Philosophy

RJP

The Richmond Journal of Philosophy

Issue Six Spring 2004

In this issue...

Richmond upon Thames College

The Richmond Journal of Philosophy

Issue six
Spring 2004

Philosophy Department
Richmond upon Thames College
Egerton Road
Twickenham
Middlesex
TW2 7SJ
United Kingdom

email: rjp@rutc.ac.uk
www.rutc.ac.uk/rjp

Contents

[Editorial]

Welcome to the sixth issue of the Richmond Journal of Philosophy. We have now entered the second year of publication and have hosted two successful conferences.

In the current issue we begin with a discussion by Peter Simons on the nature of metaphysics in the last century or so and some observations concerning future directions in which enquiry may develop. This is followed by a consideration of one of moral philosophy's fundamental questions: how are we to understand moral responsibility. In the first of a two part piece Garrath Williams tackles the issue of free will and adumbrates the approach of Aristotle to moral responsibility. In the next issue he will critically examine Kant's theory. Our third paper focuses on sleep, and in particular the challenges raised for our understanding of consciousness. James Hill's paper examines the views of three great early modern philosophers, Descartes, Locke and Leibniz. From sleep we move to trust, our understanding of which is discussed by Paul Sheehy. At the centre of Plato's *Republic* is an apparent puzzle. Why the philosopher would return to the cave having come to see the Good; why would the philosopher become king? D. J. Sheppard analyses this challenge and examines the problems it poses for Plato's account. Our final paper moves us to the twentieth century and the influential and widely challenged work of A.J. Ayer. Pierre Cruse sets out the key criticisms of the verification principle and explains how it might be defended.

Purpose of the Journal

The motivation for and ambition of the journal is to provide serious philosophy for students who are at an early stage in their philosophical studies. The style and content of the papers will be accessible to students who have yet to become hardened to the more technical and specialised journals of professional philosophy.

What do we mean by 'serious' philosophy? First, the content of the journal is not constrained by a remit to appeal to or reach the interested general public. Whilst the papers must speak to the needs of students who are relatively inexperienced in philosophy, they presuppose that their audience is actively engaged in philosophy. Second, the content is serious in its focus on the central areas of philosophy. The big or traditional questions of metaphysics, epistemology, and ethics will provide the journal's centre of gravity. The third way in which the philosophy is serious is through the scope, variety and depth of analysis that can be achieved by the accumulation of papers over time. Moreover, each paper is not simply an introduction to one of the main topics on A-level, IB or degree courses. Such papers will indeed have a role in the journal, but they will not be the only kind. Our contributors will be offering original papers based on their own research. The journal will be a forum for the kind of critical engagement and debate that characterise the practice of philosophy. The fourth way in which the philosophy is serious is in the contributors themselves. The vast bulk of the papers will be written by professional philosophers engaged in both research and teaching.

About the Editorial Board

Stephen Grant is a full-time lecturer in philosophy at Richmond upon Thames College. He has also taught at King's College London where he is completing his doctorate on the emotions. His main interests are in the emotions, ethics and political philosophy. He has published on the ontological argument.

Dr Paul Sheehy teaches philosophy at Richmond upon Thames College and King's College London. His main areas of interest are in metaphysics, political and moral philosophy and the philosophy of the social sciences. His doctoral thesis was on the ontological and moral status of social groups, and he has published papers on social groups, voting and explanation and realism.

Paul Sperring is head of the philosophy department at Richmond upon Thames College and an A-level examiner in philosophy. He completed his undergraduate and masters studies at The University of Warwick, studying both analytic and continental philosophy. He is currently working towards his PhD at Birkbeck College. His research interests are metaphysics and the philosophy of mind.

Peter Simons

Criticism, Renewal and the Future of [Metaphysics]

Let a hundred flowers bloom, let a hundred schools of thought contend.

Mao Zedong

Metaphysics, What

From the beginning of Western philosophy until the 18th century, the most important part of philosophy was metaphysics: other parts were subordinate to it. Aristotle, the editors of whose works gave us the name ‘metaphysics’, did not invent the subject, but he first clearly demarcated it, and signified its primacy over other parts of philosophy by the name he used, ‘first philosophy’. Metaphysics was first in two ways. Firstly, metaphysics set out the most general principles applying equally to all things, such as the principle that a thing cannot be both in a certain way and not in that way at the same time. Secondly, it provided an outline inventory of the entities available when describing any subject of interest, whether it be the basic constituents of matter, the kinds of living organisms, the forms of political organisation, the things investigated by mathematicians or worshipped by the devout. The first of these tasks naturally brings metaphysics into contact with logic, which investigates the formal principles of inference and necessary truth. The second brings it into contact with the specialised disciplines which examine the detail of what there is. Throughout its long and chequered history, metaphysics has wavered between these two poles, now emphasising logic, now the constraints of empirical science.

Criticism of Metaphysics

With the increasing complication of scientific knowledge in the 17th and 18th centuries, and the discovery that the world is not describable wholly in terms drawn from the experience and vocabulary of everyday life, it was perhaps inevitable that metaphysical claims should be subjected to criticism. Rationalist metaphysicians such as Spinoza and Leibniz claimed to know necessities about the world, but disagreed what these were, Spinoza holding there is but one substance, Leibniz holding there are infinitely many, and each claiming to prove his views from first principles. From Locke through Hume to Kant, more critical philosophers considered metaphysical knowledge claims should be subject to scrutiny as to their origins and reliability. The upshot of this critical movement was to topple metaphysics from its pedestal. Metaphysical claims about the nature of reality were to be subject to the same sort of critical scrutiny as any scientific hypothesis, and it turned out that they were far more fallible than had been imagined. German idealism constituted a backlash against this criticism, but its heyday was short, and its unscientific claims stirred an even fiercer backlash against metaphysics, which came at different times in different European countries in the nineteenth century. For much of the nineteenth century, ‘metaphysics’ was a dirty word. Like the British empiricists before them, critics such as Comte in France and Mach in Austria proposed a more modest role for theoretical philosophy, that of protecting science from wild metaphysical speculation. In the early twentieth century, this anti-metaphysical movement joined with two growing philosophies driven by a Kantian emphasis on critical method. On the one hand, Edmund Husserl’s phenomenology recommended abstaining from ontological commitments to uncertain ‘transcendent’ entities while knowledge claims were subject to scrutiny of the way in which they come up in our experience. By ‘purifying’ our claims of their existential weight, Husserl, imitating Descartes, promised to put science on a philosophically unshakable foundation. On the other hand, the logico-linguistic analysis created by Frege, Peirce, Peano and Russell in the service of the new mathematical logic were turned by Russell, Wittgenstein and the Vienna Circle into tools for analysing and showing the limits of our knowledge by showing the limits to what we can meaningfully put into words. The Vienna Circle, following Mach, went further and declared that metaphysical claims were literally meaningless, being unsusceptible to the kind of scrutiny that a scientific claim should sustain.

Renewal

Both of these anti-metaphysical programmes, that of phenomenology and that of logical positivism, collapsed under their own internal contradictions. Phenomenology postulated a primacy of subjective consciousness, transparently and infallibly describable, but was unable to sustain its claims because no Archimedean vantage point of pure description free of metaphysics was ever achieved. Logical positivism found its methodological strictures subject to its own negative criticism, and its unexamined metaphysical assumptions returned through the back door to undermine it.

To alert observers, these defects were obvious by mid-century. Later phenomenologists such as Ingarden and Merleau-Ponty abandoned Husserl's quest for a metaphysics-free foundation for science and accepted the best working hypothesis that there really is an external world with a plurality of things existing in it independently of human awareness. In analytic philosophy there were divergent reactions. Carnap, who had been most vociferous in rejecting metaphysics, became perfectly happy to posit numerous entities to play different semantic roles in his account of truth and meaning in the 1940s, while continuing to insist that no one such scheme gave the true inventory of reality. Carnap's relative indifference to the commitments of his semantic theory was later to be echoed by W. V. Quine. Quine criticised Carnap's sharp division between questions of meaning on the one hand and questions of fact on the other, proposing that we examine our best working theories and tease the metaphysics inherent in them out of an examination of the entities they quantify over when 'regimented' into the perspicuous medium of predicate logic. Quine's celebrated slogan 'To be is to be the value of a variable' drew attention again to the links between logic and ontology which had figured in the early analysis of Russell and Wittgenstein.

Wittgenstein himself had long abandoned any hope of a perspicuous correspondence between a neat logical language and a neat logical world, and his painstaking but unsystematised later philosophy examining the details of ordinary usage carried no general metaphysical message. Similar scrutiny of ordinary language was pushed forward by several philosophers in Oxford, notably Ryle, Austin and Strawson. Of these, Strawson introduced the term 'descriptive metaphysics' in the subtitle of his 1959 book *Individuals*. Descriptive metaphysics strives in Kantian fashion to capture those immutable aspects of the common human conceptual scheme which we all must share. Like Quine, Strawson considered that our use of such a scheme inevitably involves belief in a world of independently existing things, notably material bodies and psychophysical persons, and their qualities and relations. Categories like those of Aristotle were restored through attention to our daily linguistic practices, much indeed as Aristotle had himself found them, but tempered with a Kantian underpinning to forestall future criticism. Quine in the meantime had concluded that the fit between language and the world, mediated through our experience and linguistic behaviour, was sufficiently loose and multiply interpretable to make it impossible for us to draw definite metaphysical conclusions from any theory.

This ontological relativity sent Quine's views essentially back to those of Mach, where the role of everyday beliefs, and of science with its diverse theoretical posits was simply to make the most economical overall sense of our experience.

Both Quine and Strawson helped to revive metaphysics after the low point of positivism, but their ways with metaphysics should by rights have stopped the subject dead in its tracks once more: Quine's because ontology melted away again, Strawson's because nothing essentially new was left over to be done once the main points of descriptive metaphysics had been made. On the contrary, metaphysics has since the 1960s continued to grow in strength and confidence as a philosophical discipline. Hardly a month goes by without a new textbook or reader coming on the market, and metaphysical controversies resound through the professional journals with the liveliness of debate found in ancient Athens or medieval Europe. Why is this, and where will metaphysics go from here?

A Hundred Flowers

Up until about 1960 metaphysics had to overcome the methodological strictures of its phenomenological and logico-linguistic critics, and it did so by using the tools of its opponents, whether the search for a phenomenological basis, or for an adequate account of meaning and logic. Perhaps the ultimate version of the view that metaphysics and ontology are subordinate to considerations of language and meaning was the move by Michael Dummett to transpose questions about the mind-independent existence (realism) or otherwise (anti-realism) of objects of a disputed kind, into questions about the acceptability of the certain logical principles in the relevant area of discourse. The *realist* about such disputed entities as mathematical objects, or future events, or dispositional properties will be happy to accept that sentences about them may be true or false irrespective of our ability to actually decide the truth-value, whereas an *anti-realist* would confine the sentences accepted or rejected to those we could verify or falsify, and leave other sentences without truth-value. Dummett's subordination of metaphysical considerations to logico-linguistic ones maintains the contact between metaphysics, semantics and epistemology that characterises Strawson's and Quine's metaphysics in their different ways, but most subsequent metaphysical debate has been less subordinate to linguistic considerations. A milestone in this change was the revival of discussion of the venerable problem of universals by David Armstrong in 1974. Armstrong considered that the meaning of general terms has no relevance to the metaphysical question whether universals exist, and he preferred to rely on direct arguments. General terms might be meaningful whether or not there are universals, and the question as to which universals exist was to be settled by science rather than in the linguistic philosopher's armchair.

Metaphysics lives in contact with the special sciences, and they raise problems which leak outside their own boundaries and reveal a metaphysical side. Phenomenology grew out of Brentano's philosophical psychology, whereas analytic philosophy grew out of Frege's and Russell's attempts to provide a logical foundation for mathematics. After its establishment as a science in the nineteenth century, psychology went through its own methodological upheavals, rejecting the introspective methods of Wundt and Brentano in favour of behaviourism, a methodology embodying a deflationary philosophy of mind, influential in the middle analytic philosophy of Wittgenstein, Ryle and Quine. The first sign that metaphysics was reviving outside the logic and language laboratory came with the robust metaphysical materialism of Place, Armstrong and Smart in Australia. Debates about the nature of the mental and its relation to the physical, given its edge by Descartes' dualism, had been muted under the anti-Cartesian influences of Wittgenstein and Heidegger, and had been transposed into an issue of choice of scientific language by Carnap. Herbert Feigl's insistence that the mind–body problem is not a pseudo-problem but a genuine issue, together with the realisation that Ryle's dispositional analysis of mental acts would not work for the central cases of thinking and perceiving gave the Australian materialists their impetus. The Australian mind–brain identity theory, a metaphysical equation initially dismissed as naive by European sophisticates, led to a whole series of ever more varied and differentiated positions on the mind-body issue. At the same time the rise of computers and the prospect of genuine artificial intelligence, optimistically announced in 1950 by the computer pioneer Alan Turing, led to an increasing debate about the difference if any between human and (prospective) machine mentality, as well as the use of computer models such as the distinction between hardware and software to try and understand mind. The result was to put philosophy of mind at the centre of analytic philosophy, displacing logic and language from the central position it had occupied since the early twentieth century.

However, the strongest motivation for reviving metaphysics continued to come from considerations of logic and language. From the mid-century, logicians such as Arthur Prior in England, Georg Henrik von Wright in Finland and Saul Kripke in the USA developed modal logics, dealing with necessity, tense, belief, obligation and other subjects. The semantic analyses of modal logics, anticipated by Carnap, involved the Leibnizian idea of possible worlds. The mathematical success of such analyses convinced many that the use of logical semantics could complete for a wider range of vocabulary the rigorous analysis of language that Russell had begun, and replace the informal methods of ordinary language philosophy. Carnap's student Richard Montague extended this analysis to a wide range of features of natural language previously regarded as inaccessible to logical analysis. In the course of such analyses, logicians and linguists found themselves up to their ears in ontological commitments to times, worlds, and a host of abstract mathematical objects such as sets and functions, and in general were happy to do so.

The metaphysical high-water mark of this development was the modal realism of David Lewis, who claimed in his 1987 book *The Plurality of Worlds* that alternative possible worlds exist and are just as real as our own. Lewis confronted those who greeted this metaphysical extravaganza with an 'incredulous stare' by challenging them to find an alternative account promising equal expressive power. That dispute continues unabated today.

Lewis's ontology for the semantics of modality forced others who rejected it to offer alternative semantic accounts with alternative ontologies. Those utilising only actually existing entities were called *actualist*.

Some actualist views employed states of affairs as substitutes for possible worlds, or took them to be the objective items making modal propositions true. Another name for states of affairs is *situations*, and they were made the ontological basis of a wide-ranging semantic account of natural language, going under the name of *situation semantics*. This view sidestepped the popular model-theory-inspired semantics deriving from Tarski and Carnap in favour of an account in which the participants in a linguistic exchange are concretely embedded in the situations on which they comment and which serve to give their utterances meaning.

Truthmakers

The idea of a truthmaker, an entity which by existing makes a proposition or other truth-bearer true, goes back to Aristotle, but it flourished in the logical atomism of Russell and Wittgenstein, where the truthmakers were termed *facts*. One lively strand of contemporary metaphysics is based on the idea that some or all true propositions stand in need of a truthmaker. The Truthmaker Principle, that every true proposition has a truthmaker, was probably first formulated by C. B. Martin in the 1960s but emerged in print much later in the 1980s. A restricted Truthmaker Principle deriving from phenomenology as much as from logical atomism was proposed independently in 1984 by Mulligan, Smith and Simons, who identified individual accidents or *tropes* as a primary source of truthmakers. Unlike the correspondence theory of truth, which requires a suspiciously cosy one-to-one correspondence between truths and what makes them true, truthmaker theories allow that truths may have more than one truthmaker, for example a disjunction, both of whose disjuncts are true, has as truthmakers those for both disjuncts, although either would have sufficed alone.

Later truthmaker theories diverge over how rich and numerous they take truthmakers to be. Truthmaker *maximalism* requires at least one truthmaker for every truth, whereas John Bigelow's principle that *truth supervenes on being* requires only that there be some sufficient reason, resting ultimately on what there is and is not, for why a given proposition is true rather than false, but this does not always amount to a true proposition's having a truthmaker.

The other respect in which truthmaker theories differ is in what entities they evoke as truthmakers. The most popular choice has been states of affairs, following the lead of Russell and Wittgenstein. Indeed David Armstrong, a leading truthmaker maximalist, contends that the world is composed ultimately of states of affairs, a view found also in situation semantics. Other truthmaker theories have looked for other entities to do the truthmaking role, either a single kind, such as tropes, or a mixture of kinds.

Trope Theory

The term 'trope' is due to Donald C. Williams, an American philosopher whose work achieved its deserved prominence long after it was written in the mid-century. Williams, influenced in some measure by Husserl, proposed a single-category ontology of tropes. Tropes are individual instances of properties, such as individual rednesses or roundnesses, located where their particulars are, and different in differing particulars. A similar conception of properties as 'thin' or 'abstract' particulars was proposed earlier in the century both by Husserl and by G. F. Stout, but the idea goes back to Aristotle's *Categories* and was standard in the Middle Ages and early modern period. Tropes offer nominalists a way to rebut many of the criticisms of realists about universals, avoiding some of the difficulties of earlier and more radical forms of nominalism such as that of Nelson Goodman in the USA or Tadeusz Kotarbinski in Poland. In the hands of Williams and followers such as Keith Campbell and John Bacon, tropes are proposed as the sole ontological basis category, universals being considered as classes of tropes grouped by resemblance while substances are classes of tropes grouped by a link such as compresence in space and time.

Persistence, Time and Events

An aside in David Lewis's *The Plurality of Worlds* claims that the only way to make sense of change is to suppose that objects that persist, or exist at different times, have temporal parts, as do events and processes. Such *occurrents* typically are extended in four dimensions: three of space and one of time. They thus *perdure*, that is, spread through time by having different phases, by contrast with substances such as Strawson's bodies and persons, which *endure*, that is, are present as a whole at each time at which they exist. Such *continuants* are typically three dimensional, being extended only in space. Lewis claimed that continuants cannot be said to have contrary properties at different times without either making properties relations to times, or indulging in *presentism*, the view that only the present is real, and past and future do not exist. He thus advocated a *four-dimensionalist* account of ordinary continuants, giving them distinct temporal parts to bear the contrary properties.

