about her motivational system; they can thereby preserve both substantive neutrality and the internalism requirement. However, this comes at the considerable cost of preventing them from even acknowledging the existence of a further crucial concern, because "a legitimate question exists about why [a person] ought to accept [her] own fully informed verdict over the advice proffered by some other fully informed individual."⁴⁷ Even if idealized preference theorists do succeed in giving an account of a person's good, *modulo* some fixed features of her actual motivational set, they deprive her of the resources even to raise questions about what sort of person to be.⁴⁸

This gives rise to a version of Moore's open-question argument. A person realizes that she would embrace a particular preference structure, and thus be that sort of person, if she were to accept the advice of her own ideal advisor who began the process of epistemic purification with certain fixed elements in her prior motivational set. However, why should she take as the measure of what sort of person to be, the advice of an ideal advisor who is so complacent about the appropriateness of her prior motivation set, as an Archimedean point for fixing what sort of person she should be? It ought to be open to her upon reflection to try to become a kind of person whose most important features are beyond the advisory reach of any idealized advisor who begins the process of epistemic purification with fixed features of her current motivational psychology.

Yet again, Rosati's point strikes deep. The reply, again, is that historicizing the ideal preference program solves, or at least softens, the problem, without at the same time giving up what its proponents find so attractive about it in the first place. Two theoretical aspirations of idealized preference theorists set up this new open question, substantive evaluative neutrality and the internalism requirement. No neo-Humean idealized preference theorist will wish to give up the first by incorporating genuinely substantive ideals of a person into his conception of the rational constraints that constitute factual considerations as practical reasons. That way lies some kind of value realism, perhaps even Moore's type of intuitionism. Therefore, the first path must be avoided.

Whether or not the second aspiration must go by the board instead, as Rosati seems to suggest, depends on how we construe internalism. The brand in question here is existence internalism, which metaphysically grounds the existence of practical reasons in some motives of the agent.⁴⁹ The question is which ones. Rosati construes the requirement in a way that renders it incompatible with the very possibility that a person might engage in the kind of naturalized self-invention that sometimes results from a radical review of the kind of person she currently is, the viability of which is what is supposed to set up the new open question. But can a suitably designed form of idealized preference internalism accommodate radical self-invention? The answer is not easy, because the precise relationship of practical reasons to a person's actual motives in idealized preference theories is rendered unclear by the idealization itself. No sensible neo-Humean wishes to ground practical reasons for a person exclusively in the person's actual motives. Williams, whose view of internal reasons is influential, articulates the idea behind that reluctance in a useful way for present purposes. In his theory, a factual consideration counts as a practical reason for a person only if there is a possible route through rational deliberation from the person's actual motivational set to a motivational set that includes the factual consideration among its objects.⁵⁰ The important point is that the route can be long and circuitous, easily long and circuitous enough to provide room for the kind of naturalized self-invention that often results from someone's asking: "What kind of person should I be?"

A person's particular history of motivational development naturally arises here, because the process is fraught with opportunities for self-invention, as the young person proceeds to transform merely maieutic ends into her own more determinate final ones of adulthood. Therefore, historicizing idealized preference theories as we have considered goes some way toward accommodating the very aspect of human agency that Rosati argues is incompatible with idealized preference theories, self-invention driven from a particular evaluative point of view.

Rosati might doubt that whether a historicist idealized preference theory goes far enough. Consider her principal example, Sandy, a cautious, distrust-

ful, tidy person, who wonders whether she should be more like her spontaneous, messy, madcap friend Madelyn. As Rosati understands Sandy, Sandy is not concerned to know how her own cognitively idealized advisor would answer this question. For her idealized advisor merely summarizes the epistemic purification of Sandy's own psychological traits. What Sandy wants to know, by contrast, is whether she should leave those traits behind once and for all. Rosati's idea is that Sandy should not consult her own idealized advisor at all, but rather the non-idealized Madelyn.

These alternatives are not, however, really mutually exclusive. Suppose that by the time Sandy is ready to ask her life-challenging question she has already undertaken a cognitively unimpeachable history in which she has transformed determinable maieutic ends into her own final ends. Assume also that she has undergone a regiment of Brandtian cognitive psychotherapy that has effectively freed her from recognizably obsessive-compulsive preferences. Nonetheless, she still feels the need to ask the life-challenging question.

