Forthcoming in *The Naturalistic Fallacy*,
Cambridge University Press, 2019.
Neil.sinclair@nottingham.ac.uk

1. The Naturalistic Fallacy and the History of Metaethics

Neil Sinclair

The Cambridge philosopher G.E. Moore is a central figure in the development of analytic philosophy and his charge of "the naturalistic fallacy" the first great monument of analytic metaethics. The connections here are deep, since the alleged fallacy cannot be adequately understood without reference to Moore's wider views on semantics and metaphysics – views integral to the history of the analytic tradition.[1] Both types of influence are exemplified in the book which was the culmination of Moore's early work: *Principia Ethica* (first published in 1903). Although *Principia* addressed questions of both meaning and substance in ethics, Moore's strict separation of the two laid the ground for the modern subdisciplinary division between metaethics and normative ethics. What follows explores the derivation of this division and the subsequent influence of the fallacy on metaethics.

I. *Principia Ethica*

The principal claims which Moore argues for in *Principia* are as follows (the italicized labels are mine):

(1) *The Primacy of Good.* The common element to all ethical judgements is their concern with the predicate "good" (or its converse "bad") and hence the core question of the "science" of Ethics is: *What is good?* (1993: §§1-2, 89). Other ethical predicates can be distinguished from, yet defined in terms of, "good" (§§17, 89, 101-108).

[1] Skorupski 1993: ch.4.

(2) *The Separation Thesis*. The question *What is good?* is ambiguous between the further questions: (a) *What kinds of things are good in themselves?* (b) *What kinds of things are (causal) means to things good in themselves?* and (c) *How is "good" to be defined?* This last is "the most fundamental question in all Ethics" (§§3-4, 15-17).

(3) *The Indefinability of Good.* The answer to (2c) is that good cannot be defined. Rather it is simple (i.e. not composed of any parts) and unique. To suppose that "good" can be defined is to commit "the naturalistic fallacy" (§§5-14).

(4) *Anti-naturalism.* Naturalist theories of ethics – which hold that (a) the kinds of things which are good in themselves are those which possess some natural property and (b) this is true because good can be defined by reference to such a property – commit the naturalistic fallacy and so should be rejected (§§24-28). Such theories include Rousseau's recommendation to live "according to nature" (§27) and Spencer's "evolutionary hypothesis" that conduct is good in proportion to how evolved it is (§§29-34).

(5) *Anti-hedonism.* Hedonism – the view that only pleasure is good in itself – is often a naturalist theory of ethics and as such commits the naturalistic fallacy (§§36-44). In contrast to Bentham and Mill, Sidgwick was a hedonist who did not commit the fallacy, but his version of hedonism is implausible for independent reasons, principal among them being its failure to accommodate the value of beauty (§§50, 53).

(6) *Anti-metaphysicalism.* Metaphysical theories of ethics – theories which hold that (a) the kinds of things which are good in themselves exist in some supersensible reality and (b) this is true because good can be *defined* by reference to such a reality – also commit the naturalistic fallacy (§§66-67). Guilty parties include the Stoics, Kant, Spinoza, and Green (§§70-84). The naturalistic fallacy applies equally to naturalistic and metaphysical theories, though for convenience it is given "but one name" (§25).

(7) *Conservativism.* The answer to question (2b) depends both on empirical facts about the typical effects of actions and ethical facts about which things are good in themselves. Nevertheless it is likely that following the rules of conventional morality (for example those forbidding stealing and murder) have such effects (§§86-109).

(8) *Axiological Pluralism.* There is a plurality of kinds of things which are good in themselves, including pleasure and beauty (§§50, 113). However: "By far the most valuable things, which we know or can imagine, are certain states of consciousness, which may be roughly described as the pleasures of human intercourse and the enjoyment of beautiful objects" (§113).

(9) *The Principle of Organic Unities*. The intrinsic value of a complex whole (such as taking pleasure in a beautiful object) "*bears no regular proportion to the sum of the values of its parts*" (§§18-22, 112).

(10) *The Isolation Test and Intuitionism.* In order to establish which things are good in themselves "it is necessary to consider what things are such that, if they existed *by themselves*, in absolute isolation, we should yet judge their existence to be good" (§112, see also §§50, 55, 57). Furthermore, propositions concerning which things are (in general) good in themselves are incapable of direct proof or disproof. For these "no relevant evidence whatever can be adduced: from no other truth, except themselves alone, can it be inferred that they are either true or false" – and in just this sense they are "Intuitions" (Moore 1993: 34, see also §§45-46, 49).

Other notable claims in *Principia* include the incoherence of both ethical egoism and utilitarianism (§§58-64), the falsity of psychological hedonism (§42), the impossibility of intuitions of duty (§90), the asymmetry between the positive value of pleasure and the disvalue of pain (§§127-8) and the primacy of truth over system when constructing ethical theories

(§134). Moore also explores connections between these various claims. *The Separation Thesis* is key to making progress in other areas (1993: 33). *Anti-naturalism, Anti-hedonism and Anti-metaphysicalism* all flow from *the Indefinability of Good*, and are used by Moore to illustrate and elucidate the naturalistic fallacy. *The Indefinability of Good* lays the ground for *Axiological Pluralism* since, Moore holds, an important reason why previous philosophers favoured monism was their implicit assumption that good can be defined in terms of a single property (§24). Finally the careful application of the *Isolation Test* supports the *Principle of Organic Unities* (§112), which in turn can help diagnose the mistakes of hedonism (§§55-56).