Lewis's argument and the responses to it brought it into connection with ongoing discussions of the dispensability or otherwise of the idea of real tense, that is, an ontological distinction between past, present and future, put forward under the name 'A-Series' as essential to time in John McTaggart's famous 1908 argument for the unreality of time. Tensers, or proponents of the A-Series, contrast with detensers, those who say time consists only of the B-Series, times connected by relations of earlier and later. Often tensers defend the three-dimensional account of change while detensers defend the four-dimensional account, though in fact the two oppositions are independent of one another. An unintended side effect of Lewis's criticism was to prompt several philosophers to defend the otherwise improbable doctrine of presentism.

Lewis's view of change echoes similar ones put forward by Russell, Whitehead, Carnap and others earlier in the 20th century, in response to the conceptual strains set up by Einstein's relativity theory. Whitehead's and Russell's view that the world is composed of events was a revisionary metaphysic which went into abeyance in the mid-century, to be revived by Lewis's arguments and somewhat earlier by semantic arguments from Donald Davidson. Davidson contended that the best way to account for the logic of statements about action, such as the inference from *Sam sliced the salami in the bathroom at midnight* to *Sam sliced the salami*, was to follow a suggestion of Frank Ramsey's that such statements contain a tacit quantification over events: roughly speaking, There was an event whose agent was Sam, whose object was the salami, whose instrument was a knife, whose temporal location was midnight and whose spatial location was the bathroom. Dropping one or more of the conjoined clauses and translating back into usual idiom would reveal the inferred sentence as an instance of conjunction elimination, not of some esoteric logic of adverbs. Davidson's analysis rapidly won converts to an ontology of events and has become the standard view among linguists. It coincided with a vigorous period of investigation into the ontology of action, and a range of divergent ontologies of events emerged, some like Davidson saying events were *sui generis* individuals, Quine holding with Russell that they were simply the contents of any portion of spacetime, others such as Jaegwon Kim saying they were property exemplifications, and yet others such as Roderick M. Chisholm saying they were states of affairs. This debate subsided without being resolved, the variety of theories being later enriched by the suggestion of Jonathan Bennett that events are tropes. Nevertheless, whereas in the mid-century events were much less respected than substances as primary entities in the metaphysical menagerie, now the boot is on the other foot: few metaphysicians question the existence of events, whereas the concept of substance is much less central than it was.

The Future of Metaphysics

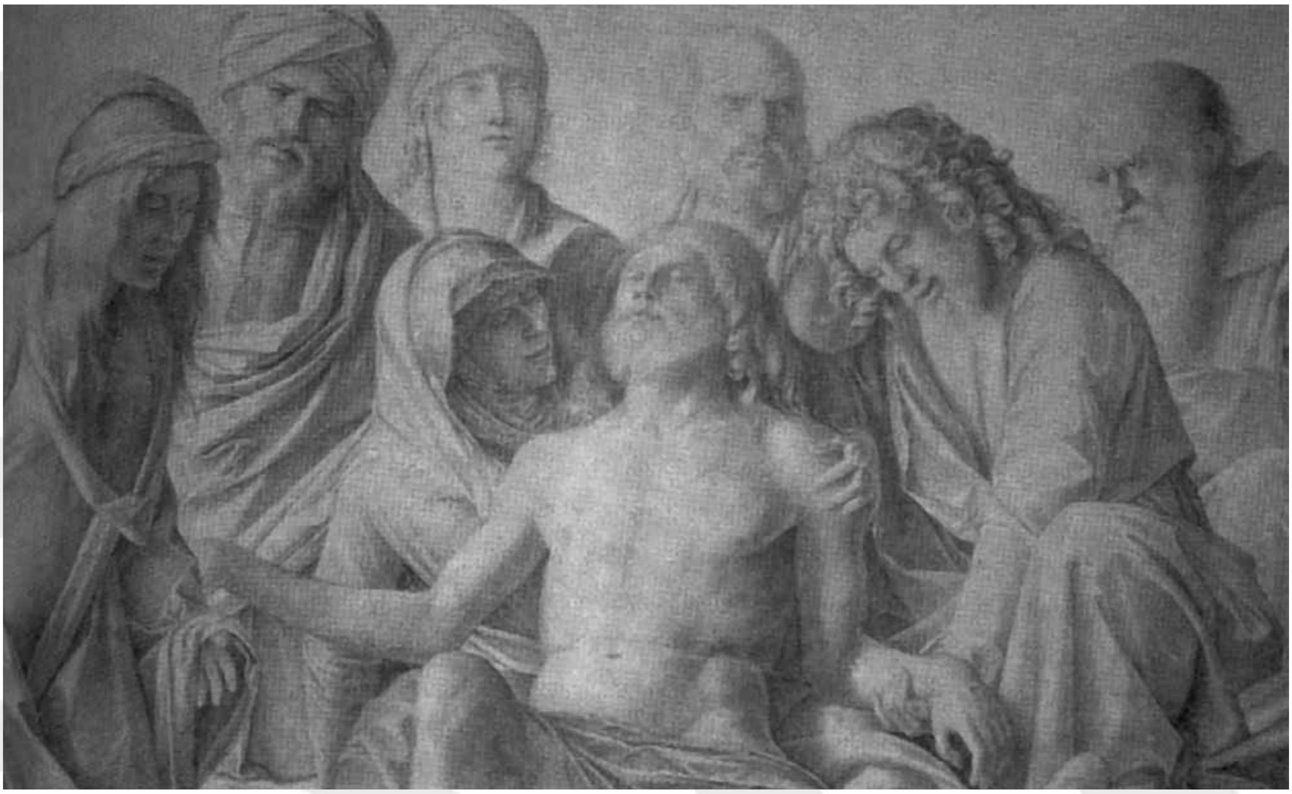
Metaphysics re-emerged reinvigorated in the twentieth century. Its debates lie at the centre of philosophy, as they did in ancient and medieval times. Metaphysical issues now permeate philosophical discussion of areas once considered to provide the scientific replacement for metaphysics, such as the philosophy of mathematics and the philosophy of physics. Computer scientists and intelligence artificers have borrowed the word 'ontology' to designate platform- and implementation-independent representations of objects in many domains of interest. Their use of the term touches the metaphysician's use only tangentially, but the task of providing computer-based representations for objects from all walks of life raises philosophical questions about the relative merits of alternative schemes for describing things, of the sort that linguists discussed in the 20th century. It is no accident that the large-scale CYC project of artificial intelligence led by Doug Lenat organises its data, intended to capture the commonsense knowledge that humans instinctively enjoy, around a systematic ontology.

Computer representation is bound to raise anew the kinds of question metaphysicians and philosophers of language have long pondered. In fact many of the more serious, rigorous and systematic attempts to investigate the natures of things from many domains now take place among computer scientists and cognitive scientists. Such attempts to represent the knowledge that human beings bring to bear on everyday situations and everyday language bid fair to reiterate positions and concerns of mid-century ordinary language philosophy, only this time with machines.

By contrast, the kinds of metaphysical issue raised by the application of metaphysics to the sciences is likely, because science does not confine itself to the explication of common sense, to result in revisions of our conceptual scheme, of the sort envisaged by the revisionary metaphysicians of the early 20th century, like Bradley, McTaggart, Alexander and Whitehead. Metaphysics in the 21st century could go in one of two directions. It could retreat again to the modest, descriptive and conservative variety, metaphysics within the bounds of epistemology, proposed by Kant and seconded by Strawson. Or the metaphysical enterprise could boldly go into new areas of application, such as medicine, biology, chemistry, engineering, economics and management, where computer modelling requires more than commonsense knowledge representation.

Because of the increasing specialisation of all sciences, including philosophy, there is the risk that metaphysics too could become compartmentalised into disjoint areas with barely any communication between them. Philosophy tends to resist compartmentalising more than the special sciences, but social and market pressures apply to philosophers and promote specialisation at the expense of synoptic visions. Certainly the trend of much late 20th century philosophy has been to narrower specialisation. The sheer volume of knowledge makes it ever more unlikely that a future Leibniz or Hegel could synthesise knowledge into a single system.

Nevertheless, though philosophers make poor prophets, I beg to state how I *would like* metaphysics to develop over the next hundred years, even if I am unsure whether it will do so. My vision for future metaphysics is that it should be untrammelled by the need to conform to ordinary language



or the restrictions of our everyday view of things. It should be bold and revisionary. It should abandon the reliance on mathematics and logic that has narrowed its vision in the twentieth century. It should treat the need to find place for a credible theory of linguistic meaning as a constraint rather than a method. It should draw on the wisdom of the great philosophers, and not pretend it is a kind of advanced science needing to refer to nothing more than five years old. It should reject *a priori* methods and certainty and be thoroughly fallibilistic, open to revision from within by argument and from without by scientific advance. It should encourage team effort and cooperation among metaphysicians and others, to counteract the pressures of specialism. It should be pluralistic, not because all the different views are somehow relatively right, but because, as Mao (all too briefly) recognised, the truth is more likely to emerge from the contention of competing theories than from the dictates of orthodoxy or fashion. It should be prepared to consider and engage in applications in areas hitherto considered remote from its concerns. It should aspire to be integrative and systematic, putting the various kinds of entity as putatively disclosed by all the special disciplines into a single overarching categorial scheme. If the late twentieth century was a heyday of analytic metaphysics, may that of the twenty-first be synthetic.

University of Leeds.

Suggested Reading

To reference all the philosophers mentioned in passing in this article would double its length. If anyone wishes chapter and verse for any allusion, they are welcome to e-mail me at p.m.simons@leeds.ac.uk. Instead I will list some of the best books available for getting to grips with the lively modern metaphysical literature.

Hales, Steven D., ed. *Metaphysics: Contemporary Readings.* Belmont, CA: Wadsworth, 1999.

Kim, Jaegwon and Sosa, Ernest, eds. *Metaphysics. An Anthology.* Oxford: Blackwell, 1999.

Laurence, Stephen and Macdonald, Cynthia, eds. *Contemporary Readings in the Foundations of Metaphysics.* Oxford: Blackwell, 1998.

Loux, Michael J. *Metaphysics. A Contemporary Introduction.* London: Routledge, 2001.

Loux, Michael J. *Metaphysics. Contemporary Readings.* London: Routledge, 2001.

Loux, Michael J. and Zimmermann, Dean W., eds. *The Oxford Handbook of Metaphysics.* Oxford: Oxford University Press, 2003.

Lowe, E. Jonathan. *A Survey of Metaphysics.* Oxford: Oxford University Press, 2002.

Van Inwagen, Peter. *Metaphysics.* Boulder, CO: Westview, 2002.

Van Inwagen, Peter and Zimmermann, Dean W., eds. *Metaphysics: the Big Questions.* Oxford: Blackwell, 1998.

Garrath Williams

Two approaches to [Moral] responsibility

Part 1

Introduction

The American legal philosopher, Joel Feinberg, once observed that 'moral responsibility... is a subject about which we are all confused.' (1970: 37) Here I want to contrast two influential philosophical accounts of why we make responsibility attributions – for instance, by praising and blaming people, in saying that someone deserves to be punished, and so on. In these practices, we respond to other people – and ourselves – as the authors of their actions. If they act well, we feel they deserve our admiration and sometimes gratitude or loyalty. If they act badly, we will tend to resent them, and feel they ought to make up for their actions, or perhaps even be punished. And we often feel guilty, remorseful and sometimes proud of how we have acted ourselves. In short, we think of people as morally responsible.

One very influential approach to this subject is broadly *Kantian*. This view sees responsibility for actions as stemming from our ability to exercise *self-contro*l. On this account moral responsibility exists because a person freely chooses her actions, and tends to lead us toward the idea of free will. In addition, because praise and blame respond to the person as the chooser of her deed, they *recognise* her dignity as a rational agent, as modern followers of Kant tend to put it. A much older approach goes back to *Aristotle*. This view situates attributions of responsibility in terms of our on-going relationships with one another. This more nuanced approach stresses the importance of *mutual accountability, moral education*, and *assessments of character* in terms of the many vices and virtues.

I will not try to convey the exact details of these philosophers' accounts. What I do want to show are two things. First, how their ways of looking at mutual accountability capture important parts of our everyday commonsense. One modern commentator claimed that, in our attitudes to moral responsibility, 'we are all Kantians now' – by 'we' meaning not just philosophers but all Western persons (Adkins, 1960: 2). Another central figure in this debate, Bernard Williams, agrees that Kant captured a widespread tendency of modern moral thinking, but also claims that there exist important counter-tendencies in our practices of responsibility. For Williams, ancient Greek understandings are actually more realistic and helpful than the Kantian one. I think Williams is quite correct in this, and the second point of these articles will be to suggest that so far as our ideas of moral responsibility actually make sense, they are best captured by a (roughly) Aristotelian account.

In this first part of the article, I want to sketch two things. First, I will say something about the idea of free will. The paradoxes involved in this idea often occur to people even before they come to philosophy, and these difficulties will be central to Kant's account. But second, before turning to Kant, I would like to tackle Aristotle's broad approach, and show that, before free will was invented by Christian philosophers, there was a quite different way of thinking about moral responsibility – one that has much to teach us.

In the second part of the article, to appear in the next issue, I ask why Kant's account continues to attract many people who would not dream of calling themselves Kantians – indeed, many who have never even heard Kant's name. Kant's theory involves a powerful idea of moral worth based on choice. This idea, though problematic because of the idea of freedom it seems to depend on, does account for many of our intuitions about moral responsibility. But it is not the only explanation of these intuitions, nor – I will argue – is it the most plausible.

The problem of the will

The free will debate has become an old chestnut of modern philosophy. It is an intuitively plausible way of approaching the issues – familiar to many even before they encounter philosophical texts. It is perhaps surprising, then, that this debate is actually a rather modern one.

The basic gist is this: if I am to be responsible (*really* responsible) for my conduct, then it must be within my control. However, if it is true that every event in the universe is determined by causal laws, then this must be true of the events that constitute my actions. Therefore, my conduct cannot really be within *my* control; therefore, I am not *really* responsible for my conduct. Two conclusions immediately suggest themselves. One is that it is incoherent to praise or blame me – and everyone else – for our actions, because it is so difficult to doubt the causal well-orderedness of the universe. The alternative conclusion, scarcely more appealing, is that the human will somehow sits outside this causal framework – ie, we have free will – because it is unthinkable that our moral ideas be so desperately incoherent.

Both lines of thought are *incompatibilist*; that is, they see the ideas of responsibility involved in praise and blame as incompatible with the causal well-orderedness of the universe. But while both attract some limited support among philosophers, the overwhelming consensus now lies with *compatibilism*. This is simply the thesis that responsibility and causal order are compatible. Most philosophers agree that the alleged incompatibility results from some important confusions, although there is much less consensus about what these may be. At least one area of confusion is clear, however, and forms the central issue of this article: what sort of responsibility for conduct is involved in praise and blame? Several familiar points in the free will debate are helpful for approaching this.

In the first place, it is well-known that this debate does not turn on the truth of determinism as such. Determinism is the idea that every event is determined by fixed causal laws. Yet it may well be that every event is somehow random in origin. One interpretation of quantum physics claims that causal laws are the product of statistical regularities, while these regularities stem from a near infinite number of random events. So far as the human will is concerned, this makes no difference. If my conduct is the product of chance, this makes *me* no more responsible for it than does its being generated by causal laws. The point is that if *I* am to be held responsible, then *I* must control my conduct – not causal laws, nor mere chance, nor some particular combination of the two.

Second, the free will debate bears a disquieting similarity to an older controversy. In medieval philosophy it used to be asked how God's omniscience – his knowledge of everything that has happened and will happen – could be reconciled with our being subject to his moral judgment (that is, being sent to heaven or to hell). If God knows what we will do then this seems to imply that it is already decided whether we will act well or badly. And this, in turn, suggests that it makes no sense to punish or reward us. Theologians developed various doctrines to overcome this difficulty, but few sound convincing to modern ears – perhaps because the problem itself is no longer a live one, even for most believers. However that may be, it is interesting that many modern versions of the debate seem to take at least one of the planks of Christian theology for granted: that individuals have wills that can be bad or good, usually now expressed by philosophers in terms of people's 'blameworthiness' or (less often) 'praiseworthiness.'

In this way, the modern American philosopher Joel Feinberg ironically referred to 'a moral bank account' that we carry through life, which sums up our moral credits and debits in a single figure (1970: 20). Whether or not such an 'account' makes sense, it is at least clear that the idea of 'the will' is by no means self-explanatory.

For Kant, as we shall see, it was obvious that all my choices can be summed up in a single moral evaluation, whether I have a 'good' or 'bad' will. Kant is equivocal, however, as to whether only God might make this evaluation, or whether human beings might also form reasonable opinions on the matter. But especially if we take the point of view of mutual, *human* accountability, it is far from obvious why we should believe any such single evaluation to be possible, or what role this evaluation might play in our individual or collective lives. Certainly, we usually praise and blame in terms of particular actions and particular vices and virtues – *not* a good or bad will.

Third, this way of framing the issues creates a gulf between the conduct of normal moral agents (adult human beings of sound mind) and the conduct of other creatures – animals and children. At some stage of evolution, and at some stage toward maturity, certain animals become 'free,' whereas before they had all been 'determined' in their conduct. Although it is grossly implausible that there are no relevant moral differences between the other animals, children, and human adults, it is no more plausible that the free will simply pops into existence at a certain stage of human development. Within a Christian framework this issue was less problematic: human beings, and only human beings, have souls. Thinkers have always been aware that animals can show profound care and concern for one another, but this poses philosophical difficulties given our awareness of evolutionary continuities and the fact that the Christian idea of the soul is no longer something we can take for granted. More than this, within Christianity the moral demands upon adults could be interpreted in terms of obedience to God-given laws. Yet in the modern world demands for obedience are those we make of children. Adults are expected to obey certain basic rules, but we also expect something more – a sense of responsibility that involves judgment and initiative. Despite all this, however, we tend to think there is something sufficiently distinctive about human action, so that many non-religious people find the idea of free will plausible, and almost everyone assumes that only (mature?) human beings can be responsible for their actions.

