1(2) - March 2006 ISSN 1748-9393 (Online)

British Journal of Undergraduate Philosophy

Editor: Robert Charleston
Open University

Journal of the British Undergraduate Philosophy Society

British Journal of Undergraduate Philosophy

Journal of the British Undergraduate Philosophy Society

Editor: Robert Charleston, Open University
rc3673@student.open.ac.uk

bjup@bups.org
www.bups.org/BJUP

ISSN 1748-9393

A big thank you to:

Edward Grefenstette, Sheffield University
Andrew Stephenson, Cardiff University

for their late-night copyediting skills.

Contents

British Journal of Undergraduate Philosophy, 1(2), March 2006

Babies and bathwater

Editorial

At the time of writing, education and the principles behind the British education system are attracting a considerable amount of political and media attention. Once again, 'selection' is the buzzword right at the heart of the debate. Philosophically, it seems unlikely that 'selection' is what really worries people. After all, it is difficult to imagine that many people would object to a system if it always selected them, the people they love, and the people they rate. What worries us is that someone we care about or rate – whether ourselves, a loved one or a stranger – might *not* be selected. That they might be *rejected*.

Whether you philosophically support or oppose selection/rejection and the criteria currently being proposed in UK government plans, one thing is clear: university-level philosophy is riddled with opportunities for rejection. You face one opportunity before you arrive, of course, with your UCAS application. But this is really just the beginning. Only a few essays are selected by tutors and examiners to receive the highest marks. You may see a few people in your department rejected outright at the end of your first year. If you make it to your final year, you will face another grading/rejection exercise – a co-operation between your department and an external examiner. If you want to continue further in philosophy, you open yourself up to rejection by a whole range of new institutions: other universities for a postgraduate place, the funding bodies, local scholarship and studentship boards, conferences, journal reviewers and so on. Professional philosophers have survived many levels of rejection by the time they get their first post, and then have to submit papers to journals which – like the BJUP – may only accept about 20% of submitted pieces. Nor does this stop with seniority. The York philosopher Professor Tom Baldwin recently revealed that despite being the new editor of *Mind*, one of the most prestigious professional journals, he *still* gets rejected by other publications on occasion. This is depressing (it never stops!) or

encouraging (the anonymised submission process *does* work!) depending on how you look at it.

All of this raises the question: if rejection is endemic in academic philosophy, why isn't 'dealing with rejection constructively' taught as a key skill? Why the lack of advice, the humming and haaing, the 'try harder next time, bad luck' vagueness? Partly, it seems likely there's a feeling that if you were rejected, you probably deserved it. Either you aren't good enough, or you didn't try hard enough. Second, the very human reluctance to associate yourself with failure – implicit in the statement 'Oh, don't worry – I know how to deal with rejection…' – can only be amplified in such a competitive, combative subject as philosophy. Rejection hurts, there are some philosophers you have to work with who never seem to be rejected, and nobody wants to lose prestige in the eyes of their colleagues. So it seems better to just avoid the topic of rejection, never really focus on it, just hope it doesn't happen.

All of which is wholly inadequate. If you've *never* been rejected in any of the ways described above – well, yah, boo, sucks. I've got great grades, several publications, good references, a couple of prizes and a scholarship. *I've* been rejected. I sometimes write a really duff essay. I've even – in this issue – rejected one of my own reviews for being too boring. This – really – is quite normal. I'd like to suggest: i) that there is a risk of fragility in philosophers who have not experienced, and survived, rejection; and ii) that there's a philosophy of rejection itself.

So what should your reaction be to rejection? What does it mean? Let's quickly consider the most salient case at this time of year. What should your reaction be if you fail to gain a postgraduate place at your university of first choice, with funding?[1] This is a paradigm case of rejection. It's the sort of thing that looks like it could change your life, or at least your career. People get very worried about it, and upset if it goes wrong. Does such a rejection

[1] Note that I'm omitting 'What if I fail to get a postgraduate place at all?' This is not a problem in the UK – if you have an undergraduate philosophy honours degree, 2:2 or above, and can pay your way, there are universities that will *always* take you for a masters degree. If you doubt this, and need one, email me and I'll list a couple.

mean you are not good enough for what you are trying to do? That you need to change your work? Or stop?

Well, *maybe*. There are no restrictions on who can download the forms to apply to universities or the AHRC for funding. Anyone can do so, so it *could* be the case that you have been completely unrealistic in doing so. You *may* simply be reaping the experts' opinion against your over-optimistic application. The AHRC and universities are all, by definition, experts in their fields. They *are* their fields. If they sent you a letter saying 'Your grasp of normative ethics borders on the infantile, you have your technical definitions confused, and your handwriting is atrocious,' you would *have* to take their judgement very seriously.

But – first – the relation between you and the people you apply to is *not* that simple. To apply you must have referees who know you and your work. If they were prepared to put in the time to write a reference, it was clearly not a *hopeless* idea to apply. And – second – if you have been rejected, you will *not* have been told that you're useless in the way I've described above. Most of the time – and this is *criminal* – you won't have been told anything at all, other than that you've been rejected.

The AHRC has a twenty-eight page application form, with fifty-six pages of explanatory notes, for each of its two funding competitions. The application procedure takes six months *after* you've submitted all the material. Oxford and Cambridge have extensive application forms, require two 2,500 or 5,000 word essays, full, certified transcripts, and two or three full references each, all in duplicate or triplicate. Other top-rated departments' application procedures can be just as involved. You must invest a couple of months' work if you want to really go for these things, will need to have your application together by the Christmas before you want to go, and will need a fairly detailed idea of what you want to write your final dissertation on at the *end* of your masters, *nine months* before you even start the course. But you will *not* receive detailed feedback if you are rejected. Some institutions will not even tell you if you have been rejected – they will simply say 'If you haven't heard from us in two months, you did not get in.' You will – in all probability – not know why you were rejected, nor will you be given an idea of what you could do better next

time. See? Criminal. But it *does* deny the idea that you are being told you are useless. You are not being told anything at all, other than that you were not one of the selected.

There are many, many applicants for places at top postgraduate departments, and for AHRC funding. Success is relative to the quality of the other applicants, not absolute, and you will be unable to second-guess the selection boards' priorities. If you are a pretty-good aesthetician applying in a year of Wittgensteins, to a board with a preference for philosophy of mind or science, you may be rejected when another year would see you accepted. There are problems with the system. Many people think a UCAS-style answer needs to be put in place.

Being rejected in these circumstances is difficult to learn from. Do *not* assume it was your fault – the system may well have made the wrong call. Really, really try to get a friendly tutor (at your current department, or the postgrad department you join) to go through your application and give you some detailed advice and feedback. And overall, don't take it too seriously – over 50% of my favourite professional philosophers have told me they were turned down for funding or a place they wanted. If it's philosophy you are interested in (rather than getting a specific university's name on your CV), you will do well at *any* of the UK's departments if you choose one that suits you and your work. It is counterproductive, and missing the point, to get hung up on the Philosophical Gourmet report (or other league tables) as you will be pushing back what most postgrads I have spoken to say is the *most* important aspect of department choice: whether you will work *well* within that department.

It also has to be said that if you are serious about philosophy or an academic career, department choice will be of fairly transient interest. The real business of academic philosophy is lecturing, teaching, writing and publishing. The former two will depend largely on you and how you interview, not your certificates. The latter two depend on an anonymised review cycle, so *cannot* be fast-tracked through simply on the basis of your (now former) university's reputation.

So do the lessons learnt from the above generalise for rejection in academic philosophy as a whole? I think they do. Not getting the grades you want in your essays? Well, there is a great repository of expertise in your department. Always pay attention to tutors' comments in the first instance. But there are always contextual and systemic reasons why you may not get the grades you want. There may still be something very important in your work, that is being missed due to departmental or tutor's preferences on style or content. No human system is perfect. But you can check whether this is the case to a certain extent by submitting carefully-written papers to conferences and journals. BUPS and the BJUP, for instance, draw reviewers from across many good departments. If you still get rejected, there is a chance we are *also* missing something when we read your work. But it looks less likely; and it certainly seems – at least intuitively – that learning to make sure people *see* the genius in your work is something a keen philosopher ought to be able to do, if you turn your mind to it. Maybe it is time to rethink, to think philosophically about *how* to write to best effect, as well as *what* to put into your essays.

But what if such efforts do not work? What if the worst comes to the worst, and you fail your course? Are you any less of a philosopher if you cannot do well in the academic system? I don't think so. Just as department is less important than publication in *professional* philosophy, so ideas and engagement *must* be more important than *papers* in philosophy itself. You don't need to be in a university to do philosophy. You don't even need to write. How much patience do you think Socrates would have had with our current system? Go and see *Pi* or *Memento* or a Goya exhibition and tell me philosophy has to be formally written and academic. Even conversation, friendship, parenthood are avenues for analysis, creation, application, refinement – the key philosophical skills. In the UK, we also have institutions such as the BBC, the OU, BUPS and the WEA if you want the formal content as *well* as a freer, better-remunerated non-academic life. Academic philosophy is *not* philosophy itself, and its selectors are *not* the ultimate arbiters of what will be rejected or retained by the world at large. Quite the contrary, if history is any guide. So rejection is not a reason to feel fear, depression or shame. It is an opportunity to learn, to improve, or to change direction towards an avenue of expression that suits your talents more closely.

Just remember: even the most famous living academic philosophers are known to less than 2% of the population at large.

I hope you are always selected, or only very rarely rejected, and stay within academia, building a successful career. But if things do get difficult academically, or you get fed up with the lifestyle, you *can* still incorporate great philosophy in your life, thinking carefully, trying to speak clearly about even rejection *itself*. This must surely be the primary response, activity and duty of philosophy. If you follow it, it doesn't matter what babies the academic infrastructure decides to raise – you and your ideas can never be just bathwater.

British Journal of Undergraduate Philosophy, 1(2), March 2006

Vanity and virtue

Milen Ganev
University of Bristol
mg3618@bristol.ac.uk

Do not be disheartened by the dullness of our title – this thing I bring before you is neither an angry tirade nor a long-winded sermon. It is a strange little item, written with intentional candour, and hopefully it will bring you some pleasure.

I begin with a story which takes place some six-hundred years ago, during a time of chivalry and honour, knightly virtues and courtly love. The hero of our tale was the son of a very wealthy and influential governor, who oversaw the wellbeing of a prosperous city in a distant kingdom whose name we need not mention. Our hero lived a life of comfort and complacency, surrounded by the affluence which his father's position guaranteed – a position which he was eventually to assume himself, or at least that was the intention. However, there came a time when the blind arrow of desire pierced through the arrogant heart of this young man – and he fell deeply and hopelessly in love.

The lady of his heart was so exceptionally beautiful that she was an object of universal and incontestable admiration. Our hero marvelled at her peerless beauty from afar, never having spoken to her, and was thrown into a state of anxious deliberation. He asked himself the following: 'what if I were to lead my life as if she were looking down upon every second of it – every desire, every act and every consequence? If I had the strength to live in such a way, would this not be the paragon of virtue? For how could she esteem something in me that was reproachable? How could she be impressed by something contemptible?'

By virtue of these thoughts and others like them, and after much consideration, the young man decided to leave behind the opulence of his situation and the security which it ensured him, and to become a wandering

knight errant. He would venture across the lands helping orphans and defending widows, liberating the oppressed and educating the ignorant, overcoming every danger and laughing in the face of fear; in short, doing all those things which he guessed that his beloved lady would find admirable; proving his worthiness by the most thorough and exhaustive means. And this is just what he did – although in actuality, as is often the case, the extent of his courage and valour fell short of the greatness he desired.