Despite this internal coherence, the influence of *Principia*, was, and continues to be, fragmented. To understand it, one must remember that many of the claims defended in *Principia* were not new with Moore.[2] Elements of the *Primacy of the Good* were previously defended by Rashdall, Brentano and Moore's Cambridge teacher McTaggart. Moore attributes the *Indefinability of Good* to Sidgwick (§§14, 25).[3] Brentano endorsed axiological pluralism and McTaggart defended the intrinsic value of love (in the preface to *Principia*, Moore acknowledges similarities with Brentano, noting that he read Brentano's work only after *Principia* was finished). And the *Principle of Organic Unities* was accepted by Bradley, albeit embedded in an Idealist metaphysics against which Moore – followed by his Cambridge contemporary Russell – famously led the charge.[4] The influence of *Principia* is therefore not wholly explicable in terms of the originality of its ideas. Rather the explanation must invoke something of the manner, or perhaps timing, of their presentation. British philosophy at the turn of the 20th Century was undergoing a paroxysm of Idealism (exemplified by Green, Bradley, and McTaggart), but there was a certain weariness to these debates: as Skorupski puts

[2] For discussion and references see Hurka 2003.
[3] As Baldwin notes (1990: 69) indefinability can be traced back past Sidgwick to Price. Note that Sidgwick's text suggests the indefinability of ought rather than good.
[4] Moore 1903.

it, the time was ripe for a "re-imposition of distinctions...[and] a reassertion of the world's independence from the knowing subject" (1993: 129). Moore's brand of non-theistic Platonic realism – applied to the ethical case in *Principia* (and by Russell to mathematics) – was timely in this regard. (See Preti's chapter in this volume for more details of Moore's anti-psychologism, and Ruse's for Moore's appeal to secular Victorian sensibilities.) Elements of *Principia*'s style also explain its influence. There is the way it combines its individual theses into a single work that can thereby claim to settle *all* the fundamental questions of ethics. There is the sometimes blunt and manifesto-like way Moore puts forward his claims, especially those concerning intrinsic value (this bluntness itself being explicable in terms of Moore's intuitionistic methodology).[5] There are bold sweeping denunciations of most previous ethicists (except Sidgwick), whose theories are often rudely accused of being not just false, but irrelevant or incoherent (§§40, 84, 106). And there is the title itself, chosen to emphasise *Principia's* foundational ambitions.[6] It is perhaps no surprise that many of Moore's disciples bought into his own hype, according to which *Principia* both exposed the mistakes in all previous systems of ethics, and laid foundations for all future ethical theorising.[7] Recent commentators (e.g. Hurka 2003) may have expressed doubts about the legitimacy of Moore's influence on ethics, but its actuality is beyond question.

Initially though, the most significant influence of *Principia* was outside of academic philosophy, on the members of the Bloomsbury Group of writers and artists, including Lytton Strachey, Leonard and Virginia Woolf, Desmond MacCarthy, E.M. Forster, Clive Bell and John Maynard Keynes.[8] Virginia Woolf once asked a friend "did you ever read the book that

[5] Baldwin 1990: 129.
[6] Baldwin 1990: 68.
[7] Baldwin 1990: 66.
[8] Baldwin 1990: 132-4, Regan 1986.

made us all so wise and good: *Principia Ethica*?"[9] The wisdom-imparting thesis here was *Axiological Pluralism*, and more particularly Moore's extolling of the value of the admiring contemplation of beauty and romantic love. This occurs in chapter VI on 'The Ideal', described by Russell as "the best in the book" and much later by Hurka as Moore's "most brilliant piece of ethical writing".[10] Moore's essentially personal, somewhat unworldly and apolitical ideal was summed up by Bell thus:

> The best man – if good mean anything – will be he who is capable of the best states of mind and enjoys them longest. It is among artists, philosophers, and mystics, with their intense and interminable ecstasies of contemplation and creation, that we must look for our Saints.[11]

Other *Principian* claims had less intense but more long-lived influence. The *Primacy of Good* was the spur for much debate about the logical structure of ethical concepts – in particular, debate concerning which, if any, is fundamental.[12] The *Separation Thesis* ushered in the now orthodox distinction between metaethics and normative ethics. The *Principle of Organic Unities* connects with the development of holistic theories of reasons.[13] The *Isolation Test* has become a standard method in axiology, evident, for example, in Sylvan's "Last Man" argument for the value of nature.[14] And Moore's ethical epistemology has been influential on various intuitionist schools, starting with Ross at Oxford.[15]

---

[9] From Baldwin 1990: xiii.
[10] Cf. Baldwin 1990: 129, Hurka 2015.
[11] Quoted in Baldwin 1990: 134.
[12] See Broad 1930, Ewing 1947, Ross 1930, and for modern discussions Danielsson & Olson 2007, Hurka 2003, and McHugh & Way 2016.
[13] Dancy 2006.
[14] Sylvan 1973.
[15] Ross 1930, cf. Huemer 2005.

Other parts of *Principia* were less influential. And some were later abandoned by Moore himself. For example in his *Ethics* – published in 1912 – Moore rejects the *Primacy of Good* and narrows the scope of axiological pluralism. But by far the most influential claim of *Principia* within academic philosophy was *The Indefinability of Good* and the accompanying charge of naturalistic fallacy.[16] Although Moore later admitted that what he said about this point in *Principia* was "a mass of confusions", he nevertheless held it to contain a core that was "true and important" (1993: 3). In an unfinished preface to the (abandoned) second edition of *Principia* it is this issue to which Moore devotes most space. To further understand this point and its influence it is first necessary to examine the fallacy in detail, and to elucidate its connections with the supporting "open question argument" and Moore's general semantic views.