Given her history and her therapy, a practical reasons internalist will be puzzled. He will want to know how the epistemically upright Sandy forms the suspicion that her own rigidity and caution are preventing her from incorporating important values into her life, values which are readily available to the madcap Madelyn. An internalist will suspect that the story has been inaccurately told, that Sandy is not so epistemically upright after all, but has only now managed to discover about herself that she is already disposed at some accessible level of her psyche to be attracted to a path of daring spontaneity. Moreover, he will suspect that Sandy could have made this discovery on her own, without the actual example of Madelyn, if only she had taken a different course in her early history of transforming maieutic ends into final ends, or perhaps had encountered a more effective cognitive psychotherapist. The possibilities suggest that the alternatives Rosati offers Sandy are not mutually exclusive, and so idealized preference theorists can maintain both the substantive evaluative neutrality of the rational constraints in their analysis and a psychologically realistic version of the internalist requirement.

Rosati ends her most explicit discussion of idealized preference theories and open-question arguments on a negative, and decidedly Moorean, note:

a plausible form of naturalism about a person's good will need to account for our character as creatures who construct and guide ourselves by ideals of the person. Such naturalistic programs, when fully developed, are likely to be quite rich and complex. But I suspect that they may encounter a problem which would have us return to our starting point. The normativity of goodness for a person, I have argued, is in part a function of ideals of the person. Yet we disagree about which ideals to endorse. . . . But our disagreements about which ideals to embrace may not all turn on any factual, logical or linguistic mistakes. Thus, it may be open to a person to challenge

even naturalistic accounts of good for a person that incorporate an ideal of the person. And when she does so challenge such accounts, she will be asking Moore's question in just another form, "But why endorse this ideal?"⁵¹

Rosati regards this as a bleak prospect for any advocate of a new form of idealized preference naturalism who is prepared to inject evaluative ideals of the person into his account of practical reasons. However, it is excellent advice for a neo-Humean internalist, who is concerned to retain the evaluative neutrality of his theory, while also accommodating the signal facts about the human condition that Rosati so insightfully identifies: that it is constitutive of a person to have a particular evaluative point of view and to engage in suitably naturalized self-invention.

Historicizing idealized preference theories goes a long way towards accommodating these deep facts, while leaving intact the essential features of such theories, especially their evaluative neutrality and existence internalism. Another notable feature untouched is the relativist idea that Rosati sounds in her valedictory, when she remarks that "disagreements about which ideals to embrace may not all turn on any factual, logical or linguistic mistakes." A neo-Humean could hardly put it better and might, accordingly, will insist that this newest version of Moore's open-question argument has no bite when set against idealized preference theories in meta-ethics.

Notes

1. See G.E. Moore, *Principia Ethica* (Cambridge, England: Cambridge University Press, 1903).
2. George Nakhnikian, "On the Naturalistic Fallacy," in *Morality and the Language of Conduct*, Hector-Neri Castaneda and George Nakhnikian, eds., (Detroit: Wayne State University Press, 1963), p. 149.
3. See Connie Rosati, "Persons, Perspectives, and Full Information Accounts of the Good," *Ethics* 105, January 1995, "Naturalism, Normativity, and the Open-question argument," *Nous* 29, 1995, and "Brandt's Notion of Therapeutic Agency," *Ethics* 110, July 2000.
4. Moore, op. cit., p. 60.
5. See J.S. Mill, *Utilitarianism*, George Sher, ed., (Indianapolis: Hackett, 1979), ch. 2.
6. Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Indianapolis: Hackett, 1981), pp. 111–112.
7. Richard Brandt, *A Theory of the Good and the Right* (Oxford: Oxford University Press, 1979).
8. Roderick Firth, "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research* XII, March 1952, p. 329.
9. David Lewis, "Dispositional Theories of Value II," *Proceedings of the Aristotelian Society* suppl. vol. 63, 1989, pp. 113 & 116.
10. John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), p. 417; Peter Railton, "Moral Realism," *Philosophical Review* XCV, April 1986, pp. 173–