From this story we take away with us the following question: 'if a man lives as if someone he loves is forever watching over him, then how can his acts be anything but thoroughly virtuous?' Bearing this in mind, let us consider two well-known moral maxims:

1. Treat others as you would wish to be treated yourself[1]
2. Act only according to that maxim whereby you can at the same time will that it should become a universal law[2]

1, in its deepest signification, tells us to treat others in a way which we think *they* would find desirable. A husband who is in moral doubt as to the manner in which he is treating his wife must ask himself: 'How would I feel in her position?' rather than 'How would I feel if she treated me in the same way?'[3]

2 is the first formulation of Kant's Categorical Imperative. According to this maxim, the question that we should ask ourselves in deciding how to act is: 'What if all rational beings were to act this way?' If any particular act, when considered as a universal law, would cause a logical contradiction or else bring the world into a state of despair or disharmony, then Kant says we have a duty as rational beings to abstain from such an act. One should only act in a manner one would desire all rational beings to follow.

[1] See Matthew 7:12, or, if you like, see chapter II of *Rob Roy*, where Sir Walter Scott calls this "the fundamental principle of all moral accounting" – and I'm sure there are a hundred other places where you could find this golden rule.

[2] Immanuel Kant – *Groundwork of the Metaphysics of Morals, Section II*

[3] We could clarify our maxim as: 'Treat others as you would wish to be treated yourself, *if you were in their shoes*' – yet it matters little for my purposes in this paper, whether the reader agrees to this or stands by the more straightforward interpretation – both alternatives are equally susceptible to the criticism I will offer shortly.

Before challenging both our maxims I will give a definition of a concept central to this paper: that of vanity. Hume speaks of vanity as the love of the "fame of laudable actions"[4]. I define the term very simply as: 'the desire for admiration'. Not as conceit or shallowness, but just the desire to be admirable in the eyes of others. Now here is my first proposition: *no human being is without some degree of vanity.*

If this is accepted, then the problem with maxim 1 comes to light almost immediately. One would wish to be regarded with admiration, and presumably one would wish to be regarded in the same manner by everyone – not with disdain from one person, envy from another and esteem from a third. So maxim 1 prescribes that the individual treats everyone else with unrestrained admiration, not distinguishing between the corrupt and the virtuous, the murderer and the martyr. For, of course, the criminal too has vanity. To say that he 'desires reproach or punishment' is to stray far from the truth. By maxim 1 we are allowed no preference from one person to another; an entire half of the emotional spectrum is denied to us: jealousy, hostility, contempt – all these sentiments must be renounced, and with them, our capacity for any sort of condemnation, be it moral or otherwise. It is both impossible to live in this way, and absurd 'to prescribe living in this way'.

The problem here is twofold. Firstly, there is a disparity between asserting that everyone should receive the treatment they desire, and at the same time reserving the capacity to call persons or actions into reproach when it is clear that this reproach or punishment is undesirable to the subject. Secondly, we have the disparity between the boundless vanity of each and every person and the feeble quantities of admiration that they are willing or able to extend towards others. Of course it could be argued that without possessing the ability to *choose* what we admire, we cannot be expected to conform to maxim 1 in this respect: namely to hold others in as much esteem as they desire. However this does not overcome the tremendous discord between the sentimental nature of the human heart and the reality of the treatment that it can possibly receive. Anyone who insists on prescribing that we 'treat others as we would wish to be treated ourselves' – is overlooking the undeniable

[4] David Hume, *Of the Dignity or Meanness of Human Nature* [An essay from: *Essays: Moral, Political and Literary* ed. Eugene F Miller (Indianapolis: Liberty, Fund, 1985), pp. 80-86]

impossibility of such a state of affairs. Furthermore, the feeling of admiration can be nurtured over time, and the feeling of contempt can certainly be hidden or otherwise suppressed. So we again return to a situation where the maxim asks us to treat others in accordance with their vanity. This request is unreasonable and unjustified in the limited realm of its possible application; and fundamentally undermined by the general sense of discord between our endless vanity and the underwhelming reality of the admiration this world offers us. In light of all this, one must either give up the maxim or else reject my first proposition – the vanity of the human heart. Anything else is problematic, to say the least.

Let us move on to maxim 2 which – although it appears different from 1 – actually gives a very similar code of practical morality. Here I will offer my second proposition: *in some cases, there is nothing reproachable or immoral in someone acting against a law they approve of.*

To defend this, let me return to the theme of love and to the wondrous landscape of our 15th century kingdom. My example involves a young lady, her would-be lover and her protective father, who – being a cautious man and moreover a Baron reputed for his honour and honesty – complies with the traditions of the age in acknowledging the laws of chastity and in safeguarding his daughter's virtue. The young nobleman who has become enamoured with this beauteous maiden is continually attempting to win her favour and, as it were, reap the fruits of his desire. She would happily go along with this – but the vigilant father is forever standing in their way.

I maintain that the two men could see completely eye-to-eye, they could have identical moral outlooks, and at the same time the lover would continue his attempts to 'actualise' his desire and the father would continue to guard against this. There is nothing reproachable in either man's actions, nothing hypocritical or insincere about the young man admiring the father's protectiveness, but at the same time attempting to undermine it.[5]

[5] Or, of course, the converse: the father admiring the young man's ardent efforts and at the same time guarding against them.

I further maintain that an impartial observer could simultaneously sympathise with both the father and the young man. For although there is a conflict of interests here, there may not be a moral conflict at all. The young nobleman may wholeheartedly agree with the values of honour and chastity which the father so resolutely enforces, while at the same time attempting to bypass or undermine the father's influence. There is nothing immoral here. Nor is there any contradiction is saying both: 'Fathers should protect their daughters' and 'Young men should attempt to undermine this' – again, the interests are conflicting, but the morality which grounds them may not be. To extend this example, we could replace the father with a protective brother who has fallen in love with a woman from the neighbouring province of our distant kingdom. The brother possesses similar youth, virtue and nobility to his beautiful sister. He can do everything within his power to attract or impress the woman that his heart desires – while at the same time guarding his sister from such efforts, however honourable they may appear to be, and reproaching those who seek her favour.

His moral outlook may be perfectly consistent. There may be nothing in his actions deserving reproach, nothing is his manner that could be called insincere – and yet this apparent 'double-standard' emerges. This is because a single moral outlook can produce apparently conflicting interests, as long as they are both admirable. But I am getting ahead of myself – let me return for a moment to a hypothetical observer of the above state of affairs. I claim that he could simultaneously agree with both the father's actions, and their accordance with the principles of honour and chastity; *and* the lover's actions, and their accordance with the principles of romance and gallantry; *and at the same time*, he could formally criticise one or the other or both. By this I mean that his official, or public, or formal opinion – the one he makes known to others – could wholly conceal his genuine moral outlook on the situation. Again, there is nothing reproachable in the conduct of this observer. Condemning an act which you are in moral agreement with, or approving of an act which is against your moral outlook, is not necessarily immoral or hypocritical or reproachable. This is because your 'moral agreement' is grounded in how much you admire the act being committed – while your 'formal agreement' is grounded in how admired you think your response to the act will be. The former is a product of the observer's admiration, the latter

of his vanity. We may admire both the lover and the father – and at the same time, formally criticise one or the other or both. This is all in keeping with a consistent moral principle, which I will shortly explain.

I have tried to show that two moral outlooks can produce conflicting actions while being morally compatible – or indeed, identical. This means that, very often, when you attempt to universalise the maxim of certain actions, you produce a law which condemns people with a morality compatible or identical to your own, or even yourself. By maxim 2, such acts would be immoral – yet I claim that there are many exceptions and offer the above example as evidence of this.[6]

What then, is moral virtue? Both our maxims miss the mark, and this is not because of the selfishness or baseness, destructiveness or wickedness, of human nature – but rather because of the unyielding vanity of the human heart. In the first case, vanity prevents us from extending the same unrestrained degree of esteem to others as we would wish to receive ourselves. In the second case, vanity encourages us to break away from our own laws and conventions –

[6] In the case of the lover, the universalization of the maxim of his action would produce undesirable results whatever he chose to do – by this I mean: whether he pursued the woman or not. However many qualifications we make about the circumstances of this pursuit in an attempt to arrive at an objective definition of virtue, the universalization of the maxim would never be desirable. This is because the most desirable and beneficial state of affairs would be if *only some people* followed this course of action. The father does not want to keep his daughter isolated from men forever – but he only wants certain men, who he deems to be worthy, to approach her, and this only after they have proved themselves against his resolute vigilance. This means that even if we reinterpret Kant's maxim so that it allows two conflicting actions to both be moral, by saying that the universalized laws may be such that one goes against another, we still have the further problem of ethical choices which do not allow of satisfactory universalization at all. This problem is in fact different to the one raised by those who criticize Kant's approval of a law which says: 'everyone should tell the truth' by giving the example of a murderer asking for the location of his intended victim. The case of the murderer admits of qualification – 'everyone should tell the truth unless the following situation occurs…' – but in the case of the lover no such thing is possible. Now if you take the additional liberty of allowing qualification about the *kind* of people who are warranted in following a particular moral law – if you produce maxims along the lines of: 'this type of person in this type of situation should follow this rule' – then you have to deal with the potentially irresolvable argument about who qualifies and who does not.

because this breaking away is also rising above, and with this comes admiration.

Accusations made against human nature – produced by dejected hearts and designed to throw the human spirit into disgrace – are not only fruitless but tremendously harmful. Our nature is immutable and our 'being here' is unquestionable – the more we turn away from it, the more we damage ourselves. On that, far too serious point, let me now present my definition of moral virtue – although it will be difficult to defend in the space that remains. Here is its most general formulation:

> *A morally virtuous act is one which is simultaneously admirable in the subject's eyes, and in the eyes of those he cares for.*

This is, admittedly, somewhat ambiguous – and I should take a moment to elucidate the definition itself, before proceeding to defend it.

Firstly, who exactly are these others who I have called 'those he cares for'? Rational beings, human beings, the subject's gender or race or nationality – the people in his community, his friends, his family, his partner? Here, we must strike a compromise between the extent of the subject's vanity and the degree of moral agreement in those he desires admiration from. My answer is: primarily, all the people that the subject is sentimentally attached to, and feels benevolence towards; and secondarily, all those whose opinion he esteems without knowing them personally, and only in relation to that opinion itself. In a word: all those he admires. In fact, 'being admired' cuts both ways – from one side it is a testament to the subject's personal virtue. From the other, it is a means by which the subject's moral outlook is multiplied and effectuated in the world – for those who admire the subject will have to act in accordance with his moral outlook if they are to sustain their own virtue.

Secondly, are we talking about the desire which motivated the act, the intended act, the intended consequences, the act itself, or the consequences of the act itself? Which of these must be admirable? A simple answer: all that can possibly be worthy of admiration, must meet this mark – and if any of these aspects are found lacking, then the act will be less virtuous, or fall past the

neutral point into reproach. Since admiration is sentimental, the various elements are unified into a single emotion: whether it is admirable intentions producing hateful acts, or hateful means to admirable ends, the subject stands or falls by an emotional reaction to the collective whole. Having said this, I maintain that the act itself – not the intent behind it, or the consequences following on from it – holds the most weight in terms of moral virtue, inasmuch as the act itself produces the most forceful emotional response, and hence eclipses the other elements in dictating what is or isn't admirable. In fact, intentions and consequences can be seen as contingent upon a sequence of choices – where these choices are essentially *stirrings of the will* which only become finalised as a series of actual events.

Thirdly, must this admiration from others be actual or merely potential? I must be very clear here: we are talking about 'being worthy of admiration *and* admired'. Not 'believing that you would be admirable if others perceived your acts'; not 'receiving actual words of praise from others'; not 'being admired for acts you did not commit'; but rather: 'being admired, whether this is made known to you or not, for your own acts, in their entirety'[7]. This is what vanity ultimately desires – actual admiration based on truth, not empty words of flattery, not esteem based on false impressions, not a tremendously admirable life lived in secret. However, more often than not vanity is misled and we fall away from virtue.