II. The Naturalistic Fallacy, Definition, and Analysis

It is often said that the naturalistic fallacy is neither a fallacy nor naturalistic.[17] And there may be some truth in this. As Moore admits, the fallacy does not name a bad type of *reasoning* (such as confusing the "is" of predication with that of identity), nor a view which is the exclusive preserve of ethical naturalists (1993: 20; §§12, 25). But nevertheless it does name a view that is (according to Moore) both false and often based on a fallacious type of reasoning. And it does name a view that was (in Moore's time at least) most *conspicuously* held by ethical

[16] Given the previous point, the scope of Moore's claims of indefinability and fallacy were fairly quickly extended beyond the narrow notion of good, to other moral (or evaluative or normative) notions such as right and ought. For ease of exposition, I will stick to goodness. For one instance of this generalization, see FitzPatrick's chapter in this volume.
[17] Frankena 1939: 464, Williams 1985: 121.

Neil.sinclair@nottingham.ac.uk

naturalists. (I should note that even the foregoing is not completely uncontroversial – see Feldman and van Roojen's contributions here.) But in any case the label which Moore invents in the early sections of *Principia* has stuck, and it is to those sections that one must look first in understanding the alleged fallacy.

In §6 Moore writes:

> If I am asked 'What is good?' my answer is that good is good, and that is the end of the matter. Or if I am asked 'How is good to be defined?' my answer is that it cannot be defined, and that is all I have to say about it. But disappointing as these answers may appear, they are of the very last importance.

This is Moore's first statement of the *Indefinability of Good*. To deny this claim – that is, to hold that good *is* definable – is to commit our fallacy, which Moore introduces in §10:

> But far too many philosophers have thought that when they named [properties possessed by all things which are good] they were actually defining good; that these properties, in fact, were simply not 'other,' but absolutely and entirely the same with goodness. This view I propose to call the 'naturalistic fallacy' and of it I shall now endeavour to dispose.

Initially then, the naturalistic fallacy can be understood as the claim that good is definable – a "fallacy" because it is false that good is definable. But we must remember Moore was writing more than a century ago, and we cannot assume that his understanding of terms like "good" and "definition" fit modern frameworks. To properly understand the fallacy we need to consider the following questions. (i) What is this good which Moore took to be indefinable? (ii) What, in Moore's sense, is it to *define* a thing like good?

(i) It is clear that when Moore claims that good is indefinable, he is not talking about the word "good". He writes that his concern is "solely with that object or idea...that the word is generally used to stand for" (§6). In the preface to the second edition Moore expresses considerable scepticism about the claim that the type of good he is concerned with is the one picked out by *common* usage of the word, but he still holds that there is an *important* (ethical) use of the word in which it stands for the object or idea of good, and that this idea is his primary concern (1993: 4-5). Two further queries arise: How is the object or idea of good distinguished from others? And what does it mean to say that good is an *object* or *idea* anyway? The first query – in effect, the question of what makes ethics distinctive – seems difficult for Moore, since his *Indefinability* thesis debars him from distinguishing good by offering a definition. Nevertheless Moore does sometimes treat "good" as synonymous with other locutions, as when he writes: "I have tried to shew exactly what it is that we ask about a thing, when we ask whether it ought to exist for its own sake, is good in itself or has intrinsic value..." (1993: 34; see Baldwin 1990: 77-80). As this passage demonstrates, Moore's use of "good" is typically non-attributive, so objects or states of affairs are good *in themselves*, not good *as* an example of a type (whose mention imports standards for evaluation – see Geach 1956). In the preface to the second edition Moore claims that the sense of good he is concerned with is the sense that can be "specified" by its "unique and fundamentally important relation" to "the conceptions of 'right' and 'wrong'" (1993: 4). The use of "specified" here suggests that Moore is trying to mark his quarry not by defining it, but by getting his readers to think thoughts involving it. The answer to the second query – What does it mean to say that good is an *object* or *idea*? – requires introducing some of Moore's wider philosophical views. In *Principia* Moore also calls good a "notion", "predicate", "property", "quality", "meaning" and "sense". To the modern ear the interchangeable use of these terms glosses many important distinctions, most obviously

between sense and reference.[18] On a popular Fregean view, senses or concepts are elements of thoughts (or propositions which minds grasp) whereas properties or qualities are elements of reality – both are distinct from linguistic types such as the word "good". Deploying this framework generates an uncomfortable ambiguity in the Moorean claim that good is indefinable, for this may be a claim about our concept of **good**, or about the property of goodness.[19] But in *Principia* itself there is in fact no such ambiguity, because Moore rejects the Fregean distinction between sense and reference. For Moore, sense *is* reference; the meaning of a word just is the object it refers to.[20] Thus the meaning of the word "good" just is the property of goodness. To say that good is indefinable is to say that this property (which is also the sense or meaning of the term "good") cannot be defined.

(ii) What is it to *define* a property like good? Armed with the Fregean distinction, the modern mind is apt to associate "definition" with conceptual analysis – whereby a particular concept is decomposed into constituent concepts – and to distinguish this clearly from any claims about identity, composition or constitution relations between properties. But given Moore's rejection of the sense/reference distinction, this framework does not fit *Principia* (as Baldwin notes (1990: 45), Moore does not describe good as a "concept" there, except for once in the preface). For Moore, there is no distinction between the semantic analysis of sense and metaphysical decomposition of referents: to define a property just is to identify "the parts which invariably compose" it (§10; see Baldwin 1990: 61-65). To say that good cannot be defined is therefore to say that the property of goodness (which is also the meaning of the term "good") cannot be identified with any group of simpler properties which compose it. This, at least, seems to be the most straightforward interpretation of *Principia's* claim that good is indefinable

[18] Frege 1892.
[19] I use bold when speaking of concepts, affix "ness" when speaking of properties, refer to terms like "good" using inverted commas, and speak of plain old good when not wishing to beg any interpretative questions.
[20] Moore 1899, cf. Baldwin 1990: 40-47.