Taking the last three points together generates a further point. If the idea of the will is complex, and there is no straightforward moral dividing line between children and adults, between humans and other animals – together, these ideas suggest that a 'will' is not something we all straightforwardly 'have.' In other words: it is implausible that all adult humans have the *same capacities*, all to the *same extent*, that are involved in controlling action. One way of retaining the idea of the will might be to think of it as the *bundle of capacities* that are needed to control action in the light of moral concerns, these capacities being set only at such a level that all adult human beings of sound mind really seem to possess them. But two points need to be kept in mind about such a strategy. First, it remains the case that people will vary in how far they possess their capacities, and this variation will largely be a product of upbringing and natural qualities – that is, *not* something within an individual's own control. Second, the sort of ultimate control over one's moral character supposed in Kant's or any other 'free will' account is unlikely to be vindicated in this way.

The Aristotelian approach

For an analysis of the basic set of capacities needed for moral action, philosophers continue to go back to an ancient account of moral responsibility. The terms of the free will debate are new, arising with the birth of modern science, and the theological debate about free will arises only with Christian thought. But the question of responsibility for action has always been known to philosophers, and the most famous discussion of when people can be praised and blamed for their actions remains Aristotle's. Many have noticed that Aristotle and his contemporaries saw no need to talk about responsibility in terms of free will. Aristotle asks whether acts are *voluntary*, and whether we *attribute* them to a person or to other factors. Some have ascribed this way of framing the issues to a lack of moral or scientific sophistication on the part of the ancient Greeks. However, a number of modern philosophers, most prominently Bernard Williams and Martha Nussbaum, have suggested that an Aristotelian account is actually more coherent and sophisticated than those typical of modern philosophy – and, indeed, more coherent than our modern, 'common sense' intuitions about moral responsibility.

Aristotle *assumes* that we are responsible for our actions (so that others can reasonably praise or blame or punish us), and proceeds by pointing to various conditions that lessen or cancel this responsibility. He discusses force of events, threats and coercion, ignorance, intoxication and bad character. (He also remarks on the continuities and differences between children's agency and that of normal adults in ways that illuminate our practices of responsibility, a point I do not consider here.) Taken together, his account shows the basic elements involved in being a person who can reasonably be praised or blamed.

The first limitation upon voluntary action that Aristotle discusses is force of circumstances. His well-known example concerns a ship caught in a storm; the sailors must throw goods overboard if the ship is not to sink (NE 1110a). In this case the action is not fully voluntary, and we would not blame the sailors for their actions. (Nor, of course, would we blame the storm: the undesirable consequence, the loss of the goods, must be chalked off as the result of natural events, for which no one is responsible.) Note that such cases are extreme examples of the force of necessity under which we always live – we are always constrained in our actions by natural facts, although we only tend to notice this when the constraint is sudden or unexpected.

In fact, it tends to be the interference of other people which causes us the most grief – and which really causes problems for responsibility attributions. Such interference can take many forms, but its paradigmatic forms are coercion and manipulation. Regarding coercion, Aristotle's judgment is balanced. It depends on what action my coercer is demanding of me, and what threats he makes. Some actions are so heinous that we should be blamed for doing them, whatever we are threatened with (and whatever blame also attaches to our coercer) – thus Aristotle dismisses the idea that a man might be 'compelled' to kill his mother (NE 1110a). This makes it clear that a central issue at stake in attributions of responsibility is the *expectations* that people have of one another. There are some forms of coercion we do not usually expect people to resist, but there are also some sorts of action that we think people should never undertake, regardless of such factors. In such cases praise and blame are clearly working *to clarify and reinforce these expectations* – in other words, they provide for a form of *moral education*.

Aristotle does not comment on manipulation, where other people support us in a false view of our circumstances. But he does discuss ignorance of these circumstances, and how it undermines our responsibility. If we are ignorant of who someone is, for example – as was Oedipus, who did not know that the old man obstructing him was actually his father – we may commit heinous acts we would otherwise abhor – thus Oedipus committed patricide, killing his own father. For Aristotle, such actions are not to be blamed (at least, when the ignorance is not itself culpable and the killing was otherwise justified).

What decides good or bad character is how a person reacts when he finds out the truth – if we fail to regret our deeds, then we can certainly be blamed, even if the original choice was justifiable. Among other things, Aristotle makes it clear that our praise and blame is often not about an individual act, but about the *character* of the person who acted.

Importantly, it is not every form of ignorance that excuses. Moral knowledge is very different from factual knowledge. What if a man did not know murder was wrong? Would this make his murders morally innocent? Aristotle says not: there are certain things we can and do expect people to know – above all, basic moral truths such as the wrongness of murder. But this knowledge is not as straightforward as it might appear: it must include a fairly good capacity to judge which sorts of killing count as murder. Nazi bureaucrat Adolf Eichmann organised the killing of thousands, without a sense of its wrongness. Aristotle is clear: such moral ignorance, an inability or failure to judge, excuses no adult. Eichmann should be held responsible for murder. But why should moral ignorance not excuse, when factual ignorance does? We must recognise that moral knowledge is actually rather different from factual knowledge. *If a person is morally ignorant it is his whole character, his lasting ability to judge and act well that is impaired* – and presumably very difficult to set right. Isolated errors in factual knowledge, on the other hand, can be easily corrected. So long as we are subsequently able to recognise and regret what we have done, *factual mistakes involve no lasting corruption of character.*

Still, if a person *is* morally ignorant it follows that they are unable to choose well; and Aristotle concedes that many people of settled bad character – be they morally ignorant or otherwise – can no longer choose to act well. Does this mean that blame is incoherent or misplaced? He claims not. Even if the vicious person cannot now choose to act otherwise, there was a time when her vices were not fixed, when she *could have chosen* not to be vicious. Therefore, Aristotle says, she can be blamed. This is neat but rather unconvincing. Aristotle is famous for emphasising the importance of good upbringing and habituation, and presumably many vices are formed in childhood, before people have formed capacities for deliberating reasonably. Indeed, many vices undercut the capacity for rational deliberation. So it is a clear implication of Aristotle's own account that the badly brought up person may never be in a position to choose *not* to be vicious. Note, further, that this unconvincing move represents Aristotle at his most *Kantian*: blame is justified by reference to control, to a 'could have done otherwise' – even when his own account of character formation suggests that such control may well never have existed.

It is also interesting that many vices take the form of moral ignorance – of not knowing that certain things are wrong or failing to recognise that certain actions represent some sort of wrong-doing. (This is often a failure only with regard to one's own actions: Bishop Butler once observed how common it is for people to condemn others for the vices they themselves are most notorious for!) The difficulty is that the vicious person cannot, or will not, see her own vices as such – in which case she is in no position to 'take control' and will see no reason to act differently in the future. But this does not mean that we have no reason to blame her, most obviously because we might hope that blame will help educate her, morally speaking.

What are we to say, though, where a person seems incorrigible – quite settled in some particular vice, either because she cannot understand the criticism or because she is unable to alter her character or habits? (In real life it's often somewhere in between: 'Yes, I know I shouldn't behave like that, but I can't help it, and it's really not as bad as all that.') Such cases are very common, and – unless we suppose that they are not morally deplorable – seem to undermine the modern, Kantian assumption that blame must relate only to conduct under our control. Clearly, if we think a character trait is really beyond alteration, by us or by the person concerned, our blaming won't involve an attempt to reason with the person we condemn. But our condemnation might have another rationale, for example, to clarify what sort of standards we expect of *others*. And it is clear that praise often has this rationale, too: a virtuous person might be quite *unable* to do certain things – commit cruelty, for example.

In sum, Aristotle's account is not entirely self-consistent. Generally his focus is upon the qualities of character revealed by acts, in terms of our overall moral expectations, and it is these that responsibility attributions attend to. However, he sometimes suggests that bad qualities are to be blamed because they are, or were, subject to choice, even though this quasi-Kantian claim is not really supportable. Despite this, philosophers have returned to his account again and again to illuminate the main ingredients of responsible agency:

- The capacity to respond to others' censure and encouragement, whether expressed emotionally (eg, as resentment), institutionally (eg, as punishment), or in the various forms of praise and blame.
- The capacity to exercise deliberate, sustained control of one's conduct. One reason why young children are not responsible agents is their inability to sustain control over time, partly owing to a lack of

emotional self-understanding. (That we do praise and blame children, however, emphasises the *educative* and *encouraging* role that praise and blame can play, both in developing such control and in inculcating shared moral standards.)

- A reasonable grasp of how actions impinge on others and how they are socially understood – that is, of our mutual moral expectations.

Taken together, these capacities allow us to participate in forms of mutual accountability, whereby we inculcate and to some extent enforce shared standards of action.

This list may not be comprehensive, but it serves to illustrate the underlying point of an Aristotelian account: moral responsibility seems to rest on these sort of fairly basic capacities, which do not seem to demand any strong metaphysical elaboration. Indeed, if we approach the matter this way, the puzzle seems to be inverted. Not, 'how might free will and determinism be reconciled?'; rather, 'why should we feel there is a metaphysical issue at all?' It is to this question, why so many people feel that metaphysical issues are involved here, that I will turn in the second part of this article.

University of Lancaster.

Acknowledgements

I would like to thank David Archard for comments on a version of these articles. Much of this material also appears, in rather different form, in an entry on 'Praise and blame' for the Internet Encyclopedia of Philosophy: www.utm.edu/research/iep.

References and further reading

Adkins, AWH (1960) *Merit and responsibility*, Clarendon Press, Oxford

Aristotle *Nicomachean ethics* (the most readable translation is Roger Crisp's, Cambridge University Press, Cambridge, 2000)

Feinberg, Joel (1970) *Doing and deserving: essays in the theory of responsibility* (Princeton University Press, Princeton NJ) – a set of classic essays on responsibility for action

Fingarette, Herbert (1967) *On responsibility* (Basic Books, New York) – another set of classic essays, including the argument that blame is intelligible only insofar as it addresses a person's pre-existing concern for others

Kant, Immanuel (1784) *Groundwork to the metaphysics of morals* (the best translation is Mary Gregor's, Cambridge University Press, Cambridge, 1998)

Korsgaard, Christine (1996) 'Creating the Kingdom of Ends: reciprocity and responsibility in personal relations' in her *Creating the Kingdom of Ends* (Cambridge University Press, Cambridge) – a sophisticated Kantian account of holding people responsible

Midgley, Mary *Can't we make moral judgments?*

Skorupski, John (1999) 'The definition of morality' in his *Ethical explorations* (Oxford University Press, Oxford)

Smiley, Marion (1992) *Moral responsibility and the boundaries of community: power and accountability from a pragmatic point of view* (University of Chicago Press, Chicago) – criticises conventional discussions of freedom and determinism, claiming that they fail to investigate the idea of responsibility

Strawson, Galen (1991) *Freedom and belief* (Clarendon, Oxford)

Strawson, Peter (1974) 'Freedom and resentment' in his *Freedom and resentment and other essays* (Methuen, London) – this famous essay resituates the free will debate by highlighting the importance of 'reactive attitudes' such as resentment to interpersonal relations

Williams, Bernard (1993) *Shame and necessity* (University of California Press, Berkeley CA) – a sustained argument that the ancient Greeks had a nuanced and sophisticated account of responsibility attributions

– (1995a) 'How free does the will need to be?' in his *Making sense of humanity and other philosophical papers, 1982-1993* (Cambridge University Press, Cambridge)

– (1995b) 'Voluntary acts and responsible agents,' in his *Making sense of humanity*

James Hill

The Philosophy of [Sleep]

The Views of Descartes, Locke and Leibniz

In the last decade or so consciousness has once again become a focus of interest in philosophy of mind, but so far sleep has barely been mentioned. Sleep raises special issues for any theory of consciousness. By this I am not referring to dreams and the sceptical difficulties that surround them – those difficulties always attract at least a moderate amount of attention. I am referring to *dreamless* sleep, the dark episodes of the mind that seem to leave no trace in us. In the seventeenth century there was lively controversy over the nature of dreamless sleep and philosophers attempted to incorporate their understanding of sleep into a more general view of the mind and consciousness. Here we will explore and contrast three philosophical accounts of sleep – those of Descartes, Locke and Leibniz – before assessing some of the problems and insights the debate about sleep provides for an understanding of consciousness.

I

It was for René Descartes and his followers that sleep first reared its head as a philosophical problem in the modern period. In the *Meditations* Descartes found his mind to be essentially a 'thinking thing', *res cogitans*. To say that the essence – or principal attribute – of the mind was thinking, meant also to say that the mind could not lose this attribute and still continue to exist. The existence of the mind without a thought was no more conceivable than of a piece of matter without extension. Since Descartes used the term 'thinking' to refer to all conscious states, this meant that so long as my mind exists I must always be conscious, even during a fainting fit, or in the deepest sleep. In the Second Meditation Descartes memorably asserts,

> I am, I exist – that is certain. But for how long? For as long as I am thinking. For it could be that were I totally to cease from thinking, I should totally cease to exist.[1]

Now, it is true, one option still seems to remain for Descartes if he wanted to deny the conclusion that we are conscious the whole time we are sleeping. He could take the view that in sleep the mind stops thinking *and* existing and on waking returns to existence and thought: he might have opted, that is, for a pause in the existence of the mental substance. After all, it is already a feature of his system that the human mind is, in common with all finite substances, continually conserved in existence by the action of God, and he asks us to look upon this conservation as a kind of continuous recreation. So what stopped Descartes from saying that when a mind goes to sleep God takes a pause before recreating that same mind on waking?

It was his doctrine of substance that closed off the option of the existential pause. Descartes held the mind to be a substance, and a substance is a thing that is able to exist independently of the activities of all things other than God. An existential pause during sleep would mean that (with God's connivance) the mind could be temporarily destroyed by, say, the action of a sleeping-pill, or the voice of a certain lecturer, and that it could be brought back into existence by a loud noise or a wet flannel. Its existence would be contingent on the activities of other finite existences and it would therefore be quite unfit to qualify as a substance.

So, taken together, Descartes' metaphysical doctrines of substance and essence meant that he had to commit himself to the controversial view that even in the deepest sleep we are really conscious. An obvious objection immediately suggests itself: why, if we are always conscious, do most of us think that we are not so in dreamless sleep? Why is deep sleep looked upon by practically all people, outside Cartesian circles, as a gap in mental activity? Descartes tried to meet this objection, when it was put forward by his critic and adversary Pierre Gassendi, in the following way:

> So long as the mind is joined to the body, then in order for it to remember thoughts which it had in the past, it is necessary for some

> traces of them to be imprinted on the brain; it is by turning to these [...] that the mind remembers. So is it really surprising if the brain of [...] a man in a deep sleep, is unsuited to receive these traces? [2]

During dreamless sleep, his argument runs, the mind can lay down no new memories. Thus, on waking up we are unable to recall any of the thinking that was in fact going on while we were sleeping. In fact, even if we are woken in the midst of a dreamless sleep, we will still be convinced that we were conscious of nothing: our brains, the physical organs in which memories are stored, do not retain, even for a split-second, the thoughts in question.

But why this memory-failure? Descartes seems to hold that it is the result of the soul withdrawing – so to speak – from the body (and in particular from the brain).[3] The consciousness that goes on in this state of withdrawal, or retirement, does not engage the physical mechanisms of memory in the brain – it wafts by without being recorded. In fact, 'memory-failure' may be a misleading phrase for what happens here: the thoughts of the sleeper never even enter the memory and therefore there is really no possibility of recall succeeding. The experiences are just not available for recollection.

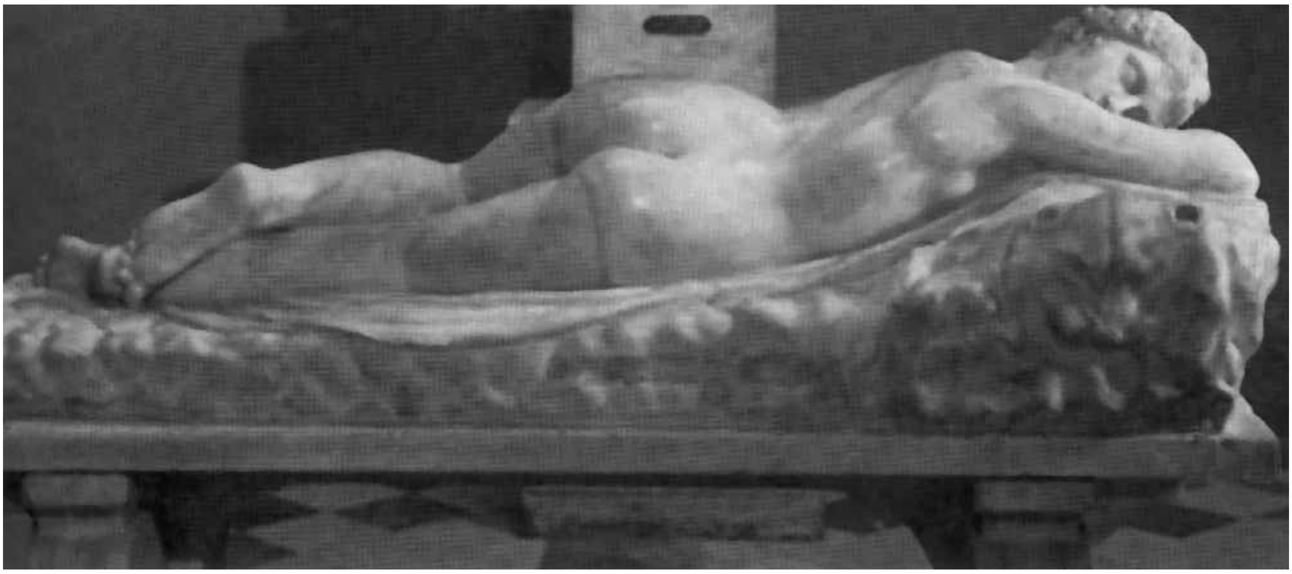
Descartes' understanding of sleep as the conscious mind retiring, or withdrawing, from the brain finds explicit expression only in the reply to Gassendi that we have just mentioned. And here there is just one paragraph in which the tone is somewhat speculative. Other Cartesians, however, developed the position found in embryo in their master. Nicolas Malebranche, for example, offered two different accounts of why there was non-recollection. The first adds to Descartes' own view. Malebranche explains the suspension of memory by the fact that only 'pure intellection' takes place, and that such thought – which deals with the abstract concepts of maths, logic and metaphysics – has no imagery associated with it and therefore, unlike sense and imagination, does not involve the animal spirits, and thus leaves no traces in the brain.[4] This shows skilful employment of the Cartesian doctrine to bring a more precise understanding of the soul's thinking in retirement from the body, though of course it would be unappealing to those who were downright sceptical about the faculty of pure intellection in the first place.[5]

Malebranche's second explanation departs somewhat from the original suggestion made by Descartes:

> It sometimes happens that we have so many different thoughts that we believe we are thinking about nothing at all. This is seen in the case of people who fall into a swoon. The animal spirits, swirling irregularly in their brain, stir up so many traces that no one of them is opened sufficiently to excite a particular sensation or distinct idea in the mind. As a result of this, these people perceive so many things simultaneously that they perceive nothing distinct – which leads them to think they have perceived nothing.[6]

Here we can more legitimately talk about 'memory-failure'. What Malebranche seems to be describing is thinking that is so fragmented and confused that it doesn't stick in the mind. Whatever traces are laid down they are too indistinct to be the subject of recall. As a result the subject concludes that he was not thinking at all in the episodes in question. In this second explanation Malebranche comes close to the view of Leibniz which we shall examine in a moment. But it is Malebranche's first explanation, in which the soul contemplates ideas of pure intellection in retirement from the body, that is most quintessentially Cartesian and which was generally recognised as the orthodox Cartesian view of sleep; so I shall refer to the Cartesian view, put forward in Descartes' replies and in Malebranche's later *Recherche*, as the *retirement* view.