Lastly, *when* must the act be admirable? After all, people's emotional dispositions change and they may grow to find different things admirable. I will give a brief answer: 'when the act took place' or perhaps 'during the period when the act took place'. I will refrain from any further clarification, although admittedly there are interesting points to be raised here. If you did grow to find the same acts contemptible that before you found admirable, then you would be in the position of disagreeing with behaviour that, in the

[7] Of course, I am not claiming that the virtue of any particular act is contingent upon whether it is made known to 'those who the subject cares for' – from the perspective of the *all-seeing, all-knowing eye*, the circumstance of this knowledge becomes irrelevant and all that matters is the '*admirability'* of the act. However, from the perspective of the subject, and in accordance with the nature of vanity, the virtue of the act must be approved or verified by the *actual* admiration of those he cares for.

past, was undeniably virtuous. Although that may sound bizarre, I maintain that statements such as 'I was wrong in the past' are not applicable to this situation; say rather: 'I was acting in accordance with virtues that I have now overcome'. The acts for which the phrase 'I was wrong in the past' is appropriate are precisely those which were *immoral at the time* – namely those that were not admirable at the time. These are the things that we should really regret – although our sentiments are not so discerning as to exclude those things that we 'used to find admirable'[8].

Having tried to bring these finer points into the light of clarity, what remains is to ask 'How do we defend this definition?' If *vanity* is the desire for the admiration of others, and *moral instinct* is the desire to be admirable in your own eyes, then by our definition a virtuous act is nothing but an act motivated by, and actualising the desire of, vanity and moral instinct. If there is an accordance between the deed itself, the self-esteem of the subject and the admiration of those he cares for, then his virtue cannot possibly be denied and the deed is 'morally irreproachable'. By this last phrase I mean that our condemnation of his behaviour has no genuine warrant beyond our emotional response to what he has done.[9] Anyone who wants to challenge this will have to produce an example where a person deserves be called immoral, even though they take pride in what they have done and are admired by all those people who they are emotionally involved with. If someone thinks that they have found such an example, let them ask themselves the following:

Firstly, is my disagreement with the actions of this hypothetical person moral or merely formal – in other words, do I genuinely find this person

[8] Furthermore, even a deed which took place over an extended period of time can be seen as a sequence of instantaneous decisions – and moral blame must be directed at one or more or all of these decisions but not at the period of time as a whole. We make choices, *all at once*, so to speak – and then foolishly go on to regret long stretches of our past. Yes, consequences of the act must also be admirable; yes these consequences may influence long periods of time; but in finding them reproachable we are in fact indirectly criticizing the acts that initiated them – and again, we are dealing with a sequence of choices. The choice is only finalised with the advent of the act itself.

[9] Whereas a morally reproachable act would be one where we can tell the subject: 'the people you care for find your behaviour to be reproachable'.

reproachable, or do I fear that I will be reproached for my approval or neutrality in the matter? Secondly, in labelling him as immoral, what am I actually prescribing that he do? What would be the alternative for him – to go against what he feels he should do, to turn away from the people closest to him, to overcome his moral instinct and his vanity, to follow a moral law dictated by strangers? And thirdly, why is the way I would act in his situation not admirable in his eyes, or in the eyes of those he cares for? What would all the people I care for, all the people whose opinions I value, say if they saw me acting in this man's shoes, in the manner I am prescribing to him? would they admire me? Or might they too find me reproachable? If, after having reflected on these three questions, my challenger finds that his example still holds its weight, then my definition of moral virtue is worth nothing at all.

Of course my opponent does not have to agree with the actions of this hypothetical person, nor does he have to find them admirable – all I am asserting is that the virtue of this person cannot possibly be dismissed or disregarded – that his actions are *morally irreproachable*. There is no difference between us instructing him to change his ways, and him instructing us to stop being honest or charitable or any other thing we have deemed to be admirable. Our reproach of his actions would be the equivalent of him reproaching our values of modesty, temperance or diligence, or whatever else we believe to be virtuous. We have as little warrant for our moral outlook as he does for his – in the end, all we can say is that we find certain things to be admirable and others contemptible, and so do all the people that we admire.

We may formally criticise or even abuse our moral enemy, we may wage war on him with all the weapons at our disposal, we may level every sort of reproach at him and administer every sort of punishment – yet ultimately, if we cannot change the things that he and those close to him find admirable, we have not been able to overcome his moral outlook, and his virtue remains wholly undamaged, and thoroughly undeniable. Conversely, the fact that our moral enemy disagrees with us inevitably calls our own moral outlook into question, and makes us wonder if the way we are living is really admirable if this person and those around him find it reproachable.

If, however, we introduce after-worldly punishment and reward – if we begin to speak of the 'eternally admirable' or 'admirable in the eyes of God' – then we have departed from the circumstances that this paper has taken upon itself to consider. For an atheist – or more precisely, for anyone who believes that this life is the only life – that which is admirable must be so because of its significance in *this world*, or more generally, because of its affinity with *purpose* or *nature* of man and his mortal condition. Very broadly speaking, virtue can be equated to wellbeing, health or flourishing – and anything which simulates or accelerates such a condition can be seen as admirable.[10] For those who believe in the after-worldly – those for whom the virtue of an act is divinely sanctioned – what is admirable is given artificially and directly by the laws and teachings of the particular religion that they have subscribed to. They need no warrant for their moral outlook beyond the word of God – their sentimental inclinations as to what is admirable give way to the unquestionable laws that they have artificially taken up, and they are exempt from the kind of moral conflict that I spoke of in the above paragraph. The description of virtue I have given in this paper is in no way prescriptive – rather I have attempted a *metaethical exposition* of the *form* of earthy virtue. And here by 'earthy' I mean disregarding the possibility of after-worldly reward or punishment. I have avoided making any claims about how we *should* lead our lives. Furthermore I have resisted the desire to expose the nature of what is *generally or often admirable* – this would mean an enquiry into what brings health and flourishing, or to use a Greek word of particular significance – into *eudaimonia*. All I have done is related the virtuous to the admirable – and in doing so, taken up a position which is only relativistic to the extent that different people find different things admirable. Nowhere have I denied the possibility that certain acts are *always admirable* or *always reproachable*.

Returning to my definition of a morally virtuous act – and to my challenge to find an example which contradicts it – I can say that, for my part, I have arrived at only two possible cases which appear to threaten my position. These

[10] This is intentionally vague, for I have not the time or space to say anything further – and the purpose of this brief characterisation is merely to provide a counterpoint for a religious notion of virtue.

are: the man who lives in emotional solitude, admiring no one and not desiring admiration; and the man whose closest relatives and friends bear a moral outlook which is in direct conflict with his own. In the first case I answer: social, moral and emotional isolation can indeed produce immorality, but they cannot subdue the thing I call vanity. Among other things, this hypothetical outcast will feel the desire to alleviate his loneliness – his moral instinct will be opposed to the act of 'living in alienation', and this disparity will blemish all his deeds. Whether this man is a criminal or a saint, his deeds will be overshadowed by his moral solitude. Whether it's the 'act of stealing while living in alienation' or 'being charitable while living in alienation', there will be a discord between how this man is living and his moral instinct. In other words, he will not find his deeds admirable, and therefore there is no potential for moral virtue.

Moving onto the second case, it looks as if the subject could be acting as a shining exemplar of conventional morality, yet due to the corrupted moral outlook of those close to him he would be labelled immoral under my definition. To this I say: the act of being emotionally attached to someone you don't morally admire is reproachable in itself. Moral outlooks must be reconciled or else people must move away from each other – and this is in keeping with the definition, inasmuch as people will always feel guilty in loving someone they don't admire, or in hating someone they do admire.

What then, is the cause of all the corruption and dissolution we see in this world of ours? And why are we finding it increasingly difficult to look upon humanity with admiration, if vanity is indeed the powerful force I have made it out to be? Here is my explanation, as foolish as it may sound: immorality is not a 'turning away from virtue' but rather an 'abuse of vanity'. The majority of immoral acts are committed in search of an admiration which never arrives, or which is only superficial, or which is given based on false impressions. As for the rest, it is merely 'weakness of the will', and perhaps even this can be characterised as a series of misguided reflexes aiming for admiration – elements of our nature which have their own unique motivations but actually damage the individual as a whole. It is not clear where instinct ends and volition begin, nor is it clear if we possess free will at all – in the end, the warrant for all moral reproach is grounded in sentiments and these sentiments

are grounded in our *nature*. Blaming someone for acting against our moral outlook, our particular nature, is a necessary and customary part of our lives and the conflict that arises from this kind of blame assists us in refining our notions of what is admirable. Blaming someone for acting against *their* moral outlook, *their* particular nature, is what I mean by the phrase 'morally reproachable' – and it is only in these cases that you can legitimately deny the virtue of their actions. The first kind of blame is directed towards your 'moral enemy'. The second kind identifies 'moral dissolution'. One does not necessarily imply the other.

Finally, a brief and general word about what we find admirable, looking to the *form* rather than the *matter* of the issue. Admiration is intuitive, primal, sentimental, a matter of the heart – yet at the same time guided by the intellect, since we can admire those we have never met. Admiration is subjective and often fleeting – yet at the same time conforms to archetypes that stretch across great periods of human history. What may be admirable in a young person could be reproachable in someone who is older – what we esteem in a man, we may disprove of in a woman. This is one further barrier which prevents us from making claims about the objective virtue of certain acts – at least without recognising a contingency upon other factors. Admiration encompasses not only morality, but other virtues: we admire the strong, the intelligent, the beautiful as well as the kind, the brave and the forgiving. The first type of virtue could be called *passive*, while the second *active*; and we could regard *all deeds* as prone to moral judgement, inasmuch as all deeds admit of being admirable or contemptible. From this perspective, *doing nothing* does not safeguard your virtue – doing the *most admirable thing* at *every single instant* is the only way to seize the immensity of the virtue that lies within your reach, and indeed, within the reach of every single person.

Admiration is somewhat reciprocal – we find it more valuable when received from those we esteem. And so, as a counterpart to the 'vanity of the human heart', we have another desire: namely the wish to see every human soul as commendable and beautiful. Of course, our feeble powers of benevolence always fall short of this – just as our feeble virtues always fail to satisfy our vanity.

British Journal of Undergraduate Philosophy, 1(2), March 2006

Could there be thin particulars?

Gareth Pilkington
University of Durham
g.b.pilkington@dur.ac.uk

If one were to subtract all the properties from a given concrete particular, that is from a given material object such as a chair, what – if anything – would be left over? Is the chair wholly identifiable with its properties such as its colour, its mass and its extension in space, or is there something to which these properties adhere, or in other words, something which acts as the bearer of these properties? In uttering seemingly substantive sentences such as 'the chair is blue' we seem to *speak as though* we are attributing some predicate, in this case 'blue' to a distinct subject, in this case 'the chair'. Is this apparent willingness to attribute to a concrete particular some 'subject' or 'substance' above and beyond its properties merely a quirk of natural language, or have we tacitly hit upon one of the fundamental features of reality? Substance-attribute theorists, such as David Armstrong, would be inclined to agree with the latter. According to such theorists, concrete particulars such as the chair *are* composed of something above and beyond their properties. This further constituent is commonly called 'substance'.

In his books, *Universals: An Opinionated Introduction* and *A World of States of Affairs*, Armstrong gives his own account of substance which involves the notion of a 'thin particular'. A thin particular is crudely characterised as a 'property-less property bearer'. It is defined in opposition to its 'thick' counterpart (which is the thin particular plus the properties it instantiates). Before analysing Armstrong's specific account, I must briefly outline how the notion of 'substance' has been used in the past and for what purposes it has been invoked. Locke[1] tends to equate the notion of 'substance' to that of 'substrata' (*unknowable* particularising entities which are the bearers of

[1] Locke (E2.23.1, E2.31.13).

properties), whereas Aristotle[2] takes 'substance' to refer to individuals such as a dog ('primary substances') and classes of such individuals ('secondary substances'). The common thread that binds these definitions is the claim that substance is first and foremost a concept of that which needs nothing else for its existence. That is, substance has independent existence. Secondly, substance is regarded as a concept of that which particularises a given entity. Substance is sufficient to render two concrete particulars (that is, two objects such as a couple of chairs or a couple of molecules) numerically distinct. I intend to show that although Armstrong's notion of a thin particular adheres to this common thread found in other accounts of substance, it is, at the same time, difficult to provide a coherent, intelligible description of a thin particular and furthermore, difficult to provide a constructive, positive reason that would motivate one to believe in such a notion. I wish to argue that the intelligibility of a thin particular is compromised by its lack of properties. It is difficult to render a concept intelligible without ascribing it some property. However, it will become apparent that as soon as one ascribes a property to the thin particular, the role for which it was intended is undermined. Furthermore, I wish to show that regardless of its unintelligible nature, Armstrong's *argument* for thin particulars renders his theory no more plausible than that of the Aristotelian antireductivist or the metaphysical deflationist.