(§§8, 12-13). Moore adds that there is a close connection between the claim of indefinability and that of unanalysability, since:

> ...we cannot define anything except by an analysis, which...refers us to something, which is simply different from anything else, and which by that ultimate difference explains the peculiarity of the whole which we are defining: for every whole contains some parts which are common to other wholes also. (§10)

Thus the Moorean sense of "analysis" leans towards analogies with scientific analysis (§8) and away from analogies with linguistic same-saying (1993: 9; §6).

On this straightforward interpretation, the naturalistic fallacy is just the claim that good is definable, that is, that goodness can be identified with a group of simpler properties which compose it (and thereby explain its distinctiveness). Since it concerns primarily properties, this might be called the "ontological form" of the fallacy. Unfortunately (as Moore himself notes: 1993: 17), *Principia* does not consistently present one version of the fallacy, even if that fallacy is understood purely ontologically. There are at least three distinct theses that, on the basis of *Principia*, might lay claim to the title of "the naturalistic fallacy". They are:

(a) Good is definable (or analysable).

(b) Good is identical with some property other than good.

(c) Good is identical with some natural or metaphysical property.

Each interpretation can be given textual support. I have already quoted Moore's first use of the term "naturalistic fallacy", which suggests (a) (see also §§12, 36, 46). (b) is suggested by *Principia's* famous epigraph ("Everything is what it is, and not another thing") and passages such as the following: "the [naturalistic] fallacy...consists in identifying the simple notion

which we mean by 'good' with some other notion" (§35, see also §§12, 25, 36, 84, 104). And (c) is suggested by: "[The naturalistic] fallacy...consists in the contention that good *means* nothing but some simple or complex notion, that can be defined in terms of natural qualities" (§44, see also §§25-26). It is also clear that (a), (b) and (c) are logically distinct. Good could be identical with some simple (i.e. partless) property without being definable – since definition requires decomposition. Hence (a) is distinct from (b) and (c) (Moore 1993: 14). Good could also be identical with some property that is neither natural nor metaphysical. Hence (b) is distinct from (c). As to the question of which of (a), (b) and (c) is the correct way of understanding the fallacy, Baldwin counsels despair: "Moore's discussion is hopelessly confused on this matter" (1990: 70). Still, Baldwin later suggests that it is (c) which predominates, and the same view is defended by Feldman in his contribution here.

This interpretation is given some credibility by Moore's own attempt to clarify these issues, in the preface to the second edition of *Principia* (although it is an open question how far this preface is *simply* a clarification of claims made in *Principia*). This preface clearly suggests that the naturalistic fallacy involves an error of *identifying* good with something other than good. But it is not the gross logical fallacy of taking distinct things to be identical. Rather it is a fallacy of identifying good with some member of a *particular class* of properties, namely those which are "natural or metaphysical". Thus Moore writes:

> This proposition, then, that [good] is different from any natural or metaphysical property, is one which I still think to be true and important: and I think it comes much nearer to what I now really want to say about [good], than does the proposition that [good] is unanalysable. (1993: 15)

However, this raises a further problem. For if the claim that Moore wants to make is that good is not identical with any natural or metaphysical property, it behoves him to define

the categories of natural and metaphysical properties. In *Principia* itself Moore begins by noting that "By 'nature,'...I...mean...that which is the subject-matter of the natural sciences and also of psychology. It may be said to include all that has existed, does exist, or will exist in time". He goes on to suggest that the test for a natural *property* relates to this second condition of existence in time: natural properties are such that their instances are located in space and time and are independent of the objects which possess them: "If they were all taken away, no object would be left, not even a bare substance: for they are in themselves substantial and give to the object all the substance that it has" (§26). In the preface to the second edition Moore regards this definition as "hopelessly confused", instead reverting to the original condition in *Principia* and defining a natural property as one "which it is the business of the natural sciences or Psychology to deal, or which can be completely defined in terms of such" (1993: 13). The definition of a metaphysical property is one "which stands to some supersensible object in the same relation in which natural properties (as now defined) stand to natural objects" (*ibid.*), where supersensible objects are those that are not objects of perception (or inferable from objects of perception by induction) and which, if they exist at all, have a non-spatio-temporal existence – numbers and Kantian noumenal selves being Moore's paradigm examples (§§66, 76). As Moore admits, these conceptions of natural and metaphysical properties are still deficient, because it remains unclear which properties are the business of the natural sciences and psychology (1993: 15).

It is at this point that Moore's text begins to manifest doubts about framing his key claim in terms of the non-identity of good with any natural or metaphysical property. He writes:

> It is, I think, far more certain that [good] is not identical with any natural or metaphysical property *of a certain limited class*, than it is not identical with any *whatever*. I wish, therefore, to define a limited (but still very wide) class of

> natural and metaphysical properties, and to insist only that [good] is not identical with any of *these*. (1993: 15)

What is this more limited class of properties? Moore seems to settle on the class of "contingent" or "intrinsic" properties. Thus Moore's important claim emerges as this: that good is "neither a contingent property nor yet an intrinsic one" (1993: 22). The naturalistic fallacy, as before, the fallacy of denying this claim. Thus a fourth claimant to that title emerges:

> (d) Good is identical with some contingent or intrinsic property.

Finally, a contingent property, says Moore, is one that can be possessed by some instances of a kind of thing without being possessed by them all and intrinsic properties are those which "tell you something about the *intrinsic nature* of things which possess them" (1993: 23). But it is unfortunately at this point that the unfinished nature of the preface to the second edition starts to undermine coherent interpretation. It *is* clear that Moore wants to claim the non-identity of goodness with a *certain class* of natural and metaphysical properties; it is less clear precisely what this class is.