[Sleep]



II

Now, it is hardly surprising that the Cartesian view should be subject to the sharpest criticism by an empiricist such as John Locke. In the first section of the second book of his *Essay Concerning Human Understanding* Locke launches a swingeing attack on the Cartesian position. He makes his first and most important move against the Cartesians by treating their thesis (that the mind always thinks) quite independently of its metaphysical background – the definition of mind as *res cogitans*, the doctrines of essence and substance. For Locke such a thesis was, like any other sweeping statement about the actual content of our minds, an empirical hypothesis. And, as such, it went against all the evidence and was grossly improbable. Anyone will tell you that they spent much of last night without thinking at all. If metaphysics leads us to deny this commonplace, Locke implied, then so much the worse for metaphysics.

Locke held, broadly speaking, that we should accept the verdict of sleepers themselves that dreamless sleep constitutes a gap in thinking – the mind does not retire to contemplate ideas of pure intellection in some phantom realm, it simply blacks-out. Let us call this the *black-out view*. Now Locke thinks that he is on particularly strong ground when it comes to hypotheses about sleep – he takes it that there is no higher authority for whether a person is thinking or not than the consciousness of that person themselves. The common-man is in a much better position to know what he did or did not think about last night than an armchair hypothesis-monger.

Some of what Locke says against the Cartesians is in a satirical vein. At one point he declares that 'every drowsy Nod shakes their Doctrine'.[7] But Locke makes more subtle moves too. His dissatisfaction with the view that the mind is thinking throughout dreamless sleep but remembers nothing, leads him to the question of personal identity. Who is this person that is thinking in me while I sleep? It is not I, myself, he argues, because there is no continuity with my present thoughts. Memory, as Locke makes more explicit in the chapter on personal identity,[8] is constitutive of the self as a continuing, reflecting person. If there are periods of thinking in me that I can have no conceivable access to when awake, then the thinking in question is really that of another person. If, say, Socrates asleep is busy thinking, but remembers nothing on waking, then this night-time thinking no more concerns Socrates than does the 'Happiness, or Misery of a Man in the *Indies*, whom he knows not'.[9]

Generally, Locke holds that it is a much more probable hypothesis that we are without thoughts in dreamless sleep. But realising that he cannot definitively refute the Cartesian hypothesis, he is content to at least point out that thoughts deemed to take place in sleep do not belong to the waking self.

III

In Descartes and Locke we have seen a fairly straightforward contradiction of opinions – thesis, antithesis. It is Gottfried Leibniz who, in commenting on Locke's *Essay*, comes up with a synthesis of the two conflicting positions. Leibniz's view of sleep is, in my opinion, the most promising and fertile of the three views we are considering.[10]

Leibniz begins by agreeing with the Cartesians that the mind is always thinking, even in dreamless sleep. Just as there is always motion, however imperceptible, in bodies, so there are confused and indistinct thoughts continually passing through the mind of the sleeper. However, Locke is also right to say that there no *conscious* thinking goes on in sleep. The thinking in question is unconscious. It is unconscious by virtue of being unfocused, fragmented and unattended to. Leibniz therefore drops the major assumption of both the Cartesians and Lockeans that thinking is by its very nature conscious. He saw that if he allowed unconscious thoughts, or perceptions, then an acceptable third way between the opposing views of sleep opened up. In Leibniz's terminology 'perceptions' occur in sleep, but not 'apperception' – his term for self-conscious thought. These unconscious perceptions he called *petites perceptions*, or 'minute perceptions' and he held that they were too faint and indistinct to be the subject of awareness. Let us call Leibniz's view of sleep the *confusion view*.

Leibniz's reasons for arriving at this synthesis were manifold and he drew on metaphysical principles, just as Descartes did, as well as on empirical observation, like Locke. Among the empirical reasons, two stand out, appealing to the phenomena of waking and falling asleep respectively. In the first of these Leibniz notes that it is sometimes easier to wake up a sleeper than at other times. It is natural, he argues, to treat this as being because someone sleeping lightly has more sense of what is going on around him. His minute perceptions grow into larger, conscious ones more readily. If this is so, the implication is that there are degrees of being asleep – a continuum from waking to the deepest slumber. And this continuum is to be understood in terms of the relative distinctness of the minute perceptions in the different stages of sleep. Secondly, Leibniz notes that a good way of getting oneself to sleep is to allow one's thoughts to wander. We all know how thinking about a problem with too much singlemindedness stops us dropping off. By letting one's attention be divided between many perceptions, one may induce sleep. It then seems natural to treat the induced state as unattentive, unfocused thought. Thus the confusion view has the virtue of suggesting that there is a continuity between extreme tiredness and the process of falling asleep on the one hand and sleep itself on the other.

Leibniz goes on to draw a comparison between sleep and the periphery of consciousness. In any waking experience there are many perceptions that are too unfocused to be consciously registered. To use an example not in Leibniz, but which is in harmony with his thought: if the clock in a room stops ticking, we 'hear' the silence it leaves. This implies that although we were not conscious of the ticking of the clock, we must have been perceiving it in some unconscious way all the time, otherwise we could not notice its absence. In sleep such unregistered perceptions become the *sole* contents of the mind. In other words the periphery of consciousness is a kind of partial sleep that becomes more general when we are drowsy and finally takes over completely when we drop off. As Leibniz puts it, it is as though we had been 'selectively asleep' with regard to objects at the periphery of consciousness, 'and when we withdraw our attention from everything all together, the sleep becomes general'.[11]

IV

Three well-defined and contending views of the mind's activity during sleep emerge from Descartes' original discussion:

(i) The *retirement* view of Descartes and Malebranche. The soul is consciously thinking throughout dreamless sleep, but no new memories are laid down because the thinking is disembodied 'pure intellection'.

(ii) The *black-out* view. Locke's more common-sense contention that the mind simply does not think in dreamless sleep and is therefore quite unconscious.

(iii) The *confusion* view. Leibniz's view that the mind continuously thinks in sleep but, because the perceptions involved are too confused and fragmented, it does not do so consciously.

I will now say something about how these views, and the reasons used to support them, relate to the problem of consciousness.

Firstly the retirement view raises a problem for the most obvious and popular definition of consciousness which is favoured by, for example, John Searle:

> 'consciousness' refers to those states of sentience and awareness that typically begin when we awake from a dreamless sleep and continue until we go to sleep again, or fall into a coma or die or otherwise become 'unconscious'.[12]

This is a definition by contrast: consciousness is what goes on when we are not sleeping, comatose etc. One problem with such a definition is that it relies on the reader *not* being a Cartesian. For, on Descartes' conception, the mind is permanently conscious (even after death), and so there is no contrast to be had with states of unconsciousness. Searle would surely reply that even for the Cartesian, it *seems* as if we are unconscious during sleep because of the gap in our memory and that that is enough to get the contrastive definition off the ground. But this would be to imply that any memory gap will do the job of making a contrast with consciousness just as well – I cannot remember what I was doing on the afternoon of June 29th, 1987 for example, so that would be an example of my being unconscious. This brings us to the crux of the problem. There seems to be no way of imagining unconsciousness that distinguishes it from a blank in the memory, a point we have seen exploited by the Cartesians. This means that a contrastive definition of consciousness is quite different from, say, a definition of light by contrast with darkness. It makes sense to define light as the negation of darkness and vice versa, because we have experience of both (for darkness just turn out the lights). With consciousness and unconsciousness the experience is inevitably one-sided.

A second important question is raised by Locke's critique of the Cartesian position that attributes consciousness to us without the faculty of retaining our thoughts, even in the shortest term. Locke describes this as 'a very useless sort of thinking'. He continues,

> the Soul in such a state of thinking, does very little, if at all, to excel that of a Looking-glass, which constantly receives variety of Images, or *Ideas*, but retains none; they disappear and vanish, and there remain no footsteps of them; the Looking-glass is never the better for such *Ideas*, nor the Soul for such Thoughts.

Locke comes close here to saying that thinking without memory is not really thinking, any more than the passing images on a mirror are perceptions. Just as someone could not talk meaningfully if they were quite lacking in short-term memory – they would forget what they had said from one word to the next – so one could not think successfully if what one was thinking about dropped out of one's mind the very instant it was thought. I suspect one should go further and say that consciousness itself is not possible without memory. In order to be conscious, I must be conscious of something. But what could be the object of my consciousness if none of my thoughts could be retained for any duration. Could I be conscious of a triangle, for example, if when I thought about one angle I forgot about the other two and, indeed, about the sides and everything else? Would I even be able to think of the angle under these circumstances? Even if I concentrated on just one thing continuously, I would not be aware of doing so without memory, since each instant I would have forgotten what I was thinking of the instant before. Mentation would be a succession of vanishing points: it would be a blind-play of imagery, less even than a dream, one would like to say, paraphrasing Kant. The faculty of memory is internal to consciousness: it is not an optional extra.

A third point at which this debate about sleep touches on the question of consciousness is seen in Leibniz's comparison between sleep and the periphery of consciousness. As we have seen, for Leibniz the unregistered peripheral perceptions, which always attend our consciousness in waking life, become the *sole* contents of the mind in sleep. In other words the periphery of consciousness is a kind of

partial sleep that becomes general when we drop off. Leibniz's view here is important because it has the advantage of understanding sleep not as a special phenomenon which calls for special treatment (as in Descartes and Locke), but as a case continuous with what is going on beyond the borders of our conscious perceptions throughout waking experience. Sleep is fitted into a larger whole. An explanation that is *ad hoc* is always suspicious. It suggests that the terms of the explanation are artificial, having been invented specially for the case in question (*ad hoc* means 'for this'). But there is no '*ad hoc*ery' about Leibniz's explanation: it accounts not only for mental activity during sleep, but also for a class of (peripheral) mental activity throughout waking life. Ideally, a theory of consciousness and a theory of sleep should be cut from the same cloth in this way.

Charles University, Prague.

References

[1] The Second Meditation, CSM II 18; AT VII 27. My references here are to *The Philosophical Writings of Descartes*, translated and edited by Cottingham, Stoothoff and Murdoch (CSM), Cambridge, 1984 and to the *Oeuvres de Descartes*, edited by Ch. Adam and P. Tannery (AT), revised edition, Vrin, 1964-76.

[2] CSM II 247; AT VII 357

[3] The terms 'withdrawing', or 'retiring' should, of course, be understood in a metaphorical way when talking of the mind in Descartes: *res cogitans* is not really in space and therefore cannot 'go off' elsewhere. What I mean is that there is a suspension in the causal relation between thinking and the body.

[4] Animal spirits were swiftly moving fluids that seventeenth century philosophers used to explain the workings of our the brain and the nervous system. In particular they were responsible for making the tracks in the brain that were the physical basis of memory. Malebranche's view of sleep and states of unconsciousness is in his *Recherche de la vérité*, translated as *The Search after Truth*, by T.M. Lennon and P.J. Olscamp, Cambridge: CUP, 1997, III.i.2 (1)

[5] The faculty of pure intellection *(pura intellectio)* is described by Descartes at the beginning of his Sixth Meditation.

[6] Malebranche, ibid.

[7] E.II.i.13. Here and elsewhere I use the standard practice of referring to Locke's *Essay concerning Human Understanding*, (originally published 1690), by the numbers of the book, chapter and section. I have used the critical edition of the *Essay*, edited by Peter Nidditch, Oxford: OUP, 1975. There is really no substitute for this edition.

[8] E.II.xxvii

[9] E.II.i.11.

[10] Leibniz's view of sleep is to be found in his *New Essays on Human Understanding*, translated by Peter Remnant and Jonathan Bennett, CUP, 1996, pp. 112-118

[11] Ibid., 115.

[12] J.R. Searle, *The Mystery of Consciousness*, London: Granta, 1997, p.5.

Paul Sheehy

A Note on the [Puzzle] of Trust

Introduction

We are familiar with the notion of trust. Trusting another person is part of what it is to stand in relations of love, friendship and co-operation. Trust is central to the faith characteristic of the membership of a religion. Trust is presupposed in the anguish of betrayal. It is natural to think that either someone is acquainted with trust or else she is in some profound sense deficient or lacking in a form of knowledge vital for a full sense of humanity. That deficiency, moreover, is not a source of criticism but of sympathy. For, an ignorance of or inability ever to trust is to be forced to endure a stunted, curtailed kind of engagement with a world of other persons. Trust is both commonplace and vital. In its ubiquity we perhaps lose sight of its importance until we find ourselves unable to act because we cannot trust the other(s) now or because our trust leads us to harm.

There is, then, one way at least in which there is absolutely nothing puzzling about trust. People do for the most part stand in trusting relations. Taking myself to be typical in this respect I trust certain others and I am in turn trusted by others. The puzzle arises when we turn to the question of why and how trust is possible. When a person trusts another or others she is displaying a confidence in them. Intuitively, we call this attitude *trust* when the confidence outstrips or outreaches the grounds which one might reasonably or ordinarily regard as giving rise to that confidence. On an influential view of human nature it is difficult to understand why such an attitude to others would arise as a widespread phenomenon. Strip human nature to its essential elements and we see that we are rational, maximising agents ultimately driven by self-interest.[1] To trust another is to expose oneself to the risk of betrayal; to be trusted by another and to act in a trustworthy fashion may be to forego an opportunity to seize something to one's own immediate advantage. The puzzle is then, first, why we trust and why it is widespread if we are driven by our own self-interested concerns. A second aspect of the puzzle is whether our intuitions about the nature of trust ought to be taken at face value. Perhaps with respect to this second part you may not have any obvious intuitions; or now you turn to reflect upon them they appear confused or uncertain; or perhaps you are firm and confident in your intuitions – but why? That is the call to conceptual analysis; to get clearer about what we mean by a term and when it is apt to deploy it.[2] If we are to understand love, friendship and the bases of social co-operation we had better attempt to elucidate the nature and role of trust in the social world.

In the present paper I cannot hope to undertake that ambitious task. Instead I shall attempt to adumbrate one part of any response to the problem of determining why trust is possible. I consider what we *mean* by the term trust.

Defining Trust

A typical dictionary definition identifies trust as 'reliance on and confidence in the truth, worth, reliability etc. of a person or thing'.[3] On this view trust is seen as a kind of reliance. For one person to trust another is for her to believe that the other can be relied upon. That is, if I am to trust some other person, then I must have some way of assessing the risk of their failing to act in the appropriate fashion. The fundamental problem is one of overcoming ignorance about how others will act. In some circumstances there may be no real difficulty. When the robber has a gun pressed against the victim's head there is little scope for doubt about how the victim will respond to the request to hand over her money. The robber can rely on the victim. For trust to be possible amongst people, then, there has to be some basis on which they can regard each other as being sufficiently reliable. The difficulties that arise when we cannot be confident in the reliable performance of others are modelled in

Hobbes' famous account of the state of nature. In the absence of an effective sovereign authority our nature brings about a situation of instability. As Hobbes observes:

> So that in the nature of man, we find three principal causes of quarrel. First, competition; secondly, diffidence; thirdly, glory. The first maketh men invade for gain; the second, for safety; and the third, for reputation. The first use violence, to make themselves masters of other men's persons, wives, children, and cattle; the second, to defend them; the third, for trifles, as a word, a smile, a different opinion, and any other sign of undervalue, either direct in their persons or by reflection in their kindred, their friends, their nation, their profession, or their name…Whatsoever therefore is consequent to a time of war, where every man is enemy to every man, [there is] continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short.[4]

The problem is plain. We need to co-operate in order to establish some degree of stability. Yet, how in the state of nature can we establish the degree of assurance that is surely required in order for individuals to co-operate? How can we come to trust each other? Hobbes approach is instructive because he is not supposing that we take the state of nature seriously as a historical situation from which we emerged into a society organised through the institutions of the state. Rather, we are presented with a thought experiment whose aim to render vivid the problems we (as socialised moderns) would encounter should there be no state. Those problems arise out of our very nature, a nature explicated (to put matters crudely) in terms of the maximising rational agent model. No one can rely on agreements to mutual restraint of competition which they might make with others. For no one has any reason to assume that others will keep their word

> If a covenant be made, wherein neither of the parties perform now, but trust one another; in the condition of meer nature…upon any reasonable suspicion, it is void. (But if there be a common power set over them both, with right and force sufficient to compel performance, it is not void.) For he that performeth first, has no assurance the other will perform after, because the bonds of words are too weak to bridle men's ambition, avarice anger, and other passions, without the fear of some coercive power; which in the condition of meer nature, where all men are equal, and judges of the justness of their own fears, cannot be supposed. And therefore he that performeth first, does but betray himself to his enemy; contrary to the right (he can never abandon) of defending his life and means of living.[5]

The answer is that the establishment (or existence) of an effective sovereign power will put in place the potentially coercive framework that will enable individuals to rely on one another (for the most part): if you fail to keep to your agreements then the law provides the sanctions to punish you. I can rely on the self-interest of my counterpart and so come to the reasonable belief that she will keep to her word. The fact that there are sanctions constraining what it is in your interest to do cannot rule out the possibility that you will cheat. Assuming the sanctions are real and you are rational I can judge that you are likely to act reliably.