- 174; Michael Smith, "Internal Reasons," *Philosophy and Phenomenological Research* LV, March 1995, p. 109; and Bernard Williams, "Internal and External Reasons" in *Rational Action*, R. Harrison, ed., (Cambridge, England: Cambridge University Press, 1979), pp. 18–19.
11. See Gilbert Harman, *The Nature of Morality* (Oxford: Oxford University Press, 1977), pp. 19–20.
 12. See Nicholas Sturgeon, "Moral Explanations," in *Morality, Reason and Truth: New Essays on the Foundations of Ethics*, David Copp and David Zimmerman, eds., (Towota, N.J.: Rowman and Allanheld, 1985); Richard Boyd, "How To Be a Moral Realist," in *Essays on Moral Realism*, Geoffrey Sayre-McCord, ed., (Ithaca, N.Y.: Cornell University Press, 1988); David Brink, *Moral Realism and the Foundations of Ethics* (Cambridge, England: Cambridge University Press, 1989); and Berys Gaut, "The Structure of Practical Reason," in *Ethics and Practical Reason*, Garrett Cullity and Berys Gaut, eds., (Oxford: Oxford University Press, 1997).
 13. Nakhnikian, op. cit., p. 157.
 14. Sturgeon, op. cit., and Boyd, op. cit.
 15. Brandt, op. cit., p. 113.
 16. Ibid., pp. 11 & 113.
 17. Ibid., p. 113.
 18. Rosati, "Brandt's Notion of Therapeutic Agency," p. 784.
 19. See Allan Gibbard, *Wise Choices, Apt Feelings* Cambridge, Mass.: Harvard University Press, 1990), pp. 18–22, and Norman Daniels, "Two Accounts of Theory Acceptance in Ethics, in Copp and Zimmerman, op. cit., pp. 120–140.
 20. Gibbard, op. cit., p. 20.
 21. Brandt, op. cit., p. 113.
 22. Gibbard, op. cit., p. 21.
 23. Ibid., pp. 21–22.
 24. Ibid., p. 21.
 25. Gibbard, op. cit., ch. 4, esp. pp. 64–76.
 26. Ibid., p. 113.
 27. Ibid., p. 110.
 28. Ibid., p. 117.
 29. Ibid., ch. 6.
 30. Ibid., p. 144.
 31. Ibid., pp. 144–145.
 32. See Daniel Dennett, "Why the Law of Effect Will not Go Away," *Journal of the Theory of Social Behavior* 5, 1974.
 33. See Martin Hoffman, "Empathy, Role Taking, Guilt and Development of Altruistic Motives," in *Moral Development and Behavior: Theory Research and Social Issues*, T. Likhona, ed., (New York: Holt, Rinehart, Winston, 1976).
 34. Brandt, op. cit., p. 141.
 35. Hoffman, op. cit., p. 126.
 36. Brandt, op. cit., p. 130.
 37. Hoffman, op. cit., p. 135.
 38. Rosati, "Persons, Perspectives and Full-Information Accounts of the Good," p. 317.
 39. See *ibid.*, p. 307.
 40. Ibid., p. 323.
 41. David Sobel, "Full Information Theories of Well-Being," p. 809.
 42. See David Schmidtz, *Rational Choice and Moral Agency* (Princeton, N.J.: Princeton University Press, 1995).
 43. See Rosati, "Persons, Perspectives, and Full Information Accounts of the Good," pp. 307–311.

44. See *ibid.*, p. 310, note 38, and Sobel, *op. cit.*, p. 808.
45. See Mark Timmons and Terry Horgan, "Troubles for New Wave Semantics: The Open Question Argument Revived," *Philosophical Papers* 21 (1992), pp. 153–175.
46. Rosati, "Persons, Perspectives and Full-Information Accounts of the Good," p. 312.
47. Rosati, "Naturalism, Normativity, and the Open-question argument," pp. 58–59.
48. *Ibid.*, p. 59.
49. See Stephen Darwall, "Reasons, Motives and the Demands of Morality: An Introduction," in *Moral Discourse and Practice: Some Philosophical Approaches*, Stephen Darwall, Allan Gibbard and Peter Railton, eds. (Oxford: Oxford University Press, 1997), pp. 305–312. Also see, Rosati, *ibid.*, p. 49.
50. See Williams, *op. cit.*, p. 19.
51. Rosati, "Naturalism, Normativity, and the Open Question Argument," pp. 63–64.