Two final points emerge from the preface to the second edition. (See Rosati's chapter for further discussion). First, after noting that his own talk of "fallacy" is potentially confusing, Moore offers his final, full-dress, definition. This version is designed to accommodate all his foregoing clarifications (and also manifests Moore's characteristic late-period hesitancy and excessive qualification):

> I should, if I still wished to use the term 'naturalistic fallacy,' propose to define 'So and so is committing the naturalistic fallacy' as meaning 'He is *either* confusing [good] with a predicate of the kind to be defined [i.e. contingent or intrinsic] *or* holding it to be identical with such a predicate *or* making an

> inference *based* upon such a confusion,' and I should expressly point out that in so using the term 'fallacy' I was using it in an extended, and perhaps improper, sense. (1993: 21)

In his contribution to this volume, Feldman suggests that Moore is too hard on himself here, and that there is nothing improper about using the term "fallacy" to label a falsehood (as opposed to an error of reasoning).

Second, perhaps the most curious thing about the preface to the second edition is that the claim which Moore considers to be most important – and important insofar as supporting or entailing it is taken to be a constraint on interpreting *Indefinability* and the naturalistic fallacy – is not explicitly presented in the original *Principia.* This is a claim, not about the *definability*, *unanalysability* or *identity* of good, but rather about its *grounds*. Moore writes: "…[good is] a property which depends only on the intrinsic nature of states of things which possess it. This proposition…seems to me to be perhaps *the* most important of those which I am anxious to assert about [good]…" (1993: 6; see also 16). This view has clear connections with the *Isolation Test* and *Principle of Organic Unities*. It also talks about a particular *dependency* which was close to the surface when Moore temporarily defined natural properties as *in*dependent of the objects which possess them.[21] Though Moore spelled out his dependency claim in more depth in his 1922 paper "The Conception of Intrinsic Value", it is difficult not to interpret its rise to prominence here, in a preface supposedly clarifying *Principia*, as anything other than an injury-time bait-and-switch. But for present purposes I can leave this interpretative controversy hanging: we must remember that the preface to the second edition

[21] Cf. Baldwin 1990: 85, Darwall 2003b. It is worth noting the similarities between Moore's dependence claim and that of ethical supervenience, deployed by another Cambridge philosopher – Simon Blackburn – to support his own non-cognitivist position (Blackburn 1971). Moore taught Casimir Lewy, who Blackburn cites as the originator of the supervenience argument.

was not published until 1993, long after the influence of the naturalistic fallacy had been felt.[22] And it is this influence which concerns us here.

III. The Open Question Argument

The final sections of *Principia* that it is essential to understand before appreciating the fallacy's influence are those wherein Moore offers his argument to support the *Indefinability of Good*, or equivalently, the claim that the naturalistic fallacy *is* a fallacy (i.e. is false). This argument is the open question argument, although like "realism" and "metaethics", this phrase is not used in *Principia*. The argument manifests in three key passages.

In the first, Moore begins by equating the claim that good is definable with the claim that "disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole". He continues by arguing that this view

> …may be most plainly seen to be incorrect by consideration of the fact that, whatever definition be offered, it may always be asked, with significance, of the complex so defined, whether it is itself good. (§13)

In a later passage in the same section, Moore returns to the issue of the significance and distinctness of certain questions. Using as an example the proposed hedonist definition of "good" as "pleasant", he argues:

> But whoever will attentively consider with himself what is actually before his mind when he asks the question 'Is pleasure (or whatever it may be) after all

[22] A synopsis also appeared in Lewy 1970.

> good?' can easily satisfy himself that he is not merely wondering whether pleasure is pleasant. And if he will try this experiment with each suggested definition in succession, he may become expert enough to recognise that in every case he has before his mind a unique object, with regard to the connection of which with any other object, a distinct question may be asked.

The natural way of understanding this argument is as applying a general condition on the acceptability of definitions to the particular case of defining good. Sticking close to Moore's own formulation, this generates the following argument:

P1. If F is definable in terms of G, then the question "Is Gness F?" is not significant.

P2. For any proposed definition of good in terms of some natural or metaphysical property N the question "Is N good?" is significant.

Hence

C1. Good cannot be defined in terms of any natural or metaphysical property N.

This is sufficient to establish the negative part of the *Indefinability of Good*. The further positive claim is established by the argument implicit in the following passage:

> In fact, if it is not the case that 'good' denotes something simple and indefinable, only two alternatives are possible: either it is a complex, a given whole, about the correct analysis of which there may be disagreement; or else it means nothing at all, and there is no such subject as Ethics. (§13)

Which can be rendered as:

P3. If "good" is not definable in terms of any natural or metaphysical property then either it denotes something simple and indefinable or else it means nothing at all.

P4. "Good" means something.

Hence

C2. "Good" denotes something simple and indefinable.

There are a number of interpretative issues worth mentioning here (although I will not resolve them). First, there is the slide between talk of goodness as a property (P2) and "good" as a term (P3, P4). Second, Moore's concept of "definition" can be understood as elucidated above (although it is applied here to terms as well as properties) and we have already seen Moore's shifting views on the class of "natural or metaphysical properties" which he wishes to distinguish from good. Unfortunately his thoughts on the nature of "significant" questions are equally elusive. One interpretation based on Moore's examples is that a question is significant – or "open" – just in case it is not settled by the meanings of the terms involved, and conversely it is not significant – or "closed" – just in case it is so settled.[23] Finally there is the issue of the form of the questions whose significance is the focus. Though the quoted passages suggest that the question is asked of the proposed defining property itself, more commonly the argument is expressed as asking the question of an object which is admitted to possess that property. In other words the pertinent question is usually taken to be of the form "x is G, but is it F?" rather than the more textually accurate "Is Gness F?"[24]

[23] E.g. Frankena 1939, Miller 2013: 12.
[24] Rosati 2003; Joyce 2006: 150-51 is an exception.