Thin particulars are a product of Armstrong's desire to reduce "coarse-grained", complex[3] concrete particulars to their "fine grained", metaphysically basic constituents. Like all so-called metaphysical realist substance-attribute theorists, Armstrong holds that concrete particulars are composed of items from two ontologically irreducible categories: instances of universal properties (such as the red exemplified by a red brick) and a further constituent which acts as the bearer of properties. For Armstrong, the bearer of properties is the thin particular. Some may be inclined to conflate the notion of a *thin* particular with that of a *bare* particular[4] or a Lockean substratum. For Locke,

[2] Aristotle, *Categories* (2a35-2b7, argument to establish primary substances as the fundamental entities of his ontology).

[3] 'Complex' in the sense that the particulars are composed of ontologically simpler items.

[4] Gustav Bergmann uses the expression 'bare particular'.

properties/qualities existing on their own – independently of any quality bearer – are inconceivable. In his words "We accustom ourselves to suppose sense substratum wherein [properties/qualities] subsist". "We have no idea of what [substratum] is, but only a confused, obscure one of what it does"[5]. The thought that we *accustom ourselves* to the existence of substrata is not equivalent to the claim that underlying, property-bearing substrata exist *in* the world, or even equivalent to the weaker claim that they could *possibly* exist. Locke is merely claiming that we, as human beings, are so constituted that upon experiencing properties/qualities, we cannot but postulate the existence of some substrata by which such properties are instantiated. For Locke, we cannot come to know anything more about quality-bearing substrata, as all we can experience are the qualities themselves (or rather, in Lockean terms, we may only experience *ideas* of such qualities). To subscribe to the notion of substrata is to subscribe to nothing more than a metaphysical article of faith.

The Lockean stance is outlined firstly to show that Locke does not reify substrata in the way that Armstrong does, but for epistemological reasons is actually critical of such reification; and secondly to show that those who conflate Lockean substrata with Armstrong's thin particulars are mistaken. Armstrong makes a much more substantive claim than Locke regarding thin particulars. In light of the empiricist objection to any reified notion of underlying substrata, Armstrong claims that we can experience concrete particulars as "particulars-having-certain-properties"[6]. That is, experiencing a concrete particular is, *eo ipso*, to experience some particularising constituent which is the bearer of instances of universal properties. Aside from the fact that Armstrong may be accused of begging the question[7], the claim that we experience particulars-having-certain-properties seems dubious when one considers the nature of Armstrong's thin particular.

[5] Locke, in Martin (1980) p.4.

[6] Armstrong (1989) p.61.

[7] Armstrong seems to assume the existence of particularising constituents, which is what he is setting out to prove.

For Armstrong, the thin particular plays a dual role: ensuring numerical diversity between concrete objects and literally bearing all of the properties that comprise the rest of the concrete object. Taking the second role first, a concrete particular such as a red brick requires a property/attribute bearer, as "For each attribute, what literally has that attribute is something whose being what it is does not involve the attribute"[8]. The red brick, in being what it is, involves an instance of the universal 'red', therefore there must be some further constituent of the brick – the thin particular – which in being what it is does not involve the property, but merely instantiates it. However, for the sake of consistency, this reasoning must also apply to the thin particular resulting in a slight variant of Bradley's Regress[9]. That is, the properties that figure in the identity of the thin particular must also have a literal bearer – a further thin particular – and so on *ad infinitum*. In light of the substance-attribute theorist's aim – to establish the fundamental constituents of concrete particulars – this regress is definitely vicious. To avoid the regress, the substance-attribute theorist must claim that "There are subjects for attributes whose identity involves *no* attributes whatsoever"[10]. But, can the idea of a subject for attributes, which is in-itself devoid of attributes, be rendered intelligible?

Sellars maintains that the idea of an attribute-less subject which exemplifies attributes is a "self contradiction"[11]. However, the substance-attribute theorist may respond claiming that to exemplify an attribute is not to possess the attribute; the thin particular is, *in-itself*, attribute-less. Allowing this semantic sleight of hand, the thin particular overcomes this criticism from logic, but the onus is still on the substance-attribute theorist to explain how something can possibly be without attributes.

Is it possible to conceive of something without attributes? It does not seem so, as whenever one attempts to subtract as many attributes as one can from some X, one is still left with some property, be it spatial, temporal, coloured or

[8] Loux (1998) p. 96.
[9] Bradley's Regress applies specifically to the instantiation relation (see below).
[10] Ibid. p. 97.
[11] Sellars (1963) p. 283.

whatever. The substance-attribute theorist may respond that the property-less particular is something *in-the-world*, not necessarily also *in-the-mind*. That is, one does not have to be able to *conceive* of the nature of the property-less particular in order for it to *possibly* exist. Furthermore, as Leibniz notes, it is inevitable that the particular cannot be conceived as "you have already set aside all the attributes through which details could be conceived"[12]. Even if the substance-attribute response is correct – that the property-less particular may exist without one being able to conceive of it – there are still problems with the notion of something being devoid of attributes, but partly constituting a concrete particular. Furthermore, this substance-attribute theorist's response seems to fly in the face of Armstrong's claim that we experience concrete particulars as *particulars-having-properties*. If one could not conceive of the particular which bears the properties, one could not possibly identify it upon experiencing a concrete particular.

To be devoid of attributes is to lack a determinate position in space. If a thin particular has no spatial attributes, it arguably follows that it is immune to causal activity. However, if this is the case it is fair to ask 'how can it be destroyed?' If the answer is 'a thin particular cannot be destroyed', then this seems to be at odds with the bounded temporal careers of concrete particulars of which thin particulars are allegedly a constituent. On the other hand, if it can be destroyed it must be part of the causal order and subsequently possess some spatial property which leads to the vicious regress.

Even if one were to accept that property-less entities could exist and that "particulars neither are nor have natures"[13], when unpacked these claims arguably unravel themselves. Loux asks "Does a thing with no essence have the *property* of being essenceless essentially?" If so, then it is not devoid of all attributes and the regress looms. If not, "Then apparently it could have had an essence, but then… there is another property that is essential to it – that of being possibly essenceless."[14] Either way, the allegedly property-less constituent possesses some property essentially. The substance-attribute

[12] Leibniz (Cambridge 1981 edition) p. 218.
[13] Bergmann (1967) p. 24.
[14] Loux (2001) p. 100.

theorist may argue that this is a somewhat weak criticism. To *lack* a property is not to *have* a property, that is, to lack an essence is not to possess the property of being essenceless. If this were the case, every concrete particular would possess infinite properties, most of which would merely refer to a deficiency of some property, as is the case with 'essenceless'. This seems absurd. However, a much stronger criticism, originally found in Bradley's *Appearance and Reality* (1893), can be levelled against the property-less property bearer. If the property-less property bearer could exist, the relation between it and the properties it instantiates would have to be outlined. A special extra relation of instantiation is needed to weld the particular and the universal together, but then an extra relation is needed to tie the relation, the particular and the universal together and so on *ad infinitum*[15].

The common thread amongst these objections is that the notion of a thin particular is ontologically dubious. However, Armstrong may respond that these objections do not necessarily apply to thin particulars, but rather to bare substrata. A thin particular "is not bare because to be bare it would have to be not instantiating any properties. But though clothed, it is thin"[16]. The thin particular is 'thin' as opposed to 'thick' (the particular plus the properties it instantiates taken as a whole), but not bare. The thin particular has the property of 'being-a-particularising-entity' and, as Armstrong notes, the relational property of instantiation. However it has already been established that the particular cannot possess any properties without incurring the vicious infinite regress. It seems that Armstrong's 'solutions' to the problems regarding the nature of thin particulars merely run into further problems. As Sellars notes, the metaphysical realist substance-attribute theorist can be "observed to leap from the frying pan of one absurdity into the fire of another"[17].

[15] Armstrong (1997) attributes this objection to Quine and F.H. Bradley. For Armstrong, this objection can be overcome by introducing the ontologically additional category of states of affairs. An advocate of the truthmaker principle, Armstrong holds that the state of affairs 'thin particular instantiating property' is the truth *in-the-world* that gives the instantiation relation ontological grounding. Subsequently, a regress involving relations for relations and so on does not occur.

[16] Armstrong (1989) p. 95.

[17] Sellars (1963) p. 282.

At present, the notion of a thin particular seems somewhat problematic. However, Armstrong offers an argument from elimination which allegedly shows that, if one is committed to a theory of universals (which Armstrong believes one should be), then one has to subscribe to thin particulars, as the only other option – the bundle theory of universals – is implausible. Metaphysical realist bundle theorists reduce concrete particulars to bundles of instances of universals 'welded together' by an ontologically primitive relation of compresence. If concrete particulars are qualitatively indiscernible, for bundle theorists who advocate a theory of universals, they must be numerically identical[18]. That is, such bundle theorists take Leibniz's Identity of Indiscernibles to be a necessary truth. For Armstrong and Max Black[19] before him, this is a mistake. Utilising a thought experiment which involves two indiscernible spheres in a symmetrical universe, Black argues that qualitatively indiscernible objects (as regards both intrinsic and relational qualities) may be numerically distinct. Armstrong infers from this that concrete particulars must possess some particularising constituent that explains their diversity. This constituent cannot be composed of properties (other than the particularising property[20]) as it may then have a qualitatively indiscernible counterpart, which would undermine the particularising job for which it was intended. However, aside from the fact that bundle theorists do not take Black's argument to be a decisive refutation of their stance, one should question whether the notion of thin particulars *necessarily* follows from the implausibility of the bundle theory.

As Martin points out, "A philosophical position draws strength from the weaknesses of the positions opposed to it"[21]. That is, Armstrong's substance-

[18] The bundle theorist holds that the concrete particular is nothing more than the sum of its instances of universals. Instances of universals are repeatable (that is, a universal may instantiate more than one particular at the same time). Therefore it is possible to have qualitatively indiscernible particulars, but on the bundle analysis, these would be numerically identical. Consequently, bundle theorists of universals must deny the possibility on qualitatively indiscernible, yet numerically distinct particulars.

[19] Black (1952).

[20] It is 'Clothed thinly'.

[21] Martin (1980) p.10.

attribute theory may draw strength from the implausible nature of the bundle theory. However, drawing strength is not equivalent to conclusive verification. Bergmann invites this objection to his notion of bare substratum in stating "The most one could say" in light of the implausibility of bundle theory "is that the dialectic directs our attention toward what is presented. But it does not and cannot tell us what actually is presented." That is to say, the notion of bare substrata "merely springs from the dialectical needs it satisfies" – allowing qualitatively indiscernible objects to be numerically differentiated – "and is not borne out by careful inspection of what is in fact presented."[22]

Even if one accepts the implausibility of the bundle theory, one is not automatically compelled to endorse Armstrong's ontology. Armstrong represents the ontological analysis of concrete particulars as involving a dichotomy of views. On the one hand, one may adopt the bundle theory of universals; on the other, one may subscribe to a substratum theory of universals. When the bundle theory is 'proven' to be implausible, one has no option but to endorse the substratum view. As Bergmann's fictional objector points out, this line of argument provides no independent, positive evidence in favour of thin particulars; it relies solely upon the refutation of the bundle theory.

Aside from the fact that Armstrong presents little in the way of positive argument in favour of the possibility of thin particulars, or even a coherent explanation of what it is to be a thin particular, his account of the dichotomy between bundle theory and substrata is unfounded. The implausibility of bundle theory could give strength to various theories concerning the ontological analysis of concrete particulars, from Chisholm's[23] deflationary approach – which suggests that a concrete particular is merely a thing which has properties and that is all one needs to say – to the Aristotelian, antireductivist view that particulars such as a dog are, themselves, metaphysically basic. Subsequently, Armstrong's negative account of bundle theory yields no positive consequences that are specific to his stance alone.