Now, one might object that there is more to trust than simply coming to a judgement as to whether someone else can be relied upon to keep their word. For on a simple reliance model of trust it seems that trusting someone becomes a function of

assessing the probabilities. Thus trust would be:

> a certain level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action.[6]

The difficulty in regarding trust as a belief about the reliability of others is in how to explain how we form such beliefs. In the case of the robber he has very good inductive grounds for relying on certain patterns of behaviour by victims, but this is in part precisely why that relationship does not look like the kind one can describe in terms of trust. Individuals trust one another under circumstances of uncertainty and risk. Trust is a response to a form of ignorance, an epistemological shortfall, that arises in certain circumstances. For example, imagine a Jew fleeing a Nazi death camp who comes to a church. He may trust a priest to assist him. There is clearly a risk in exposing himself to the goodwill of the priest for in the circumstances the escapee simply lacks the information that would enable him to come to a judgement on whether the priest is sympathetic to the Nazi regime or has other pressing reasons that would prevent him from offering help. Nonetheless, we surely want to say that trust can be placed in the priest because of *who* he is and what that entails. Of course, trusting someone cannot count as a guarantee of their performance. Trust has in this sense an asymmetrical quality. It flows from one person to the other, awaiting reciprocation but vulnerable to abuse. That *is* the risk. If I know enough about the situation to reliably assess the probability of someone performing, then I would appear not to need to trust them. I could simply rely on them. In a world in which there are psychics capable of reading minds one would surely be amazed to hear them talking of any need to trust other people. The puzzle is not how I can go about gathering enough information to do the sums working out the reliability of others, but what can justify or explain taking a risk on the goodwill of another when that information is not to hand.

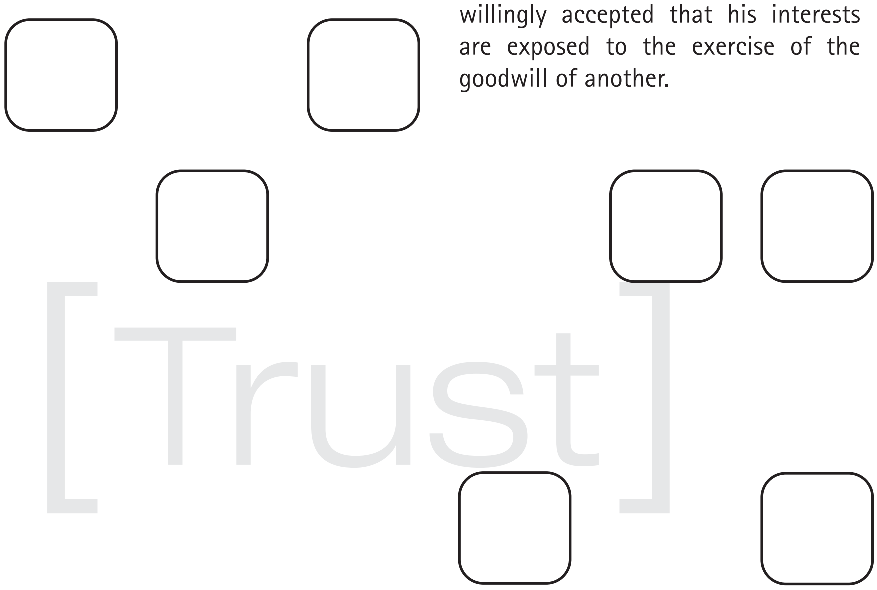
Well, one might note that Hobbes has just the right kind answer. We cannot trust each other left to our own devices. However, establish a power to force us to act decently (this is not Hobbes' way of putting matters) and we have just the form of assurance required to overcome the reluctance to take a risk on others. The difficulty with this approach is that it may just be at odds with how we think about trust. This is not to say that we should just accept our countervailing intuitions, but we ought to explore why an objection arises and whether we should stick with our intuitions or revise them.

The intuition I have in mind about our everyday understanding of trust is that it is distinct from mere reliance. The situations in which we employ the terms are not co-extensive. An account of trust just in terms of reliance may also appear to just collapse the distinction between reliance and trust. If to trust another is simply to rely on them, then there is an open ended range of ways in which I might properly judge them to be reliable in the circumstances in question. The grounds on which I determine the reliability of the other might include their stupidity. They can be relied upon to help me move flat because they cannot see how I regularly exploit their goodwill. Or, I might be able to rely on the co-operation of someone because I am her boss. I know she is ambitious and she will go to great lengths to impress me.

Consider once again the attitude of the fleeing Jew to the priest. We can hypothesise another situation of flight. A group of hardened criminals have escaped from prison. As they dash from the prison grounds they come to a wide and dangerous looking river, the howling guards and their dogs hot on their trail. Luckily the prisoners happen upon a robust rowing boat.[7] In order to effect their escape they must co-operate in the rowing of the boat and in attempting to navigate the frightening currents of the river. As rational, self-interested individuals with a common goal each of them can rely on the others to do their bit. In our story it is clear to each of them that his own singular goal is only achievable through co-operating with the others, and each knows that each of the others knows this to be the case.[8] We should say that the prisoners can rely on one another to do their part in escaping the guards. Once they have effected an escape, or at least one of them believes this to be the case, we have no reason to expect continued co-operation in the absence of some new external pressure. It is simply discordant with the way in which we usually apply the concept of trust to describe the situation as one in which the prisoners trust one another. Or is that simply to beg the question? Why ought we not talk of the prisoners standing in a relationship of trust?

The response to this challenge is to expose within an understanding of trust some feature that explains why the prisoners are not (obviously) trusting each other. The problem with the view that trust is reliance is that it fails to acknowledge that we trust others, and so see them as reliable, because we regard them in a certain way, and furthermore, in trusting them we are joined with them in a way of seeing the world. Our trust arises from our ways of seeing and interrelating with others. To see what we mean by trust we need to both generate cases where it appears to be salient to our understanding of what is going on, and to think hard about why.

First, trust may involve a willingness to be exposed to the risk of default by another because of who they are. Here one has not made any calculation about the likelihood of someone not acting in a certain way. Rather, the very thought of such a calculation does not enter into consideration. Following Annette Baier's way of putting matters trust is the willing acceptance of vulnerability to harm that others could inflict, but which we judge they will not in fact inflict.[9] A trusting agent is one who has willingly accepted that his interests are exposed to the exercise of the goodwill of another.



A second feature of trust is that it involves an attitude or feeling towards another; it is (in part) constituted by an emotional state.[10] To trust another is to have an attitude towards them that has a certain feel. The way trusting feels does not on the face of things appear to be like the way being in pain or despair is. Nonetheless, the way it feels to trust can be expressed in terms of the attitude of 'optimism that the goodwill and competence of another will extend over the domain of our interaction with her, together with the thought that the one trusted will be directly and favourably moved by the thought we are counting on her'.[11] That trust has a certain presence in our psychological state is also suggested by the sense of ill-ease, perhaps tension or even distress, experienced when one realises that trust is missing.

Trust is not just a matter of reliance because when we trust someone it is our attitude towards them that is essential in the belief or judgement we have with regard to their performance. Trust can be conceived as a complex of cognitive (belief-based) and affective (emotion-based) states. These states play an important role in securing co-operative behaviour in the face of risk, particularly in situations or contexts in which the extent, degree or probability of risk is (practically) incalculable. There is a further element in our understanding of trust: the idea that there is a non-instrumental good in trusting, which we may describe as a value that is *internal* to the practice of trusting. For I wish to propose that there is something that is good in trusting in *itself*. That is, there is a value in standing in relations of trust.

Thus far I have spoken in terms of one person trusting another, and played upon the sense in which trust is centrally characterised by the exposure to risk on the part of the trusting agent. Any elucidation of the concept of trust must attempt to cast light on the conditions in which some other can be regarded as trustworthy, and not simply adequately reliable in the circumstances.

Of course, by trusting each other we may be able to achieve through co-operative action that which would otherwise have been beyond us acting individually. Distinct from the good of the achievement of that goal is the value of being a certain way. That way is constituted by the very act of trusting and its component attitudes and beliefs. When a person trusts another – and that trust is not the subject of abuse - she comes to be assured that there is a commitment between them, a reciprocal bond upon which genuine or well-grounded trust rests and a shared understanding of the world presupposed in the act of trusting. If this is right, then by trusting we do not simply make possible the instrumental good of achieving shared ends or, indeed, of sustaining relations that are good in themselves such as friendships. Rather, by trusting we are engaging in a practice that is also in itself valuable and whose value inheres in the very relationship presupposed in the trust.

Care needs to be taken here in explaining just what is valuable about trusting. Trust can, of course, be abused. Sometimes the person investing trust in another is simply mistaken or foolish to do so. If in spite of all the evidence to the contrary I trust someone then I look to have erred. Trust is a relationship forged in the absence of reliance determining reasons, not one that can be sustained in the face of reasons that ought to compel me to judge that the other cannot or will not act as if he regarded my interests with goodwill. Sometimes it will be a close call and at others not so. For example, if my friend with a compulsion for goldfish flesh offers to look after my fish while I am away, I am a fool to trust him. Indeed, if I care about the fate of the fish I am worse than a fool. Sometimes more than goldfish hang on our decisions on whether we ought to place trust another. In other cases we are not fools or naïve to place our trust in others. Certain of the relations in which we stand seem to presuppose trust in. All we need to know to take a risk on the other is who they are – instances of such trusted parties might include parents, teachers, priests and lovers. In virtue of the relations in which one stands to them there ought to be trust. That is, to be a friend or lover is partly constituted by bearing the strain of being an object of trust. If your parent, friend or lover asks why they ought to be trusted, then one may fear that they just do not grasp the proper nature of the relationship, or that the question signals the demise of the relationship.

When one person trusts another and they are justified in doing so the trust arises because they can understand each other in the appropriate fashion for the risk to be taken by the one and the faith or confidence being reciprocated in the actions of the other. Furthermore, when one is trusted the response typically extends beyond simply responding appropriately in the sense that trust does not spring forth from a vacuum. To trust (and be trusted) is part of the typically complex array of practices and attitudes through which the domain of social interaction is constituted. On this view trust is to be understood as a way of regarding others which is embedded in the networks of social practices. Widespread trust is possible because it is presupposed in our ways of going on together. That capacity to trust is as secure or as vulnerable as those practices through which relations, institutions and groups are sustained, and in particular the capacity of those practices to withstand the pressures of critical reflection on the part of those engaged in them and the challenges of other practices.

Now, to return to the question of the value of trust. The value in trust is the value we enjoy through standing in a trusting relationship. Moreover, the value in trust is objective in the sense that the good is not a question of how I feel. Rather, the relationship both forged by and constituting the trust between us, is the source of a good in virtue of which our lives go (at least potentially) well. Trust is presupposed in the nature of certain relations; that is, trust is part of what it is to stand with another(s) in a relationship of love or friendship. In standing in these kinds of relations each of our lives is enriched.

My observations on the meaning of trust do not amount to an argument that this is how we should indeed understand that term. Rather, they represent an attempt to analyse how we can think about a concept, and as such can only represent a fragment of an endeavour to examine the nature of trust. That enterprise might need to abandon the attempt to provide a univocal analysis for the idea of trust may simply not be a settled term with a single application across different domains of discourse. I shall conclude the paper by briefly assessing if the analysis of trust offered meets certain bare requirements one might expect from an account of trust.

According to Jones 'an adequate account of trust should be able to explain at least the following three fairly obvious facts about trust: that trust and distrust are contraries but not contradictories, that trust cannot be willed, and that trust can give rise to beliefs that are abnormally resistant to evidence'.[12]

The account given accepts that at a general level distrust is a contrary of trust. Bob and Jane may just have no attitude or distrust towards each other in the relevant context. The failure to trust is not necessarily equivalent to distrusting, standing in a relation of pessimism about the goodwill and motivation of the other. In some contexts, though, the failure or inability to trust may signal distrust. If I do not trust my wife when she sets out her plans for the day then it can not be explained as lacking a particular attitude towards her, as being locked in 'neutral'. The thicker or deeper the relations in which we stand then, perhaps, the more compressed becomes the space between trust and distrust.

That trust can not be willed seems right and supported by the account given. Emerging from our interactions, to trust means that one stands with others and sees the social world (in the appropriate context) in a distinctive way. I might need to decide whether the grounds for trust are really present, I might reflect on my feelings towards the other(s). Such considerations impact upon the stance I take towards the world. However, I merely go wrong if I begin to perceive the world as if I do trust certain others; that is, if I *decide* to act as though I enjoyed the perspective of one who stands in certain trusting relationships. To simply decide – to, as it were, declare to myself - that I trust someone may induce in me the beliefs and feelings associated with trust. However, their source is not located in a relationship with the other person, and those beliefs and emotions are to be regarded as counterfeit. This need not mean that trusting relations are therefore personal in being proximate or close in social space or face-to-face. For I can trust those who I have never met through the mediating chains of relations and practices. So I may trust the leaders of my faith or community. However, to trust in a way that is appropriately grounded in the ways in which I am connected or related to the other I need to be so joined. It is not simply down to my act of will.

When individuals trust one another there is a resistance to accept evidence that undermines that trust. Again this is conspicuous where trust is deep and where it is often associated with other affective attitudes. Bob may soon learn that he is a cuckold. Indeed, the evidence of his wife's infidelity has been mounting. For an outside observer it appears nearly impossible that he could have missed so much that was so obvious. Yet, Bob trusts his wife because he loves her. Because of his understanding of the relationship in which they stand – or more accurately, the relationship in which they once stood and which he believes they continue to stand in - he interprets evidence of his wife's adultery into innocence. When we trust coherence is brought to part of the world. That coherent form serves as a constraint on how we respond and act and as a filter through which evidence is interpreted. Trust is sometimes difficult to undermine because we do not merely have a desire for coherence and stability. It is rather that the form or structure of our world is determined in large part by the nature of those relations in which trust is so central. A world that is dear to us, familiar and one that constrains our understanding is itself threatened by the dissolution of trust. Perhaps our resistance to the evidence against trusting resides in the living of a kind of life.

References

1 An oversimplifying account of this nature applied in direct empirical enquiry is doomed to misconstrue every individual act as being motivated either by the play of short-term consideration or as interpreting every motivation as being directly connected to the agent's interests. Things are likely to be considerably more complex. For example, not all social structures, practices and attitudes need be explained in terms of an agent's awareness of self-interest. It may be the case that they can arise and be sustained because they are functional for the self-interest of agents individually or for the group as whole. Alternatively, there may be *no* interest promoted by such attitudes or practices, but they do not serve to undermine or compete with such interests. The present point is that the rational maximising agent concerned with her self-interest has featured as the main character in the ways in which social interaction has been modelled in philosophy and the social sciences. The figure of the 'economic man', *homo economicus*, has loomed large in economics. This paper neither endorses nor presupposes that view of human nature, but suggests that such a view prompts a puzzlement about trust. Of course, there are alternative views about human nature. For example, one might hold that we are disposed to co-operate and be other regarding. In which case one might wonder why trust is accorded such importance.

2 Philosophical enquiry in general is interested in the concepts we employ in our ways of going on – in what we say. We are particularly interested in those concepts which apply with the broadest scope so that they frame or delineate the subject matter at hand. A cluster of apparently unrelated statements are united by their dependence on a single concept. Taking an example from metaphysics one might note that Paul is short, the dog happy and the housing market overpriced. In each of these statements a claim is made that some entity possesses a particular property. The concept of *property* is central to an understanding of what is being asserted. In metaphysics we can also include, *inter alia*, cause, space, time, entity, substance, identity in a list of basic or 'domain concepts'. See Thomas, G. (1993) *An Introduction to Ethics*, London: Duckworth pp. 2-4 for an excellent discussion of conceptual analysis and its role in ethics. A rewarding and more advanced discussion of conceptual analysis is Frank Jackson's (1998) *From Metaphysics to Ethics*, Oxford: Clarendon

3 *Collins Third Edition* (1991). It is thought that trust is derived from the Old Norse *traust*, which means confidence or firmness. This has been associated with the Indo-European root *deru* or *dreu* meaning to be firm or solid.

4 Hobbes, T., *Leviathan* Chapter 13.

5 *Ibid.* Chapter 14.

6 Gambetta, D. (1988) 'Can We Trust Trust?' in D. Gambetta (ed.) *Trust: Making and Breaking Cooperative Relations*, New York: Blackwell pp. 213–37

7 The story of the escaping prisoners is taken from Hume's example of two men rowing a boat in his

analysis of conventions. See Hume, D., (1740/1978) *A Treatise of Human Nature* 2nd Edition (ed. Nidditch), Oxford: Oxford University Press p.490

[8] For example this might become abundantly clear in their conversation and actions upon arriving at the river.

[9] See for example, Baier, A., (1994) *Moral Prejudices*, Cambridge, Mass: Harvard University Press p.152

[10] I leave to one side any further discussion of how we should properly understand an emotion, a topic which has attracted a considerable amount of attention in recent years.

[11] Jones, K., (1996) 'Trust as an Affective Attitude', *Ethics* 107.

[12] Jones op cit p15. We call terms (or propositions) contradictories if they cannot both be true and they cannot both be false. So, being alive and being dead cannot be true of anything at the very same time. If I am alive it is impossible for me to be dead, and it is impossible for me to be neither alive nor dead. Terms are contraries if they cannot both be true, but can both be false. Thus I cannot both love and hate you, but it is possible to neither love nor hate you. I may just be indifferent to you. So with trust, I may not trust you but that does not mean I distrust you. I may just have no attitude towards you at all as far as trust is concerned.

D J Sheppard

On Why the Philosopher Returns to the Cave

Introduction

It is the most familiar of scenes: hitherto confined to a world of shadows that since childhood he has been taught constitutes the only reality, a prisoner is released from his bonds and compelled to make the ascent out of the cave and into the light of immutable truth. It is the seminal image of the journey of philosophical enlightenment in the Western tradition. Summarised thus, however, it tells only half of Socrates' story: having undertaken the journey out of the shadows, the newly educated philosopher is obliged to return to the cave and rule in accordance with the vision of truth that he has been afforded. Glaucon has his doubts, but they are assuaged. The philosopher, he concedes, must return; 'there is no one else' (521d).[1]

The debate surrounding the philosopher's return to the cave is a constant of Plato scholarship. Why, having left the cave and witnessed the form of the good, does the philosopher forego his new found 'earthly paradise' (519c) and undertake the difficult task of ruling his former fellows? What is his motivation to return? The purpose of this essay is to examine Plato's answer to this question and its implications for the argument that is advanced in both the middle books of the *Republic* and, by extension, the dialogue as a whole. I shall argue that Plato provides a coherent account of the philosopher-ruler's motivation and that it is to be found in the political implications of his epistemological education in the theory of the forms. In conclusion, however, I shall suggest that there is a price to pay for the theoretical coherence of Plato's account in respect of the practical prospects for the philosopher-ruler's successful return.