Neil.sinclair@nottingham.ac.uk

Is the argument successful? There are some issues. Moore supports P2 purely on the basis of induction from examples.[25] There is also a worry that P2 is question-begging. If a significant question is one whose answer is not determined by meaning, then the claim that "Is N good?" is significant will only be accepted by those who have already determined that N *does not* give the meaning of good, that is, by those who already accept C1.[26] So the argument needs a better account of significance or "openness". One option is to say that a question is open just in case it is not immediately settled by introspecting on one's understanding of the relevant terms (an interpretation hinted at by Moore, for example, in §13). Unfortunately this renders P2 plausible only at the expense of undermining P1 – since deploying the same understanding of significance in P1 rules out the possibility of unobvious analytic truths (the so-called Paradox of Analysis).[27] A better Moorean option is that a question is open just in case we have "no difficulty imagining what it would be like to dispute it" (Darwall/Gibbard/Railton 1992: 117) or when we do not find it natural to guide our judgements in such a way that considers the question completely settled by meaning (Baldwin 1990: 89). Since we have no difficulty imagining what it would be like to dispute whether pleasure (etc.) is good, and do not find it natural to guide our judgements in such a way that the question whether pleasure (etc.) is good is analytically settled, then the relevant questions are open in this sense – so P2 stands. Defenders of this type of Moorean open question argument admit however, that it establishes only a *presumption* of the indefinability of good – for our lack of difficulty in imagining a dispute, or our reluctance to guide our judgements accordingly, may only be "evidence of the stubbornness of our attachment to an illusory conception of distinctive ethical meaning" (*ibid*.). In other words, though the relevant questions may be open in this extended

[25] Sturgeon 2003, Smith 1994: 27-29.
[26] Frankena 1939: 465, Hancock 1960, Miller 2013: 14-15.
[27] Snare 1975, Smith 1994: 37-39, Miller 2013: 15-16. See also Pigden's and van Roojen's chapters here.

sense, that they are so open is no longer, by itself, sufficient to show that their constituent terms are not interdefinable – since our sense that they are not may be illusory and apt for revision.[28]

Other worries with the open question argument concern its second half. Against P3, even if we accept that goodness cannot be defined in terms of (other) natural or metaphysical properties, it does not follow that it is not a natural or metaphysical property.[29] P3 also seems to commit Moore to what has sometimes been called the synonymy criterion of property identity: that two properties Gness and Fness are identical only if "G" and "F" are synonymous.[30] This in turn seems to follow from Moore's purely referential theory of meaning. Against this, it may be that even if "good" cannot be defined in terms of (is not synonymous with) "pleasure" or any other natural or metaphysical property-term, it nevertheless still refers to a property which is also referred to by a natural or metaphysical property-term. A model here is the identity of the properties of *being water* and *being H_20*, even though the related terms are not interdefinable.[31] On this view Moore's argument rules out any analytic property identity between goodness and natural properties (i.e. refutes "analytic" or "definitional" moral naturalism), but not a synthetic one (i.e. does not refute "synthetic" or "metaphysical" moral naturalism). (The prospects for analytic moral naturalism are discussed in Nuccetelli's contribution here, while the feasibility of the synthetic model is discussed by both FitzPatrick and Miller). Another problem with P3 is that it neglects the possibility that "good" is meaningful but does not denote any property at all – perhaps because it has non-cognitive or emotive meaning. Such views can still accept some truth in the open-question argument – viz. the point that fundamental ethical *concepts* or terms cannot be defined in natural or

[28] The idea of revisionary definitions is mooted by Ayer 1946, Hancock 1960; pursued in Brandt 1979 and Railton 1986b.
[29] Sturgeon 2003.
[30] Brink 2001: 155, Smith 1994: 27-35, Sturgeon 2003: 533.
[31] Boyd 1988, Durrant 1970, Kripke 1980, Putnam 1975. Given my earlier convention, technically I should speak here of "water-y-ness" and "H_20-ness", but that would be too much.

metaphysical terms – but deny any further implications. (This response is discussed by van Roojen's contribution here). Finally, P4 is hardly trivial, and will be rejected by those who take moral terms to be incoherent or irrevocably error-strewn. It is, however, worth noting that Moore spends far more time on the first part of the open question argument than the second. In *Principia* at least, his principal aim seems to be to defend the *Indefinability of Good* – the further cognitivist view that moral terms denote properties often seems an assumption underlying the whole discussion, rather than a conclusion of it (e.g. §§10, 13, 23).

IV. Metaethics After Moore

What then, of the influence the naturalistic fallacy and open question argument? I have already mentioned Moore's role in the genesis of metaethics. And while it would be overblown to describe the subsequent history of the subject as a series of footnotes to *Principia*, the book's influence is deep and enduring.