[22] Bergmann (1960) p. 616.
[23] Chishom (1969).

After this brief analysis of Armstrong's notion of a thin particular it seems that Locke was correct in suggesting that one may regard an underlying substratum as nothing more than a mere postulate. The dubious character of property-less particulars prevents their reification. Armstrong's notion of a 'thinly clothed' particular arguably escapes vicious infinite regress, but does little to clarify this seemingly unintelligible notion. Furthermore, there is little in the way of positive argument in favour of thin particulars, and, as has been shown, Armstrong's argument from elimination adds only as much strength to his own stance as to that of the Aristotelian antireductivists and the metaphysical deflationists. Consequently I feel compelled to agree with Hume that, at least for the time being, the notion of a thin particular remains an "unintelligible chimera".

Bibliography

Allaire, E.B. 'Bare Particulars' (1963) in Loux 2001.

Armstrong, D.M. *A World of States of Affairs* Cambridge University Press 1997.

Armstrong, D.M. *Universals: An Opinionated Introduction* HarperCollins 1989.

Bergmann, G. *Realism: A Critique of Brentano and Meinong* University of Wisconsin Press 1967.

Bergmann, G. 'Strawson's Ontology' *Journal of Philosophy*, 57, 1960.

Black, M. 'The Identity of Indiscernibles' (1952) in Loux 2001.

Chisholm, R. 'The Observability of the Self' *Philosophy and Phenomenological Research,* 30, 1969.

Crane, T. and Farkas, K. (eds.) *Metaphysics: A Guide and Anthology* Oxford university Press 2004.

Leibniz, G.W. *New Essays on Human Understanding* trans. Remnant and Bennett. Cambridge University Press 1981.

Locke, J. *An Essay Concerning Human Understanding,* 1690. Nidditch P.H. (ed.) OUP 1975.

Loux, M.J. *Metaphysics* Routledge 1998.

Loux, M.J. *Metaphysics: Contemporary Readings* Routledge 2001.

Martin, C.B. 'Substance Substantiated' in *Australasian Journal of Philosophy*, 58, 1980.

Russell, B. *An Inquiry into Meaning and Truth* Allen and Unwin 1948.

Russell, B. *Human Knowledge. It's Scope and Limits* Unwin Brothers 1948.

Sellars, W. *Science, Perception and Reality* Routledge 1963.

British Journal of Undergraduate Philosophy, 1(2), March 2006

Modern logic and how to survive it

Elaine Yeadon
Sheffield University
pia03esy@sheffield.ac.uk
Robert Charleston
The Open University
rc3673@student.open.ac.uk

Logic can be a surprisingly divisive subject. All philosophy students study logic in the sense of *reason*: 'thinking things through carefully, checking they are good arguments'; or 'making sure we are approaching our topic systematically, not obviously going wrong.' Indeed most philosophy courses start by defining the classic *modus ponens* logical technique for finding a third conclusion from two known propositions:

1. Logic is always difficult.
2. Difficult things are always painful.
3. Therefore, logic is always painful.

And its companion *modus tollens*:

1. Fun things always make you smile.
2. Logic does not make you smile.
3. Therefore, logic is not fun.

But far fewer undergraduate philosophers are taught *formal*, *symbolic* logic than twenty or forty years ago. Which can be a problem, since there are papers from the 1960s which are still important and relevant, but which take a working knowledge of logical notation as read. There is a simple test for whether you should read the rest of this paper or not. The next piece in this issue of the journal, by Alex Davies, contains plenty of formal, symbolic logic. Flick through, find the notation, and honestly consider if you: i) can read it

unaided, and ii) know how to check the formal conditions for validity. If you can, you've probably studied formal logic in your own time, or as a course option. You may even have decided you'd quite like to be a logician. With all due respect, you will probably find the rest of this paper too slow and basic. Go and prove some rules for eliminating specific quantifiers – it will be far more rewarding. The rest of us will see you at the next paper.

This paper was written for (and partly *by*) people who fall into the other camp. One of the present authors has training in advanced logic, the other has historically skipped past the sections of papers that used symbolic logic, or has been forced to blithely accept the conclusions such sections reached. Doing so is annoying. So here you will find a potted history of formal logic, and at the end of this paper a crib sheet for non-logicians who need to survive symbolic encounters.

First, it is important to realise that there are several *types* of logic used in mainstream philosophy; that 'logic' can refer to several types of thing, and any one of several systems of representing thought. The original 'logic' is of course what you are used to using when you carefully, analytically think things through. Derived from the Greek for 'word', λογος ('logos' for anyone who did not spend their teenage years knee-deep in Euripides) 'logic' refers to the Ancient Greek predilection for speeches (and therefore arguments) that took an ordered, orderly approach to their topic; that proceeded step-by-step, through justification and reason rather than mystic revelation or poetic narrative. So the word comes to its meaning 'reason' through abstraction, and by extension to the sense 'system of rules by which we evaluate the validity of arguments'. This might seem a trivial, baby-steps starting point, but it carries a serious point: the complex forms of modern, formal, symbolic logic all draw on this common root. Formal and symbolic logics are specialised shorthands for expressing structures of thought and reasoning. A crib sheet such as the one offered here is possible because *everything* even the most complex notational system expresses is *necessarily* renderable in everyday philosophical terms. It may result in unwieldy, complex prose, but a translation of these difficult systems is always possible. If you can already think philosophically, there is nothing in formal or symbolic logic that is beyond your grasp, given a little work.

In modern terms, the kind of logic covered above (implicit in all philosophical reasoning and argument) is *informal logic*. When a set of rules is identified and said to constitute, represent, or model the operation of informal logic, those rules and their application have been made explicit, codified, defined, *formalised*. So talk of premises, propositions, *modus ponens*, *tollens* and deduction are all indicators of *formal logic*.

Almost all UK philosophy students are explicitly taught the basics of formal logic, and their success in studying and writing philosophy proves that you can do most things (ethics, philosophy of mind, aesthetics, and so on) you want in the subject without being *forced* to specialise any further. However, there are occasions when you really need to look at the *general* form of valid arguments in detail, or want to put ideas into particularly complex relations. For instance, you may be studying which things count as causes of other things; or what the set of possible answers to a given puzzle must contain. In such subject areas, you may well end up needing to express those complex inter-relations in a more concise manner than longhand techniques (such as the three-line syllogisms at the start of this paper) will allow. You need, for clarity and manageable article-length's sake, to start using symbols as substitutes. Then you can stop writing 'is the same as', 'and', 'not', or 'or' all the time, and start using '=' and other appropriate symbols instead. And you can stop making up real-world examples about bachelors, runaway trains and babies, instead – just as in algebra – using letters to stand in the place of whole sentences such as 'logic does not make you smile'. This kind of symbolic logic is *propositional logic* or *propositional calculus*, or *sentential calculus*. Leibniz had a go at producing something of the kind, but Boole really got it up and running in the mid-Nineteenth Century.

In its unmodified state, though, such propositional calculus has two key flaws. First, it misses out a lot of the key terms we use in thinking. We use ideas such as 'all', 'every', 'none' and 'some' *all* the time. (See?) The propositional logic mentioned above has a problem with such *quantifiers*: namely, that if all you've got is 'true', 'false', 'and', and 'or', all you can say is: 'and this one is false, and this one is false, and this one is false…' You cannot say 'all are false'. Second, we often want to deal with ideas at a more detailed

level than an entire sentence / proposition. Sure, 'logic does not make you smile', but we have other – related – questions within the same topic, such as 'what about "logic does not make you rich" or "logic does make you more attractive to others"?' We want a set of symbols that does not see these three ideas as *entirely* different things, but rather can see something in common – 'logic' – and something different – 'smile', 'rich', 'attractive to others' – between the three.

Adding quantifiers and breaking down the sentences we consider into subjects ('logic') and predicates ('is not fun') extends propositional logic into *predicate logic*. This is the famous stuff, put into practise by Frege, developed by Hilbert and Ackerman in the early Twentieth Century, that you have almost certainly been skipping over in philosophy papers. It can be further complicated by using second-order variables, but this is the basic idea.

There are other bits and pieces. For example, *modal logic* allows you to start doing the kind of algebra covered above for something being, or not being, 'necessarily', 'possibly' or 'contingently' something else. *Mathematical logic* develops further into concepts of proof, computation, model and set theory. And *philosophical logic* – confusingly, perhaps – is a term used to refer to the study and development of the concepts – proposition, identity, meaning, analyticity, etc. – that we use in putting together formal logic systems. Furthermore, there is of course a huge debate on what rules, what descriptions, even what status we ought to accord and derive in and from our formal logic system(s). But needless to say, if you want to study these in detail, you will need to study the subject in a great deal more depth than can be included here.

For the moment, it should be enough to move on to the crib sheet for reading the notation you will find in mainstream philosophy papers. We have tried to cover the main symbols and concepts. For the symbols we cannot translate in such a plain medium, we have included the terms you need to type into Google to learn more.

A crib sheet for reading logic

Symbol styles:

Reading logic often requires you to be fairly flexible. There are, helpfully, several sets of symbols that mean roughly the same thing. There are – just as with spellings – 'English' and 'American' standards. But – as with spellings – you'll find people often end up mixing and matching the two. We've used the English standards for preference, but have also included the American versions where they seem likely to crop up.

Propositional logic – P, Q, ¬, ∧, ∨, ⊃, (), ≡, and ⊢

P, Q

Single letters such as **P** or **Q** can represent single sentences, **propositions** such as 'Sam is wearing a hat'. The author should specify what sentence each letter stands for and then use the same letter to refer to the same proposition throughout the argument.

¬ or ~

The **negation** symbol ¬ is roughly the same as saying **not** or **false** in everyday English. It goes before a proposition that is being denied and is traditionally interpreted as '**It is not the case that**…' So when P stands for 'Sam is wearing a hat', **¬P** should be read: "**It is not the case that** Sam is wearing a hat."

This reads as overwrought prose, but simplifying to 'Sam isn't wearing a hat' should only be done with great care. It works OK with Sam, but consider what happens to 'All cars are blue': you might end up with 'All cars are not blue' (no cars are blue) – which is a very different claim from the correct 'It is not the case that all cars are blue' (some cars may be blue, or may not, but not all of them are blue). Beware misleading interpretations – always use the unwieldy but accurate translation first.

$\wedge$ or $\bullet$ or &,

$\vee$,

$\supset$ or $\rightarrow$

These are **sentence connectives**, and can be read as **and** $\wedge$, **or** $\vee$ and **therefore** $\supset$ respectively. Their technical names are **conjunction**, **disjunction** and **conditional**. Simple propositions can be combined to form more elaborate expressions just as we can join propositions together in English: 'I'm late for work and the car won't start.'

Conjunction $\wedge$ works pretty much the same way as **and** in plain English. $P \wedge Q$ is true when both P is the case and Q is the case.

Disjunction $\vee$ works pretty much the same way as **or** in everyday language, with one serious proviso. Consider two questions:

1) 'Are you tired or bored?'
2) 'Is Sartre alive or dead?'

You can answer questions like 1) with 'tired' P, 'bored' Q or 'a bit of both, really' $P \wedge Q$. The question is about an **inclusive disjunction**. Whereas 2) *cannot* meaningfully be answered with 'a bit of both'. It is an either-or question, an **exclusive disjunction**. The kind of 'or' $\vee$ refers to is the **inclusive** question 1) type. As such, $P \wedge Q$ is true when P is true, or Q is true, or *both* P and Q are true.

(), [] and {}

Brackets are often used to keep terms together. Let's say P is 'Sally's in a mood', Q is 'John is hiding', R is 'The car broke down' and S is 'They are running late'. If I ask the question 'Why aren't they here yet?' and my logician friend answers: $(P \wedge Q) \vee (R \wedge S)$, I know that she means 'It is either the case that Sally's in a mood and John is hiding; or that the car broke down and they are running late' and understand that she's also allowing for *both* these contingencies to be true.