I

Having completed his account of the philosopher's ascent, Plato considers what might be assumed to be a drawback in a potential ruler: a reluctance to govern. 'It won't be surprising,' Socrates suggests, 'if those who get so far are unwilling to involve themselves in human affairs, and if their minds long to remain in the realm above' (517c-d). Socrates would appear to have a point. He has already suggested that, if they returned to the cave, they would be blinded by the dark as once they were blinded by the light and make fools of themselves as a result (517a-d). Worse still, they would rightly return in fear of their lives: given the opportunity, Socrates reflects, the other inhabitants of the cave would be sure to kill them (516c-517a).

Socrates cannot be accused of sugaring the pill. Yet far from viewing his reluctance as a problem, for Plato it is an indication of the philosopher's fitness to rule. Socrates maintains that 'the state whose prospective rulers come to their duties with the least enthusiasm is bound to have the best and most tranquil government, and the state whose rulers are eager to rule the worst' (520d). The argument is that only those rulers who have experienced a way of life preferable to political governance – namely, the life of philosophical contemplation – will be just rulers, for having experienced the philosophical life they will not seek self-satisfaction through the attainment of political power. In Socrates' words, they will not 'hope to snatch compensation for their own inadequacy from a political career' (521a). Rulers eager to hold office inevitably compete for power, condemning the state to endless internecine strife. The just state requires rulers who hold themselves aloof from their task; who possess what, after Nietzsche, we might term the 'pathos of distance.'[2] In Nicholas White's summary, 'philosophising is essential to ruling because it is the activity that is preferable to ruling, and so the activity that the ruler must have available to him if he is to wish not to rule, where wishing not to rule is, paradoxically, what makes it possible for him to rule well.'[3] In short, the just ruler is a reluctant one.

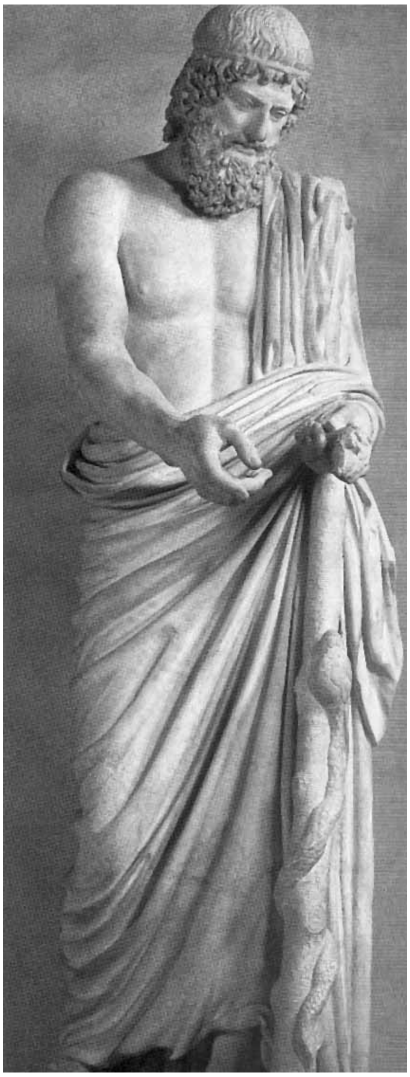
There are two points to make in respect of this argument. Firstly, whilst it may explain why the reluctant ruler is qualified to rule, it does not explain why he would necessarily feel obliged to do so. We shall return to this in due course. Secondly, Plato's insistence on the philosopher's reluctance raises a question mark against another aspect of his account. To consider this, however, it is necessary to examine the broader discussion that unfolds in the central books of the *Republic*.

II

Glaucon's agreement at 521d concludes a discussion that begins at 471c when Socrates is reminded that he has yet to explain how the ideal state might be realised. It will never see the light of day, he replies:

> ...till philosophers become kings in this world, or till those we now call kings and rulers really and truly become philosophers, and political power and philosophy thus come into the same hands, while the many natures now content to follow either to the exclusion of the other are forcibly debarred from doing so. This is what I have hesitated to say so long, knowing what a paradox it would sound; for it is not so easy to see that there is no other road to real happiness, either for society or the individual. (473d-e)

Plato sets himself an immense task in this passage, not simply in view of the contempt in which the philosopher is commonly held – the topic on which Adeimantus will soon hold forth (cf. 487b-d) – but in view of the argument that is advanced in the dialogue as a whole regarding the nature of justice. Plato's ideal state is constructed on a definition of justice known variously as the Principle of Specialisation or the Natural Division of Labour. Each individual is to perform the single task for which by nature he is best suited: 'one man one job' (434c). Yet, in his insistence that the ideal state can only come about if rulers become philosophers and vice-versa, Plato would appear to predicate the realisation of the just state on the seeming *in*justice of a 'philosopher-ruler'; of one man with, in effect, two jobs (philosophising *and* ruling). Viewed in this light, the magnitude of Plato's task in the central books is clear. For the just state to possess a just philosopher-ruler, Plato has to show not simply that philosophy and ruling are mutually tolerant roles – that the philosopher should rule because he is best qualified to 'multi-task' in this way – but that they constitute *one and the same role*. If this claim is not substantiated, then on Plato's own account of justice the realisation of the just state is significantly compromised.



In this connexion, the philosopher's reluctance to rule is of particular import. Plato argues that the philosopher's lack of enthusiasm for governance is supposed to guarantee that, upon his return, the ideal state will not be riven by internal conflict. His experience of, and preference for, the life of philosophical contemplation will deter him from seeking self-satisfaction in the life of politics. Yet according to Plato's conception of justice, such 'multi-tasking' is 'the worst of evils,' and guaranteed to plunge the state into conflict and disharmony (434c). In short, it is a prescription for injustice. Far from fulfilling the task set out at 473d–e, Plato's insistence on the philosopher's reluctance suggests a calamitous irony: the philosopher is just – according to the 'one man one job' thesis – only if he refuses to rule and remains outside the cave. Does Plato's account of how the ideal state is realised founder on this irony? Does it betray Socrates' claim that, in expecting the philosopher to return to the cave, a just demand is made of just men?

III

It is in answer to these questions that we return to the matter of the philosopher's motivation. I suggested that the fact of the philosopher's suitability to rule is not in itself a reason why he need feel compelled to do so. Nor, incidentally, is Socrates' additional suggestion at 520b–c that the philosopher will feel obliged to repay the society that provided for his education; justice as the repayment of what one owes was dismissed at the beginning of Book I (cf. 331d). To discover the true source of the philosopher's motivation we need to consider the specific details of his education.

When Socrates says that, however reluctantly, the philosopher ruler will have to return to the cave, Glaucon raises his own cry of injustice. Picking up on the implication that the lives of philosophy and politics are distinct, he objects that 'we shall be compelling them to live a life poorer than they might live' (519d–e). Socrates replies:

> The object of our legislation [...] is not the special welfare of any particular class in our society, but of the society as a whole; and it uses persuasion or compulsion to unite all citizens and make them share together the benefits which each individually can confer on the community; and its purpose in fostering this attitude is not to leave everyone to please himself, but to make each man a link in the unity of the whole. (519e-520a)

The recurrent suggestion that 'persuasion or compulsion' will be required to ensure that the philosopher fulfil his obligation is rather troubling. It is one thing to say that the philosopher will be reluctant to return but ultimately understand that it is his duty, quite another that force will be required. It is as though Plato does not fully appreciate the implications of his own account of the philosopher's education. On my argument, if, notwithstanding his reluctance, the philosopher requires 'persuasion or compulsion' to return, then he is not a philosopher; there is something fundamental he has failed to comprehend in the course of his education. To have properly understood the ascent out of the cave *is in itself* to have understood the duty to return and be motivated to obey it. Moreover, it is thus that the strict identity of the philosophical and the political demanded by the Principle of Specialisation is maintained.

What is it, then, that the true philosopher understands? It is agreed that the simile of the cave situates the ethical and epistemological concerns of the similes of the sun and of the divided line in a political context. Yet nothing specific is added in the detail of the simile that points to its explicitly political implications. The reason for this, as the true philosopher understands, is that the duty and motivation to return is inscribed in the theory of the forms itself.

From the very beginning of Plato's account of the true philosopher at 474b, the latter is distinguished from the mere 'lovers of sights and sounds' (476b) by his ability to understand the relation between the 'one' and the 'many.'[4] Each *eidos* or form, he says, is in itself 'single' or 'one', 'but they seem to be many because they appear everywhere in combination with actions and material bodies and with each other' (476a). However, the lover of sights and sounds does not understand this: 'Those who love looking and listening are delighted by beautiful sounds and colours and shapes [...] but their minds are incapable of seeing and delighting in the form of beauty itself' (476b). In short, they fail to understand the 'oneness' of the form in its 'many' sensible manifestations. The philosopher, on the other hand, 'believes in the form of beauty and can see both it and the particular things which share in it, and does not confuse particular things and that in which they share' (476d). He understands that two instances of beauty – to borrow Socrates' example – relate to each other not in terms of empirical resemblances between them, but in terms of their mutual participation in the form of beauty.

The process of understanding this relation is depicted in the philosopher's progress up the divided line and reiterated in a political context in the ascent out of the cave, culminating in the vision of the good in itself. The suggestion is that the relation between the form of the good and the other forms is analogous to the relation between the forms and their physical manifestations; the form of the good, as 'responsible for whatever is right and valuable in anything' (517c), understood as the 'one' in which the 'many' forms participate or share.

It is often argued that the philosopher understands the imperative to return in relation to this vision of the good. As Julia Annas writes in her influential introduction to the *Republic*, the philosopher-rulers 'know what is just because they have the knowledge that is based on the form of the good. Their return is demanded by the justice that prescribes disinterestedly what is best for all.'[5] On this reading, the philosopher's motivation is very abstract; he takes a wholly impersonal view of his interests and sets aside the life of contemplation because his judgements are made 'in the light of the impersonal good.'[6] Yet for Annas such an account begs the question why the philosopher would wish to sacrifice himself in this manner ('why should *I* do what justice requires?').[7] The argument that the philosopher-ruler would not perceive a conflict between justice and his own interest since he has been trained to understand himself as being 'merely' a part of the whole, is rejected on the grounds that it 'only raises the more urgently the question why in that case *I* should want to be a [philosopher-ruler].' 'Justice,' she says, alluding to the demand made of Socrates at the beginning of the dialogue (cf. 367d), 'was to have been shown to be in *my* interests. But now it requires that I abstract completely from my interests.'[8]

However, I would suggest that Annas's contention is predicated on what, from a Platonic point of view, is a false premise: that to understand oneself as part of a whole is thereby to be *'merely'* a part, and that to 'cease to care about my own happiness in a specially intimate way' – i.e. as an atomistic individual - is to 'positively stop being human.'[9] Whilst this may reflect Annas's conception of the human, it is not a reflection of Plato's view. Regardless of the offence to our liberal-humanist sensibilities, it is essential to recognise this if we are to understand the philosopher-ruler's motivation.

To this end, it is necessary to examine the philosopher's motivation not simply in relation to his crowning vision of the good, but his education in the forms as a whole. Recall Socrates' insistence that, in the just state, each man is a 'link in the unity of the whole.' It is a reminder that bears repetition, since in the light of the education in the forms it can be seen as the political counterpart of the epistemological relation between the one and the many.

By it, the philosopher understands his particularity – his belonging to the many – in relation to the one, to the community as a whole. Specifically, he understands his just participation in the whole as an obligation to rule. As in the epistemological relation between the particular and the form in which it participates, the philosopher understands that it is only in his proper participation in the whole that his own 'usefulness and value' (505a) is manifest. As a result, the philosopher-ruler does not consider his return to involve a personal loss or the means by which he surrenders his humanity and becomes *'merely'* a part of the whole. Phrased otherwise, he does not equate acting in accordance with the good with acting impersonally. Rather – and most importantly – it is only in his return to the cave that he becomes properly human. This is not to suggest, however, that the philosopher's motivation is consequently selfish, where selfishness is understood as deficiency in one's consideration for others. He understands his good *as* the good of the whole.

Such an understanding accords with the account of justice in the state and the individual in Book IV. According to the Principle of Specialisation, every individual in the just state understands the role for which he is properly suited. The latter is determined by the predominance in the individual's soul of one of its three 'parts': the rational, the spirited, or the appetitive. The majority, those in whose souls the desiring part is dominant, enter the class of 'artisans and businessmen' (434b); those in whose souls the spirited part is dominant form the military class; and those few in whose souls the rational part is dominant will become the ruling class (the class later named philosopher-rulers). Strictly speaking, there are no just individuals; only the just state in which each individual component plays its proper role. Considered apart from the whole, each individual is, qua individual, 'unjust,' since as an individual he lacks the harmony that is only to be found in the just collective. No one is self-sufficient (cf. 396b). Whilst it is indeed Plato's view that the best soul is one in which the rational part of the soul predominates (441c-442d), that 'supremacy' is only meaningful in so far as the soul understands its proper relation to the whole. There is nothing 'impersonal' about this understanding; on Plato's terms it is the apogee of self-knowledge. It is what the philosopher-ruler understands, and it is the source of his motivation.

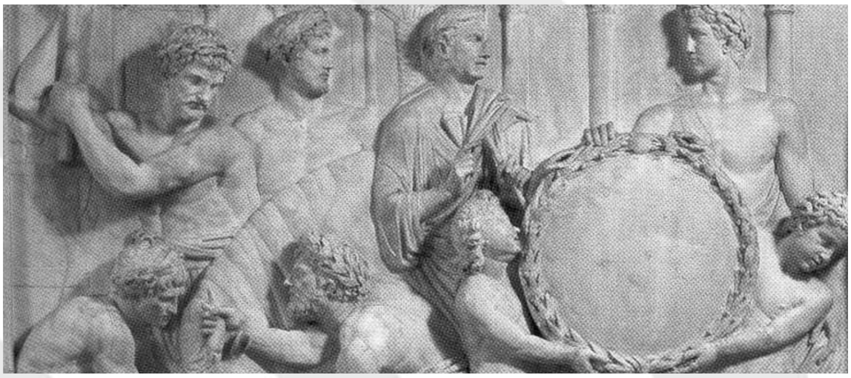
Thus, Glaucon's claim that an injustice is done to the philosopher in depriving him of the life of philosophical contemplation is a misnomer, and suggests that Glaucon has missed the point. What would be unjust is the disjunction between the philosophical and the political that his objection presupposes. However, as the Principle of Specialisation requires and as the education in the forms establishes, philosophising and ruling are one and the same task. Consequently, if, in insisting upon the philosopher's reluctance, Plato maintains that the life of philosophy is separate from the life of politics, then we have to conclude that in this instance he creates an unnecessary tension in his own account. He underestimates the extent to which the preceding account provides an explanation of the philosopher's motivation to return, and why the call to do so is indeed a just demand made of just men.

Conclusion

At the beginning of the essay I proposed that, if Plato cannot provide an adequate account of why the philosopher-ruler returns to the cave that accords with his earlier definition of justice, then the theoretical coherence of the argument advanced in the middle books of the *Republic* is undermined. I argued that such an account is to be found in the detail of the philosopher's education. However, there is, I suggest, a practical price to pay for this theoretical coherence that in its turn compromises Plato's vision.

It follows from the account of the philosopher-ruler that everything he needs in order to rule is contained in his education in the forms. It is thus that 'political power and philosophy come into the same hands,' and thus that the philosopher-ruler is motivated to return; yet it also the reason that we fear for the philosopher-ruler's chances of practical success. Why?

My concern centres on how the philosopher-ruler overcomes the hostility of his former fellows and establishes his authority. In Book IV we are told that in the ideal state governors and governed 'will agree about who ought to rule' (431e). But when in Book V attention turns to the practical matter of realising the ideal state there is a question mark over how this agreement is reached. The issue is highlighted in the simile of the ship (cf. 488a-489c). Ostensibly, the purpose of the simile is to account for why the philosopher is an outcast in contemporary society. The politicians of the day are compared to the crew who, though none of them understands the art of navigation, quarrel over who ought to command the ship. The philosopher is compared to the true navigator who has studied 'the seasons of the year, the sky, the stars,' and so on, but whom the crew regard as 'a word-spinner and a star-gazer, of no use to them at all' (489a). On this scenario, we wonder what would happen if the true navigator resolved to assert his claim to the captaincy. On the face of it, there is



little reason to suppose he would be successful. How is he to convince the crew that he is their rightful ruler? What means does he have at his disposal? The problem is that the navigator turns to the task unequipped with the rhetorical skills of persuasion or any other trick of the political trade that would appear essential if he is to obtain a hearing. It is difficult to conceive that he will not require such additional skills: the crew do not even believe that the art of navigation exists (cf. 488e). It seems that the best the navigator can hope for is to be ignored; should he persist, he will more than likely be thrown overboard.

Plato subsequently disputes this conclusion, though the basis on which he wishes to do so is unclear.[10] At one point Socrates says that the populace will have to be 'compelled' to listen to the philosopher-ruler (499b). A little further on the suggestion is that the dominion of the philosopher-ruler will be recognised 'if instead of bullying [the masses] you are gentle with them, and try to remove their prejudice against learning and show them what you mean by philosophers' (499e-500a). The latter would seem to be the only alternative available to the navigator, but again we wonder how he is supposed to go about the task.

As we have seen, there is a similar mix of optimism and pessimism in the simile of the cave. Socrates makes the point that upon his return the philosopher will 'blunder and make a fool of himself' (517d). But he adds that this stage is only temporary: in time the philosopher will re-accustom himself to the dark and 'see a thousand times better' than his former fellows (520e). Yet it is not the philosopher-ruler's sight that is at issue; rather it is his ability to convince the prisoners of the clarity of his vision. Once again, it is difficult to deny that the more realistic possibility is the suggestion that 'they would kill him if they could lay hands on him' (517a). After all, he will challenge every conception the prisoners hold dear, and to have any chance of success it would appear imperative that he possesses the most sophisticated rhetorical and persuasive skills if he is to 'turn' them.