The history of 20th Century metaethics begins with the Moorean three-step.[32] The first step is Moore's claimed refutation of all naturalistic and metaphysical theories of ethics – views which commit the naturalistic fallacy. The second is the subsequent dissatisfaction with the non-naturalist view that Moore thought alone avoided the fallacy. This is the view that good is a simple non-natural property, to assert that something is good is to assert that it possesses this property (§§10, 13, 23) and that our knowledge of such properties is non-inferential (§§45-46, 49). The second step was characterised by worries that such a view gave too much by way of

[32] Baldwin 1990: 66.

metaphysics and not enough by way of epistemology.[33] The third step is the culmination of this dissatisfaction in distinctively un-Moorean non-cognitivist views of ethics, which reject the realism inherent in both of the previous steps while honouring the spirit of Moorean irreducibility. According to these views ethical judgements are not ascriptions of properties to objects at all, rather they are expressions of emotion or disguised prescriptions. The former, emotivist, variant was the house-style of the Vienna Circle and was subsequently refined by Ayer, Nowell-Smith, Urmson and Stevenson. The latter, prescriptivist, variant was developed by Hare.[34] While non-cognitivists could not accept the Moorean ontological claim that the property of goodness was indefinable in natural or metaphysical terms (since they denied the existence of ethical properties), they could accept the semantic view that the term "good" (or concept **good**) was indefinable in natural terms, since, they held, it performed an expressive function which was inconsistent with the attributive or descriptive function of natural terms. In other words, non-cognitivists could, and did, take pride in the fact that they avoided the "semantic form" of the naturalistic fallacy. Thus, it is no surprise that non-cognitivists also tended to rely on slightly reworked, "practical", versions of Moore's open question argument – characterised by the claim that the questions linking moral and non-moral terms remain open because moral terms have a practical import which non-moral terms lack. [35] The wide currency of the naturalistic fallacy in this period is demonstrated by Nowell-Smith's remark, in 1954, that the fallacy is "too well known to require exposition in detail" (1954: 32). This was also a period where ethics as a whole was dominated by metaethical concerns (as Ruse's contribution here attests).

[33] Ayer 1946, cf. Baldwin 1990: 101-106, Hudson 1967, Miller 2013: 24-33.
[34] Ayer 1946, Hare 1952, Nowell-Smith 1954, Stevenson 1963, Urmson 1968.
[35] Ogden & Richards 1923: 125, Stevenson 1937, Hare 1952. For discussion see van Roojen's contribution here, Baldwin 1990: 89, Miller 2013: 17-22, and Sinclair 2007.

This takes us up to the early 1960s, at which point it was assumed that the true beneficiaries of Moore's charge of fallacy were non-cognitivists.[36]At this time the debate between non-naturalistic intuitionists and non-cognitivists was commonly understood through the prism of the "is-ought problem", that is, the problem: "How is what *is* the case related to what *ought* to be the case – statements of facts to moral judgements?" (Hudson 1969: 11). Non-cognitivism was associated with the view that ought-statements were "logically distinct" from is-statements, insofar as the former cannot be reduced to, nor are they derivable from, the latter.[37] The two main rivals at this point were the Moorean intuitionist view that ought-judgements were a *sui generis* species of is-judgements, and the neo-Aristotelian view – exemplified by Foot – that ought-judgements were derivable from naturalistic premises. Participants in this debate (e.g. MacIntyre 1959, Searle 1964) often held that to commit the naturalistic fallacy was to do what Hume had counselled against: to derive an "ought" from an "is". Thus it seemed that non-cognitivists were the beneficiaries of Moore's arguments primarily because (i) non-cognitivism was associated with the non-derivability of "ought" from "is", and (ii) it was assumed that the naturalistic fallacy is the fallacy of deriving an "ought" from an "is".[38]

Moore's own connection to this "inferential from" of the fallacy is debatable. For one thing in *Principia* the fallacy is almost exclusively expressed in terms of "good" rather than "ought" (although occasionally Moore does take the fallacy to concern other normative terms – e.g. "right" (§14), "desirable" (§65) and "beautiful" (§121) – and the *Primacy of Good* makes the step from "good" to "ought" a small one). For another Moore himself seemed to accept

[36] E.g. Veatch 1966.
[37] Hare 1952: 29. Sometimes labelled "the autonomy of ethics" (Prior 1960).
[38] Frankena (1939: 465) was one of the earliest to draw a connection between the naturalistic fallacy and Hume's view on "ought" and "is".

necessary connections between natural properties and goodness.[39] For a third, the preface to the second edition never considers the inferential form of the fallacy, nor does *Principia* mention Hume. On the other hand, several passages suggest that the naturalistic fallacy is at least the source of a mistaken type of *inference*. This thought is near the surface in Moore's discussion of Spencer (§31) and explicit in the discussion of metaphysical ethics: "To hold that from any proposition asserting 'Reality is of this nature' we can infer, or obtain confirmation for, any proposition asserting 'This is good in itself' is to commit the naturalistic fallacy" (§67). There is also the fact that in the preface to the first edition, Moore states that the aim of *Principia* is to investigate "the nature of the evidence, by which alone any ethical proposition can be proved or disproved" (1993: 34). Moreover *if* one has succumbed to the ontological and semantic form of the fallacy – for example if one has identified *goodness* with *pleasantness*, and **good** with **pleasant** – then one will be disposed to succumb to the inferential form – for example to infer that contemplation is good on the basis that it is pleasant. (The converse connection between the forms of the fallacy does not seem to hold.) But issues of interpretation aside, there is little doubt that the inferential version of the fallacy was hugely influential in the 1960s, and remains so today, particularly amongst evolutionary ethicists. (For examples see Joyce (2006: 146). The chapter by Ruse here continues this tradition, Feldman considers and rejects the inferential interpretation and Pigden traces the connections between the naturalistic fallacy and the supposed ban on deriving an "ought" from an "is".)