Out of interest, you now ought to be able to work out and appreciate, from looking at the above and translating the terms, why the **exclusive disjunction** is written as: $((P \vee Q) \wedge \neg (P \wedge Q))$

Sometimes you will find that authors use square brackets when nesting bracket pairs, as in the above. Here's an example where this has just been done for readability, the square brackets meaning the same as curved brackets, but hopefully making it clearer which open-bracket goes with which close-bracket: $[P \vee (P \wedge Q)] \vee [P \wedge (P \vee Q)]$

You will also see curly brackets {} used in logic. These can be used to indicate a **set** of objects such as $\{P, P \wedge Q, Q \vee R\}$ – which can be taken intuitively as 'a list of things' if you want to keep your reading quick and shallow, or investigated in depth as part of set theory and predicate logic. (In this sense they really belong on the next section, but it seemed best to mention it as early as possible). But {} can *also* be used to indicate which prior statements a later line of reasoning **depends on** for its truth. So if you see a numbered list of propositions:

{1}	1.	Blah, blah, blah.
{2}	2.	Some more blah, blah, blah.
{1, 2}	3.	Some derived blah, blah, blah.

Then you know that it is being announced that 3. **depends** on 1. and 2. as an assertion.

$\supset$ or $\rightarrow$

Hook $\supset$ (and the American **arrow** $\rightarrow$) mean that you are looking at a **conditional**. Just as in normal philosophical usage, if the stuff on the left (antecedent) is true, the claim is that the stuff on the right (consequent) *must* also be. $P \supset Q$ means '**If** P, **then** Q'.

It's quite common in logic to write out truth tables for conditionals, just to make sure they fit with our experience of the world, when P and Q are substituted with real-world states of affairs. Here's one for $P \supset Q$:

When P is…	…and Q is…	P ⊃ Q is…
True	True	True
True	False	False
False	True	True
False	False	True

The first two lines match our informal logic intuitions. But note what happens in lines three and four, when the antecedent is false. The proposition turns out true. This allows for the following conditional statements to be true:

i) 'If the Earth is flat then it might rain tomorrow.'
ii) 'If cows are green then pigs are blue.'

So begins one of the great debates in logic: conditionals, counterfactual conditionals and implication. You can look into the reasons for this, and what fixes have been proposed for the basic problem, but as a 'translator' just make sure you check conditional statements properly as you read through an argument. Check they aren't being used as misleading, arbitrarily true premises purely on the basis of a false antecedent.

≡ or ↔

This is the **biconditional**, also known as **equivalence**, familiar in mainstream philosophy as **iff** – 'if and only if'. It means the implication runs both ways, that if you've got **P**, you've got **Q** and **vice versa**. P and Q must be *both* true or *both* false, otherwise the overall **biconditional** is false, as can be seen from the truth table:

When P is…	…and Q is…	P ≡ Q is…
True	True	True
True	False	False
False	True	False
False	False	True

In line four, the infamous problem mentioned above crops up again.

It is also important to note that the only way **P** and **Q** are required to be equivalent in the statement '**P ≡ Q**' is that they have equivalent truth values. The claim is *not* that two propositions with the same truth value are *identical*, or *the same thing*. Though this might, of course, be the case. Again, be careful with your translation.

⊢

This symbol is called **turnstile** or **inference**. When translating, it is perhaps easiest to think of it as an interim conclusion. Basically, whatever is to the left of it is claimed to infer what is to the right of it. This is taken to be an active inference, rather than a conditional, as is the case with ⊃. Try substituting in '**proves**' or '**shows**' in plain English to get the meaning.

So **P ⊢ Q** means '**P shows that Q is the case**'. This is far easier if you are translating than if you are writing logic, in which case you will need to study the niceties of the concept in far greater detail.

Predicate logic – a, b, c, =, F, G, H, ∃, ∀, x, y, w, P

a, b, m, n, F, G, H – names and predicates

Lower case letters such as **a**, **b**, **c**, **n**, **m**, and **o** are usually **names** in predicate logic, specifying particular individuals or objects within the class of objects we are interested in (the 'domain of interest'). For example **c** might pick out the individual **Charlie** from the domain of all people. However, different names need not denote different objects all the time. Named objects are often stated to be identical using the notation **a=b**. In other words, one object has two names.

Upper case letters such as **F**, **G** and **H** are used to denote the properties or **predicates** that apply to the things being named. For example, where **F** is the **property of being female**, and **a** names **Anna**, the statement **Fa** can be interpreted as '**Anna is female**'. Traditionally predicate symbols are upper case letters, sometimes Greek letters.

F, G, H - relations

In addition to their use as predicate symbols, upper case letters can denote a relation between a number of objects, such as **R** indicating '**is taller than**'. So **Rab** might translate as **Andy is taller than Bob**. Be careful about how the author has defined the usage of such relations. The author might order the names the relation applies to differently from the way you would intuitively expect. If the author thinks in terms of ascending height, for instance, **Rab** might indicate **Bob is taller than Andy**. The definition should be in the paper you are reading.

∃, ∀, x, y, z, w, P – quantifiers and variables

The **existential quantifier** ∃ can be read as meaning '**There exists some thing…**' The **universal quantifier** ∀ can be translated as '**For all things…**' These quantifiers are used along with **variables** to refer to the things which are being quantified.

x, **y** and **z** are usually used as those **variables**, standing in for objects or abstract entities. In logics about possible worlds, **w** is often used to refer to a *whole* world, rather than a thing *within* a world.

So, combining the above, **∀w** naturally read as '**For all worlds…**' and **∃w** '**There exists some world for which…**'

Variables can also, in second-order logics, be **properties** – usually denoted by the letter **P**. So, **∃P (Pa ∧ Pb)** says '**there's a property such that Andy and Bob both hold that property**'.

In each case, though, it is the *function* of a variable which matters, not the symbol used. When you see a variable, you know that *something* is being proposed or described, but not necessarily *which* thing.

Rather than saying of Anna: she's my friend and she's in Glasgow **(Fa ∧ Ga)**, an author may be deliberately leaving things open: **∃x (Fx ∧ Gx)** – **there exists *someone* who's my friend and they are in Glasgow**.

Quantifiers work the same way as the negation symbol – if placed immediately outside of a set of brackets, they apply to everything within those brackets. If not immediately outside some brackets, then they apply just to the statement immediately adjacent. So the following two statements say different things:

(i) **∃x (Fx ⊃ Fa)**
(ii) **(∃x Fx ⊃ Fa)**

The first clearly states the existence of some **x**. The second makes **x** part of the conditional's antecedent claim, so allows that there may in fact be no **x** at all – a significantly weaker claim.

The order in which quantifiers occur can also be important.

If we define **Kxy** as meaning '**x knows y**'

(a) **∃x ∀y Kxy**

does not say the same thing as

(b) **∀y ∃x Kxy**

(a) holds that there is at least one person whom everyone knows while (b) holds that all people know at least one person.

As above, negated quantifiers need to be translated carefully. **¬∃x** should be read as '**it's not the case that there exists…**', and **¬∀x** as '**it's not the case that (for) all x…**'

Modal logic – ◊ and □

◊ can be read as '**it is possible that…**', and □ as '**it is necessary that…**' These modal operators depend on their order for meaning, just as quantifiers do. So **◊□P** reads as '**It is possible that it is necessary that P**' (or '**Possibly, necessarily P**'). Whereas **□◊P** means '**It is necessary that it is possible that P**' (or '**Necessarily, possibly P**').

Again, ◊ and □ affect just what they are adjacent to. In **◊((P ∧ Q) ∨ R)**, ◊ applies to the whole statement; but in **(◊(P ∧ Q) ∨ R)**, it just applies to **(P ∧ Q)**.

And again, negated modal operators need to be translated carefully. Sticking to the general form of 'It is not the case that it is possible that P…' will keep your translation on the straight and narrow, but there are other more reader-friendly equivalents for specific combinations. For example:

1) **¬◊P**: '**Not possibly P**'
2) **◊¬P**: '**Possibly not P**'
3) **¬□P**: '**Not necessarily P**'
4) **□¬P**: '**Necessarily not P**'

All of which, hopefully, will be sufficient for most mainstream papers which use some symbolic logic[1].

[1] One final tip is that if you need to include logical notation in your own work, and use Microsoft Word, you will find all the symbols you need in the 'Lucida Sans Unicode' typeface. Insert > Symbol > select the correct font, and you will then be able to scroll down to the symbols you need. Many other typefaces do not include all the operators covered here.

British Journal of Undergraduate Philosophy, 1(2), March 2006

Must a pragmatic theory of explanation mean 'anything goes'?

Alex Davies
Selwyn College, Cambridge
asd32@cam.ac.uk

Bas van Fraassen conceives of explanations as answers to contrastive why-questions. E.g. why was Sam late for School *rather than on time*? - because he missed the bus. Roughly speaking, 'because he missed the bus' is an explanation because it is an answer to a why-question. He explicitly denies that explanation is the same as understanding.[1] A model of explanation, as van Fraassen sees it, should characterise what it is that makes a piece of information, when added to that person's background knowledge, capable of improving that person's understanding. We must characterise the missing piece of the puzzle that when added to the rest of one's background knowledge, yields understanding. This is a very pragmatic notion of explanation. What counts as explanatory changes depending upon context (which includes the background knowledge and the interests of the inquirer).

I find this notion of explanation very appealing because it seems to stand much closer to reality than accounts which draw a thick line between 'description' and 'explanation'. If correct, van Fraassen's account shows how this line shifts with context.

However, in their paper 'Van Fraassen on Explanation', Kitcher and Salmon claim his model is too weak.[2] In its efforts to allow such plurality in what counts as an explanation, van Fraassen ends up allowing anything to count as explaining anything else, *and very well*. They show this by providing an answer with a logical form that passes the evaluation criteria van Fraassen uses

[1] Van Fraassen, 'Salmon on Explanation', p641.
[2] Ruben, *Explanation*, pp310-325.

to measure the quality of an explanation. This paper defends van Fraassen against their criticism by revealing a flaw in their counterexample answer, albeit with a few modifications to van Fraassen's model of explanation.

His model of explanation divides a contrastive why-question (e.g. Why was Sam late for school rather than...?) into three parts.[3]

> **Topic (Pk)** – The topic is what actually happened (e.g. Sam was late for school).
>
> **Contrast Class (X)** – The contrast class has members that are alternatives to what actually happened. (e.g. Sam was early for school/Sam was on time/Sam was late). Note that the contents of this class are determined by context. When asking the question, 'Why was Sam late for school?', I could be focusing on the fact that Sam rather than anyone else was late, or that he was late rather than early, or that he was late *for school* (rather than anything else). The contrast class includes Pk.
>
> **Relevance Relation (R)** – The relevance relation stands between an answer A and the contrast class X. i.e. R(A, <Pk, X>). The relevance relation is not constant. It can change from context to context. For example, in one context the relation might be causal relevance (i.e. A caused Pk rather than anything else to happen), in another it might be entailment (A entails that Pk happened rather than anything else).

So a contrastive why-question can be represented as: **Why Pk rather than any other member of X? Because A**.

For present purposes the only other parts of this model we need to consider are the criteria used to evaluate how good an answer is to a given question. I.e. how good an explanation of Pk is the answer A?

> **K** - this symbolises the questioner's relevant background knowledge.[4]

[3] Van Fraassen, Op Cit, pp141-143.

K(Q) – this symbolises a subset of K where knowledge that Pk occurred and all other members of X did not, is excluded.

For A to be a *good* answer to a given contrastive why-question (Pk, X, R):

a) **A must be probable** relative to K (e.g. Given what I already know, it is likely that Sam missed the bus).

b) **A must favour Pk relative to other members of X** with respect to K(Q). I.e. A must make Pk more probable than each other member of X (e.g. A conjoined with a subset of my background knowledge K(Q), must make it more likely that Sam was late for school than each of his classmates).

c) **A must compare favourably with respect to alternative candidate answers** in the following respects:

   i) There shouldn't be other answers more probable than A with respect to K.

   ii) There shouldn't be other answers more strongly favouring Pk relative to other members of X with respect to K(Q).

   iii) There shouldn't be other answers that render A irrelevant to X because they screen A off from Pk and the other members of X.