Crucially, however, Plato's account precludes the possibility of the philosopher-ruler being thus armed. The claim that the just state will only see the light of day when philosophers become rulers and rulers become philosophers, and 'political power and philosophy thus come into the same hands' (473d), demands the strict identification of the philosophical and the political: all the philosopher-ruler requires to rule must be contained in the education in the forms. For this education then to be appended by rhetorical skills that are inimical to it is to undermine this identification, and to render the saviour of the just state the epitome of injustice (philosopher-ruler *and* rhetorician). Consequently, the philosopher-ruler must return without such resources. Perhaps it is the recognition of this practical blind spot in his account that causes Plato to oscillate between optimism and pessimism when considering the prospects for the philosopher-ruler's return. In the final analysis, it is difficult to sugar the pill. Little wonder if the philosopher appears reluctant.

References

[1] Desmond Lee's translation of the *Republic* (Harmondsworth: Penguin, 1987) will be followed throughout, amended only when it is necessary to assist clarity of exposition.

[2] Cf. Friedrich Nietzsche, *Beyond Good and Evil*, §257.

[3] Nicholas P. White, *A Companion to Plato's Republic* (Oxford: Basil Blackwell, 1979), 190.

[4] Cf. John Sallis, *Being and Logos: Reading the Platonic Dialogues* (Bloomington and Indianapolis: Indiana University Press, 3rd ed. 1986), 382ff.

[5] Julia Annas, *An Introduction to Plato's Republic* (Oxford: Clarendon Press, 1981), 266.

[6] Ibid. 267.

[7] Ibid.

[8] Ibid. 268-69.

[9] Ibid. 269.

[10] Cf. White (op. cit.), 171-72.

Pierre Cruse

On [Language] Truth and Logic

Language, Truth and Logic (LTL) is nothing if not ambitious. In a little under 150 pages, Ayer aims to resolve the major questions of philosophy. He doesn't propose to resolve the questions by actually *answering* them, but by showing them to be spurious. They are spurious, he thinks, because they concern questions of 'metaphysics'. And Ayer thinks he can show that metaphysical questions, and the answers which philosophers propose to them, are uniformly meaningless.

I think few would disagree that there is something slightly naïve about Ayer's project – or that he doesn't entirely succeed in carrying it out. However, the question I want to look at here is whether the basic idea behind the project is a sound one. Ayer's guiding principle – that a statement that is not verifiable is meaningless – is often criticised on the grounds that it is either baseless, or worse, inconsistent. However, I will argue that the situation for Ayer is not as bad here as is sometimes made out. That's not to say that I will agree with Ayer: I do not agree with many of the conclusions Ayer reaches, and I do not think he provides satisfactory arguments for them for reasons I will explain. However, I also think there is more to be said in his defence than one might think, even though it ultimately isn't quite enough to save his project. In the following, I will try to explain why this is.

I The argument behind Language, Truth and Logic – The Verifiability Criterion

Ayer's aim, as I said above, was to argue against 'metaphysics'. Metaphysics, as he understands it, is the activity of trying to discover matters of fact which cannot be known on the basis of experience. Ayer thinks he can show this activity to be illegitimate, by showing that the propositions metaphysics deals in are meaningless.

In order to demonstrate this, Ayer proposes a criterion by which we can sort meaningful from meaningless statements. The criterion he calls the 'verifiability criterion of meaning' or 'principle of verification' (POV). The principle of verification says that a statement is factually significant for a person if and only if 'he knows how to verify the proposition which it purports to express – that is, if he knows what observations would lead him, under certain conditions, to accept the proposition as being true, or reject it as being false'. In other words – as he explains himself later – a statement is meaningful if and only if there are possible experiences that would count as *evidence* for or against it.

The POV as stated applies to 'factual' propositions, that is, those which aim to make true or false statements about the way the world is. In addition, however, Ayer allows statements to be meaningful when they are *analytic* or *tautologous*, that is, true simply in virtue of the words involved. Examples of such statements are 'ewes are female', or 'a numismatist collects coins'. Statements like this are meaningful, but don't strictly say anything about the world – rather, Ayer says, they express relations between the meanings of the words involved. However, Ayer thinks that if a statement is neither analytic, nor empirically verifiable, then it is meaningless.

Ayer gives as an example of the sort of claim this criterion will class as meaningless, the sceptical claim that the world we perceive is 'unreal'. This sounds at first like a meaningful (if unlikely) supposition. However, no observations could possibly count for or against its truth: observations could only confirm how we experience the world rather than the nature of the reality behind our experience. Nor is the supposition that the world is radically different to how we experience it analytically false – its falsity does not follow simply from the meanings of the words involved. Therefore, by Ayer's criterion the claim that the world behind our perceptions is unreal is meaningless; it has no chance of even counting as true or false.

Having introduced the principle of verification, Ayer goes on in the rest of the book to apply it to various areas of discourse, philosophical and otherwise. The areas he looks at are primarily areas where we might think metaphysical propositions are involved, but which seem to be perfectly meaningful. Thus, his aim in discussing them is to show how we can understand them without the intrusion of meaningless metaphysics. There isn't room here to go through everything Ayer says, but I'll mention a selection of his ideas to give the general picture.

II The Elimination of Metaphysics

The first question Ayer raises is the nature of philosophical inquiry. As he is, essentially, writing a work of philosophy, he needs to show that doing philosophy does not essentially require us to inquire into metaphysical questions – factual questions which cannot be decided by appeal to experiential evidence.

To counter this suggestion, Ayer argues that philosophy is in the business of *analysis*. Analysis, for Ayer, is simply inquiry into the meanings, or as he prefers to put it, the 'definitions' of words. A familiar example of this kind of inquiry (though not Ayer's) might be the case of analysing the notion of 'knowledge'. Traditional analyses of knowledge consist in putting forward some definition of the term 'knowledge' in terms which do not presuppose it, such as 'knowledge is justified true belief'. This definition will either be analytically true, or (as in this case) we will notice that there are cases that fall under the analysis which we would not class as knowledge, or vice versa, in which the definition will be analytically false. The important point about this kind of inquiry is that there is nothing 'metaphysical' involved in it. We are merely looking for some way of characterising the situations in which we would apply the term 'knowledge', and the knowledge we get from our inquiry is essentially just of analytic truths. If we succeed in doing this, we learn something not about the world, but about the meanings of our words.

Ayer's applies this idea in a more controversial way in his analysis of statements about ordinary material objects (see LTL, pp 64-69). The problem here is that on a common understanding material objects are thought to be 'substances' which exist in the external world entirely independently of our experience of them. But talk about metaphysical substances that exist independently of our perception of them is unverifiable, since Ayer thinks there are no experiences which would count as evidence for the truth of sentences about these substances. However, given the frequency with which we talk about material objects it would be highly implausible for Ayer to dismiss all such talk as meaningless. His solution is to propose that statements about material objects are not really about mind-independent substances at all. Rather, they are what he calls 'logical constructions' out of the immediate contents of our experience, also known as 'sense-contents'.

Ayer's idea is that we can translate any sentence that contains the word 'table' into an equivalent statement that refers only to sense-contents. Roughly , his claim is that a material object can be defined as a group of connecting sense-contents that bear a certain kind of relation of similarity to each other. Thus, what characterises the table in front of me is the fact that the sense-data in a certain (table-shaped) region of my sensory field are all similar (brown, hard, etc.) and are different from the directly adjacent parts of my sensory field (which are empty). Statements about tables, then, can be translated into other statements that refer only to relations of similarity between sense-data by referring to relations between brown, hard elements of my sense-field and their relations to other parts of my experience.

Philosophy, then, does not violate the principle of verifiability, on the grounds that philosophical knowledge is ultimately not 'factual', but just about the meanings of words. Another area in which the same applies is mathematics and logic (see LTL, pp. 75-87). The apparent problem here is that mathematical propositions like '6+7=13' appear to give factual knowledge, but are not known on the basis of sense-experience – you can know that 6+7=13 without ever counting any actual objects. Again, however, Ayer argues that mathematical propositions are true just in virtue of the meanings of the constituent symbols ('6', '7', '+', '=' and '13'), so they are not really 'factual'. The same goes for logical principles.

A further area in which Ayer applies the principle of verifiability is in ethics (LTL, pp. 102-114). As Ayer notes ethical propositions pose a problem for his criterion. Consider an ethical statement such as 'killing innocent people is wrong'. This statement certainly seems to be meaningful. However, it isn't obviously analytic, since it isn't clear you could come to know it is true simply by virtue of reflecting on the meanings of the words involved in stating it. Nor is it obviously synthetic, since it is not clear what observations would count for or against it. But if it is neither analytic nor synthetic, then the principle of verifiability would class it as meaningless. This is clearly unacceptable.

One obvious tactic that Ayer *could* pursue in accounting for moral and other value judgements would be to adopt what we might call a descriptivist theory of ethical terms. According to a descriptivist theory, ethical judgements are equivalent to statements of non-ethical fact. One such theory, for example, is utilitarianism. According to utilitarianism to say that some action is right is to say it is conducive to the general happiness. However, Ayer thinks this kind of analysis is unsatisfactory. For Ayer, we can recall, a philosophical theory such as utilitarianism must – if it is true – be analytically true, true as a matter of definition. However, he argues that since it is not contradictory to assert that something is morally right without being conducive to general happiness, utilitarianism cannot be analytically true. The same goes, Ayer thinks, for other 'descriptivist' theories that equate being right or good with the satisfaction of some factual condition.

Ayer's alternative is to deny that moral judgements express propositions at all. Or rather, they do not convey any propositions other than the factual propositions that are already involved in stating them. If I say 'killing innocent people is wrong', then I am expressing no proposition, but merely my disapproval of killing innocent people. To use Ayer's way of putting it, stating that killing innocent people is wrong is like writing 'killing innocent people' with a special kind of exclamation mark designed to show that I am expressing disapproval. If, on the other hand, I say 'it was wrong for Pol Pot to have killed so many innocent people', my assertion has some factual content – it asserts that Pol Pot killed so many innocent people – but in addition my statement only serves to express my disapproval of Pol Pot's killing those people. Ayer's theory is therefore sometimes known as the 'boo-hooray' theory of moral judgements – saying that killing innocent people is wrong is very much like saying something along the lines of 'killing innocent people – Boo!' – since 'boo' expresses disapproval of something without making a factual statement.

III Do Ayer's conclusions follow?

Ayer applies his criterion in a number of further areas, but there isn't time to outline them here. However, I hope this gives a flavour of Ayer's general project. The basic theme, as we have seen, is that Ayer wants to go through various areas of discourse (talk about material objects, mathematics, ethics, and also truth, probability, other minds, and other topics) to show how they can be interpreted in line with the basic guiding principle – the principle of verification.

There are therefore (at least) two key questions we might ask about the success of Ayer's arguments. The first question is whether his conclusions follow from the principle of verification – whether, if the principle is true, we are committed to the views that he says we are committed to. The second question is whether the principle of verification is in fact true.

Although it is a very interesting question, and crucial for a full evaluation of Ayer's views, I want to pass over the first question relatively briefly. However, I will make a couple of comments. The first is that if the principle of verification is true then it is clear that his overall strategy is a coherent one. For a start, the truth of the principle would give him a good reason for saying that many 'metaphysical' conclusions are unverifiable. It also seems a good idea, if he is to defend the principle, to show that it does not class the majority of our talk as meaningless – suggesting that it might have to be analysed in unfamiliar terms.

However, even if we grant this it is not clear that *all* his conclusions will follow. Consider for example his claim that statements about ordinary material objects can be defined in terms of sense-contents. The principle of verification says that a factual statement is meaningful only if it is verifiable, in the sense that there can be experiences that count as evidence for or against it. But on an ordinary understanding there can be evidence that counts for the existence of a real, physical table – seeing the table in front of you, for example! To prove that tables are really constructions from sense-contents Ayer needs to show not only that sense-contents don't conclusively verify the existence of a table, but that they provide no evidence for it whatsoever. Presumably Ayer does think this, otherwise he would not have felt the need to argue for this view. But he doesn't really put forward an argument for it. For this reason, I think that some of his conclusions are doubtful even if we grant the strong principle of verification.

Nevertheless, there are other areas – such as ethics – where he clearly *does* need to do some work. For example, if we think there is a plain and irreducible fact that, say, killing is wrong, then it is very difficult to see how it could be verifiable in Ayer's sense. So if he thinks this statement is meaningful at all, he needs to explain how that can be. So I think that much of what he says is at least well-motivated given his starting point.

IV Problems with the Principle of Verification

I want now to focus on whether the principle of verification is actually true. I will begin by looking at some problems with this principle.

The first question we might ask is why Ayer thinks the principle of verification is actually true. When he introduces the principle in the first chapter of LTL, he has the following to say about it:

> As to the validity of the principle… a demonstration will be given in the course of this book. For it will be shown that all propositions which have factual content are empirical hypotheses; and that the function of an empirical hypothesis is to provide a rule for the anticipation of experience. And this means … that a statement which is not relevant to any experience is not an empirical hypothesis, and accordingly has no factual content. But this is precisely what the principle of verifiability asserts. (LTL, 41)

Now this appears to be an argument for the principle of verification. The crucial premise is that the function of an empirical hypothesis is to predict experiences. The idea is presumably that an unverifiable statement will predict no experiences, and will therefore fail to be an 'empirical hypothesis', and will therefore have no factual content – in other words it will be meaningless. Ayer says that he will demonstrate later in the book that the function of an empirical hypothesis is to predict experiences. But in fact, when the issue comes up, he only says:

> What is the purpose of formulating hypotheses? Why do we construct these systems in the first place? The answer is that they are designed to enable us to anticipate the course of our sensations (LTL, 97)

This is not the argument we were hoping for. It is merely a statement of the view he needs to prove. Moreover, it is hardly the sort of thing we can just accept on trust. Everyone will agree that it is *a* function of formulating hypotheses to predict experiences. But it is a big step to conclude from this that the *only* reason we formulate hypotheses is to predict experiences, which is what Ayer needs for his argument to go through. On the contrary, someone who disagrees with the principle of verification (a metaphysician, for example) will argue that one important reason for forming hypotheses about the world is simply to try to acquire knowledge of the truth, whether or not doing so helps us predict experiences. Ayer needs some reason why someone who holds this view is wrong. But he doesn't, as far as I can see, provide one.

I think it is fair to say, therefore, that Ayer provides little *direct* justification for his crucial principle of verification. However, this isn't a crippling objection as it stands. For Ayer *could* reply that the proof of the principle is to be found in its application (see Foster 1985, 31, for further discussion of this possibility). He could say that we will see it is true simply because the results we get by applying it can independently be seen to be true. Now this would be a legitimate response, were the results it gave intuitively true. But there is little doubt that they are not; it is far from obviously true that statements about material objects are really about sense-contents, or that ethical statements are not genuine propositions. No-one is going to accept these claims without a strong independent argument to suppose that things must be that way. Now he would have a strong argument to this effect if he could independently demonstrate the verification principle. But as we have seen, he doesn't do this.

A further, related objection to the verification principle is that not only does it lack justification, but it is actually incoherent. The principle says, as we saw, that *all meaningful statements are either empirically verifiable, or analytic*. Now this principle applies to all statements. But it takes the form of a statement itself. Thus, it should apply to itself. That is, it should itself be either empirically verifiable or analytic. The difficulty is that the statement does not appear to fall into either of these categories.

On the one hand, it does not appear to be empirically verifiable. In order to be empirically verifiable, there would have to be experiences we could have that gave evidence that all meaningful statements are empirically verifiable or analytic. But it is not clear what experiences would do this. One problem here is that you don't *experience* whether a statement is meaningful or not, so it's difficult to see how experiences could help you decide which statements are meaningful. On the other hand, the principle does not seem to be analytic either. The difficulty is that analytic truths are supposed to be true in virtue of the meanings of the words involved. And this suggests that if you know the meanings of the words involved then you should be able to tell straight off whether the statement is true. For example, if you know what a ewe is (i.e. a female sheep) then you can tell straight off that ewes are female. But this doesn't seem true of the principle of verification. We all, presumably, understand the words involved in stating it ('meaningful', 'experiences', and so on). But few would claim that they could see straight off that it is true.

There are two major problems, then, with Ayer's principle of verification. The first is that he doesn't provide us with any reason to think it is true. The second is that the statement is incoherent, as it declares itself meaningless. It is, in other words, a prime bit of metaphysics. If these charges are correct, then Ayer's project is clearly deeply flawed.

V Defending Ayer – The Consistency of the Verification Principle

Let us look at whether these criticisms will stick. I will take them in reverse order.

The second criticism was that the principle of verification classes itself as meaningless, so cannot consistently be held. In fact Ayer addresses this problem briefly in the introduction to the second edition of LTL. He says,

> While I wish the principle of verification itself to be regarded, not as an empirical hypothesis, but as a definition, it is not supposed to be entirely arbitrary. It is indeed open to anyone to adopt a different criterion of meaning and so to produce an alternative definition which may very well correspond to one of the ways in which the word 'meaning' is commonly used… Nevertheless I think that, unless it satisfied the principle of verification, [a statement] would not be capable of being understood in the sense in which either scientific hypotheses or common-sense statements are habitually understood. (LTL, 16)

There seem to be two strands to what Ayer is saying here. One strand is that the principle of verification is a 'definition', and that it is open to others to adopt a different definition. On this view, it seems that the principle is supposed to be a *recommendation* about how to talk about meaning, rather than a statement about what meaning is. On the other hand, Ayer suggests that the criterion is not entirely arbitrary, and corresponds to the way in which certain statements are 'habitually understood'. On this view, it seems that Ayer is saying that the principle of verification should be seen as an *analytic* truth, that is, a statement true just in virtue of the meanings of the words involved (such as 'meaning'). Let us ask whether either of these responses could work for Ayer.

The first view is that the principle of verification is primarily a *recommendation* about the way we *should* talk about meaning. This view certainly has some advantages. For example, suppose someone objects to Ayer that the conclusions Ayer draws from it about material objects, ethics, and the rest, are not intuitively true. On this view he can merely respond that he is not describing how we *do* understand statements about material objects and ethics, but how we *should* talk about them, so it's irrelevant if his conclusions are not intuitively true – he is not describing them but proposing how they should be reformed.