Return to non-cognitivism. The common perception is that in the mid-1960s such views were decisively torpedoed by the second great monument of analytic metaethics: the Frege-Geach problem.[40] This is the problem of explaining how moral sentences can function in logical

[39] Baldwin 1990: 86 – although the connections here were synthetic rather than analytic.
[40] This is undoubtedly a simplification, since worries concerning non-cognitivism's irrationalist tendencies extended beyond Geach's narrow logical point.

contexts if they are understood on the non-cognitivist model.[41] In response (and after a resurgence of interest in normative ethics, spurred on by the work of Rawls and Nozick) the 1970s, 80s and 90s saw new brands of realist ethical naturalism developed. These views often incorporated contemporaneous developments in semantics and analysis, allowing them to reject the semantic assumptions underlying Moore's open question argument.[42] One version – so called "new wave moral realism", a version of synthetic moral naturalism – pursued the criticism of the argument mentioned above, viz., that of applying the sense/reference distinction to the ethical case and accepting the indefinability of ethical concepts while denying the non-identity of ethical and natural properties. Such views therefore avoided the semantic form of the naturalistic fallacy while denying that the ontological form *was* a fallacy. Somewhat inevitably, Moore's charge of fallacy was later loosed from its original semantic moorings and pinned upon this new version of naturalism too. Thus we have the spectacle of 20th Century metaethics both opening and closing with the accusation that the naturalism of the day is undermined by the open question argument.[43]

In roughly the same period as these new naturalistic versions of realism were being developed, naturalists of a different, methodological, type continued to develop theories in the non-cognitivist tradition.[44] The twist was that proponents of these theories – sometimes going under the label "quasi-realist" – began to argue that their views need not take the revisionary forms of earlier versions, and that a fundamental non-cognitivist understanding of ethical language was consistent with many of the "objective" features of ethical practice highlighted by Moore and other realists (e.g. §73, see Glassen 1959, Sinclair 2012). Of course such theories

[41] Geach 1965, Searle 1962.
[42] Brink 1989, Boyd 1988, Jackson 1998, Putnam 1981: 206-208, Railton 1986b.
[43] Horgan & Timmons 1992. See Adams 1999 and Taylor 2016 for the same dialectic in a theological context, and Miller's contribution here for more context.
[44] Blackburn 1993, 1998, Gibbard 1990, 2003, Sinclair 2009.

were (and still are) obliged to say something in response to the Frege-Geach problem, but they continued to approvingly cite the Moorean claim that basic ethical terms are indefinable and sometimes offered reworked "practical" versions of the open question argument.[45] Such quasi-realist views have never had a period of dominance within metaethics, but have continued to be a thorn in the side for those enamoured of the Frege-Geach point. Their relationship with Moore's views has also not been straightforward. For while they have been pleased to avoid the semantic form of the naturalistic fallacy and even deployed Moore's claims about the dependence of good in arguments to support their position, critics have suggested that a close variant of the open question argument can be deployed against their view of the nature of ethical attitudes.[46]

Since the turn of the 21st Century, metaethics has seen a resurgence in non-naturalist and irreducible realist views with strong affinities to Moore's own. Again such developments have often been spurred by contemporaneous developments in metaphysics and semantics, which are applied to the ethical case. And again, such views are frequently supported by their coherence with the irreducibility of fundamental ethical concepts or properties.[47] Moore has also been cited as an influence on non-metaphysical versions of moral realism. These are views that take ethical judgements to be objective, truth-apt and knowledge-apt, yet not requiring for their truth any distinct realm of ethical existents.[48] The connection here is that in early work Moore had distinguished between spatio-temporal existence and abstract being, and suggested that the property of goodness (as opposed to its instances) *was* but did not *exist* (§66, see Baldwin 1990: 45-47). Three final notable trends in the increasingly multi-track landscape of 21st Century metaethics are constitutive approaches (which seek to ground ethical facts in

[45] Blackburn 1998: 50, 70, Gibbard 2003: 6.
[46] Baldwin 1990: 107-8. Cf. van Roojen's contribution to this volume.
[47] Shafer-Landau 2003: 56-58, Wedgwood 2007: 68-76, Enoch 2011: 100 n.1.
[48] Parfit 2011, Scanlon 2014. The connection is suggested by Hurka (2015: §2), Miller 2013: 10-11.

constitutive facts of agency), hybrid theories (which seek a middle ground between realism and non-cognitivism) and error theories (which accept the realists' semantics while rejecting their metaphysics).[49]

Though these views are now being debated more than 100 years after *Principia*, it is still common for them to be motivated by reference to Moore – and in particular to some version of the *Indefinability of Good*. *Principia's* greatest influence, therefore, has been not in its articulation of a completely worked-out positive theory in metaethics (for Moore's non-naturalism could scarcely be described as worked-out), but in the identification of an important desideratum by which any metaethical theory should be judged. This desideratum is that of recognising the indefinability of good, or in other words, avoiding the naturalistic fallacy.

In light of this, the metaphor with which this chapter began – describing the fallacy as a monument of analytic metaethics – is not entirely apt. Monuments are static and often wandered past. The naturalistic fallacy is more akin to a constraining force or frame for metaethics. Consider then, each phase of analytic metaethics to be a rectangular piece of card, with an irregular shape cut out. Each contributes its own distinctive form and colour (problems, concerns, options, desiderata). When laid one on top of another they reveal a composite bounded form – a distinctive reverse-silhouette. The revealed shape is the form of contemporary metaethics. And lying at the bottom of the pile, the frame which frames the others, is the naturalistic fallacy.

[49] For constitutivism see Rosati's contribution here. For hybrid views see Fletcher & Ridge 2014. For error theory see Olson 2014, Streumer 2017.

Forthcoming in *The Naturalistic Fallacy*,
Cambridge University Press, 2019.
Neil.sinclair@nottingham.ac.uk