And in addition to these, A must stand in relation R with <Pk, X>. **A must be relevant**.

[4] Note that there's a vagueness as to exactly what 'relevant' means, and this cascades into the rest of the definition of a good explanation. I'm aware of this. What I am assuming is that some independent account of relevance could be constructed to fill this 'black box'. I have in mind some psychological theory that gives an account of why a person finds certain things relevant to others. This is admittedly nothing but science fiction now, but some such account could conceivably fill that step in the definition.

The introduction of K(Q) in (b) and (c)ii is of central importance.[5] If K were used instead of K(Q) then the very fact that the inquirer knows that Pk occurred would render every answer, A, no matter what it said, irrelevant. The reason? A in addition to K instantly favours Pk over all other members of X because Pk entails Pk giving it a probability of 1, whilst it also reduces the probability of all other members of X to 0. However, performing (b) and (c)ii with respect to K(Q) aims to avoid this problem.

The trouble here for van Fraassen is that his attempt to dodge this problem fails for the following reason. (For brevity I will use the symbol ¬Pi to symbolise the denial of all members of X other than Pk.)

Rather than stating (Pk & ¬Pi) in our background knowledge, which he rules out, we can do this inside the answer A. In fact, there's nothing stopping us from giving just that as our answer. But this is easy to avoid - we merely add another restriction so that we have two restrictions in total:

R1: Part (b) and c(ii) of the evaluation criteria must be performed with respect to K(Q), not K.

R2: The answer, A, must not include propositions that state Pk and ¬Pi.

But again, these are just as easy to sidestep. Here's how. Suppose we define R such that R(A, <Pk, X>) iff A entails Pk.

Let's take the answer to be:

A1: [S1 & (S1 ⊃ Pk) & (S1 ⊃ ¬Pi)]

....where S1 is some true statement that is in the inquirer's background knowledge K(Q).

[5] Van Fraassen, *The Scientific Image*, p147.

A1 entails Pk, so A1 is relevant to our question (whatever it may be).

a) A1 is probable with respect to K because K includes our knowledge that Pk, ¬Pi and S1. So, given the definition of '⊃', all three conjuncts come out true, i.e. with high probability.

b) A1 favours Pk over all other members of X because S1 is a member of K(Q). So, given that A1 is true, when A1 is conjoined with K(Q), it entails Pk and the falsity of all other members of X.

c) There isn't a better answer to the question, because A1 *entails* Pk. Entailment is monotonic – no matter what other information we consider, A1 will still entail Pk. So no other answer could be *better* than A1.

So A1 passes the criteria.

This is where Kitcher and Salmon think they have van Fraassen between a rock and hard place. And they glorify this with a concrete example where van Fraassen is committed to saying that astrology explains why JFK died when he did, and does so very well.[6]

Suppose the question is asked, why did JFK die on 22 November 1963? Then...

Pk = JFK died 22 Nov. 63.

X = {JFK died 1 Jan. 63, JFK died 2 Jan. 63,...,JFK died 31 Dec. 63, JFK survived 1963}

R = the relation of astral influence, and as they put it, 'One way to define R is to consider ordered pairs of descriptions of the positions

[6] Ruben, *Explanation*, pp317.

> of stars and planets at the time of a person's birth and propositions about that person's fate.'[7]

So, treating S1 as a true and accurate statement of the positions of the planets and stars when JFK was born, our concrete instantiation of A1 becomes:

> CA1: S1 & (If S1, then JFK died on 22 Nov. 63) & (If S1, then JFK did not die on 1 Jan. 63;...JFK did not die on 21 Nov. 63 & JFK did not die on 23 Nov. 63; & ... & JFK did not Survive 1963)

CA1 was created simply by filling in A1 (a schema) with Pk and X as I have just defined them. The relevance relation for CA1 is different from the relevance relation for A1. In A1 it was entailment, here it is a list of ordered pairs of propositions. Nonetheless, CA1 passes the evaluation criteria for the same reasons as A1.

It is relevant (given the new relevance relation). CA1 is probable with respect to K, given that K includes Pk and a denial of all other members of X – and so passes part (a). CA1 favours Pk over all other members of X because CA1 entails Pk, and the falsity of all other members of X – thus passing part (b). And there are no better answers than CA1 because CA1 is basically just an instance of the relevance relation, if not a statement of the relevance relation itself – thus passing part (c).

So CA1 passes van Fraassen's evaluation criteria with flying colours. But I don't think van Fraassen is in all that bad a position – of which I will now try to persuade you.

How does A1 pass part (b) of the evaluation criteria? A1 favours Pk over all other members of X because I said A1 conjoined with K(Q) entails Pk and ¬Pi. Note that entailment is a relation that does not concern truth (entailment is not a question of soundness, but rather a question only of validity). A1 entails Pk because it is a valid inference from 'S1' and '(S1 ⊃

[7] Ibid.

Pk)', given the formal definition of '⊃', regardless of whether they actually state something true.

However – and here's where the first modification of van Fraassen's model is to be made – for this *modus ponens* inference to be made *by the questioner*, the questioner must know both 'S1' and '(S1 ⊃ Pk)' to state something true. The questioner just cannot infer Pk if he/she doesn't know these two premises to be true. The modification to van Fraassen's model we need is to focus not on entailments but on inferences the questioner can make given his/her background knowledge. This modification blocks part of Kitcher and Salmon's counterexample. For we can accept that the questioner knows S1 to be true, because we supposed S1 was true and that S1 was some statement included in K(Q). But it's not clear that we are committed to saying that the questioner knows that '(S1 ⊃ Pk)' states something true, or even that it is true in the context of (b) – which is distinct from the context provided by (a) due to the shift from the use of K to K(Q).

So now we have two routes to explore. Firstly, there is the case where the questioner does not know that '(S1 ⊃ Pk)' is true. And secondly there is the case where it is known to be true.

Let's address the simpler case, where the questioner does not know '(S1 ⊃ Pk)'. If this is not stated in the questioner's background knowledge then there is nothing, for the purposes of part (b), which can inform the questioner as to its truth value. The reason for this is that unless the truth value of the statement is just given to us (which we are supposing it is not) the only way to gather the truth value of a truth functional statement is to know the truth values of its parts. But because this particular proposition has the truth value of 'Pk' as its consequent, this is explicitly ruled out for part (b). Why? Because this part of the evaluation takes place as though we had no knowledge that Pk occurred. Since we know S1 to be true, we cannot know that '(S1 ⊃ Pk)' is true merely because the antecedent is false. And because '(S1 ⊃ Pk)' is then true only if Pk is true and otherwise false, and that we do not know the truth value of Pk, we cannot give '(S1 ⊃ Pk)' a truth value.

The upshot of this, on the supposition that the questioner doesn't know the key conditional in the modus ponens inference (in A1) to be true, is that the restrictions R1 and R2 are enough to stop *explanans* which include statements of the *explanandum* from being counted by van Fraassen's criteria as good explanations. This manoeuvre maintains entailment as a legitimate relevance relation provided that the questioner does in fact know the conditional premise to be true i.e. provided that the questioner can *infer* Pk from K(Q) plus A.

There are two worries that may arise upon reading this suggestion.

Firstly, notice that the counterexample answer, A1, works by being an instance of self-explanation. Self-explanations have the general form, 'P because P' or less schematically an example would be, 'he ran down the hill because he ran down the hill' – the *explanandum* is given as the *explanans*. A1 is just a complicated self-explanation. It is true that A1 doesn't assert the *explanandum*, but it does imply it through a *modus ponens* inference, and this is the reason it passes van Fraassen's evaluation criteria:

A1: $[S1 \& (S1 \supset Pk) \& (S1 \supset \neg Pi)]$

Because my defence against A1 is based on ruling out self-explanations altogether, there may be worries that some self-explanations are in fact good explanations. My defence would then be misclassifying some self-explanations. An example might be this. A child says to his father 'why is it wrong to kick people?' and the father might just reply 'because it's wrong to kick people,' and it is perfectly plausible that the child find this answer genuinely explanatory. It appears then that we have a self-explanation which is a good explanation.

But I don't think this is an accurate description of what is occurring here. Just because the answer when written down and removed from its context appears to be a mere restatement of the *explanandum* (i.e. it is wrong to kick people), doesn't mean that it *is* a mere restatement of the *explanandum*. In the father-son example I strongly suspect that if a child did find such an answer explanatory, it would be because the son dubbed his father some kind of a

moral authority. And in that instance, the act the father performs by saying the phrase 'because it's wrong to kick people,' is bound up with more information than had appeared in the uttering of the *explanandum* in the why-question by the son. Consequently, this example isn't an instance of 'P because P', because the two occurrences of 'P' denote different information. I will contend that if this isn't so, i.e. if the restatement of the *explanandum* is a genuine attempt at self-explanation, then the son wouldn't find the reply explanatory. The son would continue asking why-questions. Because of this, I don't see the need to worry about the possibility that some self-explanations are good explanations.

Secondly, and more worryingly, this manoeuvre rules out the possibility that van Fraassen's model could count *self-evidencing* explanations as good explanations. Self-evidencing explanations occur when the *explanans* is given high probability only once the *explanandum* is known to have occurred. For example, suppose John sees a woman outside a church all dressed in black and he asks 'why is that woman dressed in black?' Sam replies 'because she's at the funeral of someone she knows,' and then John asks why Sam thinks that's a likely explanation. John can quite legitimately just reiterate the *explanandum* of the why-question viz. 'the woman is dressed in black outside a church'. The *explanandum* is what makes the *explanans* likely. But given my new stipulation that any conjuncts in the *explanans* must be held to be probable by the questioner, combined with the fact that in criterion (b) knowledge that Pk and ¬Pi (i.e. knowledge about of the *explanandum*) is excluded from consideration, self-evidencing explanations will often not be probable when considered inside criterion (b).

One response to this problem might be that knowledge of the *explanandum* (e.g. the woman is wearing black) is only a small part of the many beliefs that result in an attribution of high probability to the *explanans* (e.g. she's at a friend's funeral). Other pieces of information that would be important for making the *explanans* likely include what people wear at funerals in general, that most people do not casually walk around clad entirely in black, that most people do not spend most of their time outside churches, that funerals are often held at churches and so on. So, one possibility is that we could take this other information and say that the *explanans* is likely on the basis of it. The

problem with doing this is that on the basis merely of this contextual information, it is not clear that some other answer wouldn't come out with a higher probability. When the *explanandum* directly makes the *explanans* likely, it is clear that self-evidencing explanations are going to do well as compared with other possible answers. But by ignoring the *explanandum* in the way I suggest, this advantage is lost because the contextual information alone is not enough to ensure that the self-evidencing explanation (i.e. that she is at a friend's funeral) will be deemed more likely than other answers (e.g. she's going to the shops, she's lost, she's upset because her boyfriend just dumped her, or whatever).

Although there may be another way of dealing with self-evidencing explanations inside the model which I have not considered, it seems likely that self-evidencing explanations cannot be accounted for on this model.

So now for the second supposition; that the questioner does know the conditional '(S1 ⊃ Pk)' to be true. To draw conclusions from this though, we must discuss what it is a questioner knows if he/she knows '(S1 ⊃ Pk)' to be true.

When I ask what it is the questioner knows when she knows '(S1 ⊃ Pk)', I want a precise interpretation of '⊃'. Now, '⊃' has a meaning which is as specific as we could hope for. It is given in the truth table that makes 'A ⊃ B' false only when B is false and A is true, otherwise the conditional statement comes out true.

On this interpretation a person who knows '(S1 ⊃ Pk)' will indeed be able to infer Pk (because the inquirer already knows that S1 is true) and so A1 will pass part (b) of the evaluation criteria.

However, this interpretation is dodgy because '⊃' has a very specific definition that makes its truth depend exactly on the truth values of its consequent and antecedent. This means if we were to fill in the schematic symbols (i.e. Pk, ¬Pi, and S1) the statement given would say very little, if anything about the world.