But this leaves him with two problems. One is simply that it is very difficult to justify the claim that we should only regard statements that meet the principle of verification as meaningful. Presumably a metaphysician, for example, will disagree, and will think that Ayer's proposal is a very bad way to talk about meaning. It is difficult to see how Ayer could respond to this. Another difficulty is that Ayer's theory about ethical statements says that a statement like 'you should do X' just means something like 'X – Hooray!'. If this is so, then the statement 'you *should* only regard statements as meaningful when they meet the principle of verification' just means 'the principle of verification – Hooray!'. But if this is true, we would have to interpret the whole book as being devoid of arguments, and merely expressing Ayer's approval of the verification principle and his disapproval of metaphysics. This, I take it, is not how he means it.

The other suggestion is that Ayer should say that the principle of verification is an analytic truth. I agree with Foster (Foster 1985, 5-6) that this is what Ayer should say. The main advantage of this is that it fits in with his conception of philosophy. Ayer argued, remember, that philosophy is all about *analysis*, where analysis involves finding connections between the meanings of different words that we use. Thus, if the principle of verification is a bit of philosophy – which it presumably is – the natural thing to say is that *it* can be reached by a bit of analysis of the words involved: 'meaning', 'verification', and so on.

The problem that we raised above with this idea was that unlike 'ewes are female', the principle of verification is not obviously true even if you understand the words. But as Foster points out (1985, 5) this is not really a major problem. If Ayer is right, then there are many unobvious analytic truths. For example, philosophical claims that material objects are really about sense-contents, or that ethical statements do not express propositions, are on Ayer's view analytic but not obvious. So are unobvious mathematical propositions like 'every even number is the sum of two primes' (assuming this is true). In all these cases, we would need to think hard before we realised that the propositions in question are true, even if their truth was, ultimately, just a matter of the meanings of the words. So there is no reason to think the principle of verification should be any different.

I think it is *consistent*, then, for Ayer to regard the principle of verification as analytically true. But this doesn't mean that he has any *justification* for holding it. This brings us to our second point. *Is* there any reason to think that it is actually true?

VI Is there any Justification for the Verification Principle?

I think it is fair, in view of what we said in section III, to say that Ayer himself never really gives us a good reason to accept the principle of verification. However, I think there is a plausible line of argument that leads to *something like* the principle. Whether it gives us exactly Ayer's version of the principle is a more subtle issue, however, as we will see.

Let's begin by asking what reason there is for thinking that verification has anything to do with meaning at all. On the face of it the two notions are not intimately connected. A sentence means something if it succeeds in representing some state of affairs as obtaining. But it is verifiable if we have some way of recognising that the state of affairs it represents obtains. These appear at first sight to be two entirely different things.

However, there are considerations that suggest that meaning and verification are more closely linked. Here is one sort of argument that leads to this conclusion. I will put the argument in terms of 'linguistic units' for reasons I will explain shortly – by 'linguistic unit' I mean either a word or a sentence. We will see that the argument runs a bit differently depending on what exactly we take a linguistic unit to be.

The argument I have in mind has three premises. The first is that for a linguistic unit to be meaningful it must be possible for someone to understand it. This, I think, is very plausible, since it is ultimately through people using and understanding linguistic devices like words and sentences that they acquire a meaning.

The second premise is that someone can only be said to understand a linguistic unit if they are able to distinguish between cases in which it applies and cases in which it does not. This, again, I think is quite plausible. For example, consider a sentence like 'it's raining'. It is plausible to think that someone would only count as understanding it if they could (at least sometimes) distinguish between cases when it was raining and cases when it was not. Or in the case of a word, someone could only count as understanding the term 'dog' if they could distinguish between cases where a dog was present and cases where no dog was present.

The third premise is that in order to be able to distinguish between cases where a linguistic unit applies and cases where it does not, it must be possible to have experiential evidence for when it applies and when it does not. This, again, is pretty plausible. There is no way of distinguishing between when it's raining and when it isn't unless there would be some experiential evidence (e.g. feeling raindrops) by which we can detect the difference. Equally, there is no way of distinguishing between when a dog is present from when one isn't unless there can be some experiences (e.g. the sound of barking) by which we can tell the difference.

In summary, then, this argument suggests a link between a linguistic unit's having meaning, and our being able to verify that it applies. It does this because we can only be said to understand a linguistic unit if we can distinguish when it applies from when it doesn't, and to do this, there must be some evidence by which we can do this. If the argument is sound then *something* like the verification principle follows, as we can conclude that a linguistic unit can only be meaningful if it is it is possible to have experiential evidence for when it applies and when it does not. The reason I put things in terms of 'linguistic units', however, is that the argument yields very different conclusions depending on what we take linguistic units to be.

VIITwo Versions of the Principle of Verification

If we think that linguistic units are *sentences*, then the argument demonstrates that a *sentence* is only meaningful if there can be evidence for when it is true and when it is not. This is exactly Ayer's verification principle. But if we think that linguistic units are *words* then things are rather different.

To see this suppose that I am competent at recognising dogs, and therefore at recognising whether the word 'dog' applies to anything in my vicinity. If this is true the word-version of the verification principle says that I can understand the word 'dog'. But then I can go on to formulate such sentences as, 'there are dogs so distant from us that no-one has, or ever will, have any evidence for them'. Now if we assume that the other words in this sentence are meaningful, and that a grammatical sentence comprised from meaningful words is itself meaningful, then it follows that the sentence is itself meaningful. But it is unverifiable – there could be no evidence for the existence of dogs for whose existence there is no evidence! It follows that if we think the verification principle applies to words rather than sentences, Ayer's version of the principle does not follow, since there can be unverifiable but meaningful sentences.

Let us call the view that a sentence has to be verifiable to be meaningful the *sentence-principle*, and the view that a word has to have some verifiable conditions of application the *word-principle*. In order to work out whether there is any argument in what we have said for Ayer's view, we need to look at whether the word-principle or the sentence-principle is more likely to be true. There is not room here to consider the issue in detail. However, my own view is that the word-principle is much more likely to be true. I'll suggest a couple of reasons why I think this is.

One reason is simply that I think we have a very strong intuition that we *can* form unverifiable, but meaningful, sentences. One example is, 'there are dogs that no-one will ever discover', which certainly *seems* meaningful, despite being unverifiable. Another example might be the case of scientific theories that postulate entities for which no evidence is known. For example physicists currently talk about fundamental physical forces being explained in terms of 'superstrings'. But (as I understand it) no evidence has yet been proposed by which a superstring could be detected. It seems conceivable, then, there just *is* no such evidence. However, we would be reluctant to say that all talk about superstrings is, as a result, meaningless. It seems, then, that if you believe the sentence-principle, then you need some explanation of why sentences like this *seem* meaningful, but aren't – and it's not too clear how such an explanation would go.

On the other hand I don't think there are such strong intuitions that there are meaningful words that would violate the word-principle. Or rather, not after we adjust it slightly. We need to adjust it to take into account the possibility of defining words that do not satisfy the word-principle. For example suppose I defined the term 'udog' to apply to all the dogs for whose existence we will never have evidence. Talk of udogs would then be meaningful – since I gave a perfectly clear definition of what I meant by it – but we could never detect the presence of udogs. Thus, the principle would need to allow not only words that satisfy the principle themselves, but words that are definable in words that satisfy the principle. The same goes for 'superstring' – which physicists could presumably define in terms of other more recognisable concepts if called upon to do so. But given this adjustment to the word-principle there seems to be a good case for saying that any words that fail it are genuinely meaningless.

Another reason we might prefer the word version of the principle is that it gives a better explanation of the meaning of analytic sentences. Remember that the principle of verification says sentences are meaningful if they are verifiable *or* analytic. So we might ask how the different versions of the principle of verification we have been looking at might classify analytic sentences.

On the one hand, the word-principle appears to apply directly to analytic sentences. If we ignore sentences that are themselves definitions, it is plausibly true of *all* sentences – analytic or synthetic – that they are meaningful only if their constituent words have detectable conditions of application or are definable in terms that do.[1] On the other hand, the sentence-principle cannot be applied directly to analytic sentences, since there are no experiences that count as evidence for or against the truth of analytic sentences. This suggests that the sentence-principle is going to have to be combined with some other principle that explains the meaning of analytic sentences.

The problem I see for the sentence-principle might therefore be put this way. The natural account of what makes analytic sentences meaningful is that the words that form them are meaningful and grammatically arranged. But the sentence 'there are dogs that no-one will ever discover' is also apparently comprised from meaningful words grammatically arranged. So the believer in the sentence-principle needs some explanation of why this idea works for analytic sentences, but not all non-analytic sentences. I am not sure that no such explanation is available. But it isn't as clear as in the case of the word-principle what that explanation would be. I think, therefore, that there at least are some reasons to think that if any version of the verification principle is true at all, then it is going to be the word principle.

VIII The Verification Principle and the Elimination of Metaphysics

I have argued that although there are considerations that point to something like the verification principle, they are more likely to lead to the principle that words have to have verifiable conditions of application than that sentences have to be verifiable. I now want briefly to comment on how much of Ayer's project would survive if this version of the principle turned out to be true.

The fundamental aim of Ayer's book, we will remember, was to eliminate 'metaphysics' – for Ayer, the attempt to gain knowledge of matters of fact through pure thought – by showing that it is meaningless. But I do not think that *this* project could succeed if only the word-principle is true. According to the weak version, it is only necessary for a word to be meaningful that there be some cases in which it is possible to gain evidence of the fact that it applies. Thus it allows unverifiable but meaningful factual sentences – so by Ayer's definition, it will allow at least some 'metaphysics' to be meaningful.

Not only this, it is likely to leave just the sorts of metaphysical statements that Ayer dislikes most to count as meaningful. For example, consider the claim that the world is completely independent of and different from our experience of it. This is unverifiable, but the words we need to express it – 'independent of', 'different from', and so on – are clearly of a sort that would occur in other more familiar contexts, where we could recognise their conditions of application: two cogwheels can be recognisably moving independently of one another, say, or any two things could be recognisably different from one another. So we can formulate this metaphysical claim using words that the word-principle counts as meaningful. The same, I think, would go for many other traditional metaphysical statements.

This doesn't mean that the word-principle is completely devoid of anti-metaphysical implications, however. This is because there remain words which we sometimes class as meaningful, but where it is difficult to think of *any* situations in which their conditions of application could be recognised in this way, or any definition of the word in such terms. One kind of case would be when a notion was simply too ill-defined and abstract to have any detectable conditions of application or precise definition. If Ayer is right, something like 'The Absolute' might be like this (though not being familiar with the work of C. D. Broad, from which he takes this example, I would not wish to endorse this claim!)

Ethical claims might also be classed as meaningless if we adopt the word-principle, though the issue is less clear. On the one hand, anyone who understands moral terms must be able to distinguish cases when they apply from cases when they do not – you don't understand 'is morally wrong' unless you have some ability to distinguish morally wrong actions from others. But on the other hand, when someone distinguishes cases where a moral term applies from when it does not, they will do this on the basis of non-moral facts – e.g. they will judge that an act is wrong since, for example, it causes pain to an innocent person. So it is not clear whether a moral term like 'is good' will satisfy the word-criterion, since it isn't clear whether we can really have experiential evidence that this term applies in the right kind of way.

Thus, I think one could base *some* sort of anti-metaphysical project on the word-principle, by sifting out concepts that neither have detectable conditions of application, nor are definable in terms that do. But in view of the fact that the word-principle allows us to formulate apparently metaphysical statements, that project would be little like Ayer's.

IX Conclusion

In this paper I have tried to give a sympathetic reading of *Language, Truth and Logic*, and show that Ayer can be defended against some of the criticisms that are sometimes levelled at his view. Moreover, I have suggested that there are reasons for thinking that something *like* the overall project of the book is well-motivated. However, I have also tried to show that the anti-metaphysical aspect of Ayer's project relies on a premise – the principle of verification, as applied to sentences – which he does not adequately justify, and which is problematic for independent reasons.

Ultimately, I think the problem is that he never really overcomes the strong intuition that at least some unverifiable sentences *are* meaningful. Nevertheless, I think that the book retains considerable interest as an exploration of what we would have to accept if the principle of verification were true.

King's College London.

References

[1] This criterion would not work for sentences that *are* definitions because definitions contain words – the words they are supposed to define – which are not independently meaningful. For example, suppose we try to define 'a flotch' to mean 'a brown horse', by stipulating that the sentence 'a flotch is a brown horse' is true. But is this sentence meaningful? According to our criterion it is meaningful if and only if the words are meaningful, and the words are meaningful if and only if they have detectable conditions of application or they are definable in terms that do. So is 'flotch' so definable? Presumably only if 'a flotch is a brown horse' is a meaningful sentence… but it should be clear that we have now gone round in a circle. This means we need a different criterion for when definitions are meaningful. There isn't really room here to go into how we might formulate such a criterion, but I don't think it will ultimately be a problem for this theory.

A. J. Ayer, *Language, Truth and Logic*, 2nd ed., (London: Victor Gollancz Ltd., 1970). Referred to in the text as 'LTL'.

Foster, John, *Ayer* (London: Routledge and Kegan Paul, 1985)

Notes On [Contributors]

Pierre Cruse

is a teaching fellow at King's College London and has held a research fellowship at the Catholic University in Louvain la Neuve. He completed his undergraduate studies in philosophy and physics at the University of Glasgow before going on to do his doctoral work at King's College London. He has jointly edited a collection on the emotions and has published on scientific realism, which is also the subject of his current research.

D.J. Sheppard

teaches philosophy and politics at Kenilworth School, Warwickshire. He has previously taught at the University of Warwick and Staffordshire University. He is currently writing a book on Plato's *Republic*.

James Hill

lectures at Charles University, Prague. He studied at Oxford, Geneva and King's College London, and has published a range of articles on the empiricists. At the moment he is preparing a book on John Locke's critique of mechanism and working on Descartes' concept of thinking. He is also collaborating on the translation of several English-language philosophers into Czech.

Peter Simons

is professor of philosophy and head of department at the University of Leeds, and honorary professor of philosophy at the University of Salzburg. His wide range of interests includes metaphysics, logic, philosophy of mathematics and the history of Central European philosophy. He has published widely, including *Parts: A Study in Ontology* (Oxford: Clarendon press, 1987), 'Truth-Makers' (with Kevin Mulligan and Barry Smith) in *Philosophy and Phenomenological Research, 44*, and 'Bolzano, Tarski and the Limits of Logic', in *Philosophia Naturalis, 24*. He is also the director of the Franz Brentano Foundation, and edited *History and Philosophy of Logic* from 1993 to 2001.

Paul Sheehy

is joint editor of the Richmond Journal of Philosophy.

Garrath Williams

is lecturer in philosophy at the University of Lancaster. He completed both his undergraduate and graduate studies at the University of Manchester, and has taught at the universities of Central Lancashire, Manchester, and St Andrews, as well as the European Academy in Bad Neuerahr-Ahrweiler. He specialises in moral philosophy, political thought and applied ethics. His numerous published articles include 'Nietzsche's Response to Kant's Morality' in *Philosophical Forum 30*, 'Normatively Demanding Creatures: Hobbes the Fall and Individual Responsibility' in *Res Publica 8*, and 'Blame and Responsibility' in *Ethical Theory and Moral Practice 6*.

Notes For [Contributors]

Content

We welcome articles on any area in philosophy. Papers may be broad or narrow in their focus (for instance a discussion of the mind/body problem, or an analysis of Hume's treatment of causation in the *Enquiry*). We would particularly encourage contributions which reflect original research on the following philosophical themes: epistemology, metaphysics, philosophy of religion, ethics, philosophy of mind, philosophy of science, political philosophy, religious ethics; and texts, such as: *The Republic, The Nicomachean Ethics, The Meditations, An Enquiry Concerning Human Understanding, Beyond Good and Evil, On Liberty, Existentialism and Humanism, The Problems of Philosophy, Language Truth and Logic*.

The articles should be around 3000-4000 words.

Style

The language used in the articles should be as non-technical as possible whilst preserving the richness of the arguments. Where technical terms are unavoidable they should be explained and examples offered.

Notes should be presented as endnotes. Textual references should be presented in the following format: Barry Stroud, *Hume* (London: Routledge, 1977), 77-91.

Presentation

Articles should be written in *Word* (any version).

Contributions

Articles for this journal are currently written by a panel of philosophers from a variety of universities in Britain, Australia and the United States, whose work is edited by the journal's editorial board. We invite unsolicited contributions from philosophers working in any field. The contributions should be submitted via email attachment to rjp@rutc.ac.uk

Copyright

The RJP retains the option of reprinting published articles in later RJP publications. Authors may republish articles with the journal's permission provided that they acknowledge that those articles were first printed in the RJP. Papers should only be submitted if the author is willing and able to be bound by the conditions set out in this paragraph.

[Contributors]

[Subscribing] to the *RJP*

The RJP comes out in Autumn, Spring and Summer. To subscribe you need to select the appropriate price from the table and complete the mailing information at the bottom of the form. All prices include post and packaging. The bottom half of the form should then be detached and sent with a cheque made payable to Richmond upon Thames College to the address below:

RJP Subscriptions
Philosophy Department
Richmond upon Thames College
Egerton Road
Twickenham
London TW2 7SJ
United Kingdom

Please allow one week for delivery in the UK, and two weeks for the rest of the world.

Annual Subscription : Current Rates

Institutional Subscriber in the UK	£33.00
Individual Subscriber in the UK	£18.00
Institutional Subscriber from the rest of the EU	63.00 Euros
Individual Subscriber from the rest of the EU	40.50 Euros
Institutional Subscriber outside the EU	$67.50 US
Individual Subscriber outside the EU	$45.00 US

Resubscription YES / NO From issue ....

Your name: .... Your organisation's name (if appropriate) ....

Address ....

.... Post code/Zip code..

Telephone number (including full international dialling code) ....

Email address ....

I enclose a cheque for the amount of .... to purchase an annual subscription

Signed .... Date ....

[Philosophy]

The Richmond Journal of **Philosophy**

Richmond Journal of Philosophy
Philosophy Department
Richmond upon Thames College
Egerton Road
Twickenham
Middlesex
TW2 7SJ
United Kingdom

Tel: 020 8607 8000
Fax: 020 8744 9738
Email: rjp@rutc.ac.uk
www.rutc.ac.uk/rjp

Richmond upon Thames College

ISSN: 1477-6480

Design & Production Marketing for Education 01282 612222 [ref 26809 - 04/04]