We can see this if we refer back to the JFK example. Take the first conditional of CA1, '(If S1, then JFK died on 22 Nov. 63)'. Taking the conditional to have the meaning of '⊃', what makes it true is merely that the antecedent is true and the consequent is true, or else when the consequent is false. But in these truth conditions you will find no appeal to any astrological influences, no appeal to any link between stellar constellations and JFK's death.

What has slipped through the criteria is a statement that allows JFK's time of death to be *inferred* (rather than merely entailed), given background knowledge K(Q). But this inference only arises out of the strict definition of '⊃' and the definition of '⊃' says nothing about any link between the antecedent and the consequent – *all it says is that in reality it is not the case that the antecedent is true and the consequent is false*. Nothing is said about any link between two kinds of thing, namely, times of death and stellar constellations at times of birth. In other words, it is not astrology that has been let through the criteria, but a logical statement of affairs that can't plausibly be said to concern any metaphysical claim.

Nonetheless, this answer does pass van Fraassen's evaluation criteria. How can we stop it?

Notice that if we were to take the same relevance relation and apply it repeatedly in different questions, there would indeed be instances where other instantiations of A1 would come out false, because a person had not died at the time stated in the consequent of the relevance relation. This seems inevitable if there is no link between the constellations of the heavens and events in peoples' lives (including their time of death). The only way this could be avoided is if the relevance relations used were tailored to the person (P) who was the subject of each question of the form, 'Why did P die at time t?' I.e. if we were to gerrymander each relevance relation so that answers like A1 both counted as relevant and passed the evaluation criteria (as was done with the definition of R in CA1). The criteria need to be amended so that they can distinguish between answers which on one rare instance happen to

appear good but are generally bad (as judged by the criteria), and answers which are generally good.

The difference shows itself if we reuse a relevance relation with different questions. So the solution is to require that the evaluation criteria hold generally for a given relevance relation. This is what I call the generality condition. For the purposes of avoiding Kitcher and Salmon's criticism, we can say this condition requires that a relevance relation must yield good answers (as judged by the criteria (a) – (c)) more than once.

With this addition to the criteria, the counterexample fails to pass. I will not be more specific than this as to what the content of the generality condition should be, but I will make two remarks. Firstly, it needn't be so strong as to require that a relevance relation yield good explanations in all questions in which it is used. Such a condition is clearly too strong – it would rule out relevance relations that are commonly accepted. For example, it would rule out both relations of causality and of entailment. If we want an explanation for why P entails P, we cannot use the relevance relation of causality between the *explanandum* and the *explanans* because P does not cause P to be entailed by P. Causation just doesn't come into it. And we can easily reverse the situation to show that in cases of causal relevance there is no entailment involved. Secondly, it would be foolhardy to make the condition require only that a relevance relation provide good explanations in 'more than one' why-question. It is quite conceivable that a relevance relation be constructed which yields good explanations in only two (or some other uselessly small number of) why-questions. It might be therefore that rather than using an *absolute* requirement where a relevance relation must yield a good explanation for a definite number of why-questions, we instead use a *relative* requirement which operates on the basis of how many times a relevance relation can yield a good explanation in comparison with the number of times other relevance relations can yield good explanations.

It should be clear that if A1 stated something more substantive then it would succeed in saying something about the world. To say something more substantive A1 could be restated using a counterfactual conditional rather

than merely '⊃'. By making claims about counterfactual situations, A1 would be making a claim about an observable correlation between the constellation of the stars and events in the lives of us poor humans. However this statement, being a genuine statement of astrology, would fail to pass another set of criteria that van Fraassen provides. They aren't central to this discussion, so these criteria haven't been stated here, but they set standards for whether or not an instance of A is actually an answer to a why-question at all. That is to say, these criteria don't *evaluate* the attempted answers. They are preliminary to the evaluation criteria. This 'answer' criterion includes the provision that A must be true, or at least that the questioner must believe it to be true, for the questioner to consider it an answer to his/her why-question. Because it is generally known that the predictions astrological theory makes are often false, it is likely the questioner will reject a substantive version of A1, calling it not just a bad answer, but denying that it is an answer whatsoever to his/her why-question.

Thus, the counterexample provided by Kitcher and Salmon is nothing of the sort. They took their counterexample to show that a pragmatic theory of explanation is hopeless. The only way they thought van Fraassen could fix his model was to introduce an explicit division between a class of genuine relevance relations which in fact yield explanations, and another class of non-genuine relevance relations that fail to yield explanations. I have shown instead that only three changes are needed to properly characterise their counterexample, and none of these stop the model from being a model of how explanation is *pragmatic*. The first change was to introduce R2 in addition to R1. The second was to focus on inferences that can be made by the questioner given his/her background knowledge, instead of entailments. And the third change was to introduce the generality condition to the evaluation criteria. With these we can hold onto the central idea of the model (that explanation is pragmatic) contrary to the suggestion of Kitcher and Salmon.

However, as stated above, a drawback of these manoeuvres is that self-evidencing explanations will not be counted among good explanations. Some may see this as fatal to my defence. The extent to which this defence is a good

one will therefore coincide with the significance one attaches to self-evidencing explanations.

Bibliography

Achinstein P., 'The Pragmatic Character of Explanation', in D. Ruben (ed.) *Explanation*, (Oxford University Press, 1993)

van Fraassen, B. C., 'Salmon on explanation', *The Journal of Philosophy*, 82 (1985), No. 11, 641

van Fraassen, B. C., *The Scientific Image*, (Clarendon Press, 1980)

Hempel C. G., *Aspects of Scientific Explanation and other essays in the Philosophy of Science*, (The Free Press, 1965)

Kitcher P., 'Explanatory Unification', in J. C. Pitt (ed.) Theories of Explanation, (Oxford University Press, 1988)

Kitcher P. and Salmon W. C., 'Van Fraassen on Explanation', in D. Ruben (ed.) *Explanation* (Oxford University Press, 1993)

Lewis D., 'Causal Explanation', in D. Ruben (ed.) *Explanation*, (Oxford University Press, 1993)

Lipton L., 'Contrastive Explanation', in D. Ruben (ed.) *Explanation*, (Oxford University Press, 1993)

Salmon, W. C., 'Scientific Explanation and the Causal Structure of the World', in D. Ruben (ed.) *Explanation* (Oxford University Press, 1993)

Salmon, W. C., Statistical Explanation and Statistical Relevance, (University of Pittsburgh Press, 1971)

British Journal of Undergraduate Philosophy, 1(2), March 2006

Is familial partiality any better than racism?

David Marlow
Lancaster University
davidmarlow@teacher.com

It would seem to be a fundamental requirement of morality that we treat people impartially, i.e. that we treat people equally unless there is some morally relevant difference which justifies different treatment. A university lecturer, for example, should give equal grades to students who perform equally; unequal grades would be justified only if there were some morally relevant reason for the difference. A relevant reason might be that a particular student had submitted a superior piece of work, whereas it would clearly be unjustified to a give a student a better mark if they were attractive, humorous, or wore nice clothes.[1]

This idea of impartiality, however, appears to be called into question when we consider the notion of personal relationships. Our relationships with our friends, lovers and family members are inherently partial. We not only give greater weight to the interests of such people, we also expect it in return. In this essay I examine this apparent conflict. We will see that if we wish to defend the notion of familial partiality, we are faced with the difficult task of demonstrating how this type of partialism differs from unacceptable partialisms such as racism. The approach I will suggest focuses on the nature of moral judgements. I will argue that for a particular directive to count as a moral judgement, it must be shown how it contributes to some overall view of how one's life should be lived for it to be worthwhile. Thus, for any type of partialism to be a plausible ethical stance, it must be shown how giving greater weight to the interests of those involved contributes to a fulfilled life.

[1] It is important to note at the outset that treating people equally does not necessarily mean treating them the same, it simply requires giving equal weight to each individual's interests. For example, a doctor is not required by the principle of equality to prescribe the same treatment to each of her patients, only to have equal regard for their interests.

Let us begin our discussion by examining the apparent conflict between the principle of equality on the one hand, and the notion of familial partiality on the other. Two initial strategies seem to suggest themselves. Firstly, we could argue that the requirements of impartial morality should take precedence and therefore partiality shown towards family members should not be morally permissible. Or secondly, we could hold that the notion of partiality towards family members is morally acceptable, and seek to show why such an exception can be made to justify preferential treatment in the case of familial relations but not in other areas such as relations between different races. I will look at each of these strategies in turn.[2]

The first strategy appears doomed from the outset. As John Cottingham has highlighted, the practical feasibility of impartialism is very much in doubt.[3] All of us give much greater weight to the interests of our family and loved ones and it is very difficult to see how any normal human being could set about dividing up their time and resources in such a way which ignored agent-relative categories such as 'mine' and 'ours'. But the fact that impartialism may prove to be difficult does not entail that it is therefore unwarranted. No one ever said that morality was easy!

Susan Wolf highlights that to be truly impartial one would have to be some kind of moral saint.[4] She not only highlights that such an endeavour would be extremely demanding, but claims that it would constitute a model of personal well-being toward which it would not be either rational or desirable to strive. This is a different claim from Cottingham's, however. Wolf is not simply claiming that impartialism is difficult but further, that it is actually morally unwarranted. Other writers would seem to agree. As Charles Fried highlights:

[2] There are, of course, other possible strategies. For an alternative approach, see John Kekes (1981) who divides morality in a personal and a social aspect, arguing that acceptable partialisms (like familism) fall into the former classification, while unacceptable partialisms (such as sexism and racism) fall into the latter.

[3] John Cottingham (1986) p357

[4] Susan Wolf (1982)

> *…surely it would be absurd to insist that if a man could, at no risk or cost to himself, save one of two persons in equal peril, and one of those in peril was, say, his wife, he must treat both equally, perhaps by flipping a coin.*[5]

And Bernard Williams has argued that in such cases, a rescuer's choice to save his wife is justified simply because it is *his* wife and any further appeal to a moral principle which would legitimise his choice is 'one thought too many'.[6] He accepts that deep personal attachments will, by their very nature, conflict with impartiality, but maintains that without them life would not be worth living:

> *…unless such things exist, there will not be enough substance or conviction in a man's life to compel his allegiance to life itself.*[7]

So it appears clear then that partiality towards family members must (at the very least) be morally permissible, and so our task here will be to provide a conclusive argument as to why it is morally acceptable to give greater weight to the interests of a member of one's own family but not acceptable to give greater weight to the interests of a member of one's own race, for example.

Before embarking on this discussion, however, we must first pause to define the scope within which our argument takes place. To say *without qualification* that it is morally permissible to give preferential treatment to a member of one's own family is simply not correct. There are many roles and responsibilities which place one under a *duty* to be impartial. For example, a doctor who, when treating her patients, gives greater weight to the interests of her family members is clearly not acting morally. Her role places on her an obligation to act impartially and thus she would be failing in her duty if she were to show favouritism to individual patients. To take into account such 'role-related obligations' let us define partialism as the thesis that:

---

[5] Charles Fried (1970) p227

[6] Bernard Williams (1981) p18. For an interesting analysis of Williams' argument see Alasdair MacIntyre (1983) and also Barbara Herman (1983) pp245-6

[7] Ibid. p18

> *…unless one is under a direct or indirect duty to be impartial, it is morally correct to favour one's own.*[8]

But this raises the problem of what exactly we mean by the phrase 'one's own'. It seems to imply that the individuals whose interests we may permissibly give greater weight to are not defined in terms of some morally relevant descriptive quality that they possess, but solely in terms of the relationship they have to the agent. So in Fried's example, a decision to save my wife simply because she is *my* wife is based solely upon a non-eliminable agent-relative element. This, however, does nothing to clarify the scope of the phrase in question. My wife counts as one of 'my own' (whose interests I may legitimately favour) because she is a member of my family. But if being a member of my family is a legitimate reason for my preferring their interests then what about being a member of my race? Accepting familial partiality seems to commit us to agreeing to all kinds of arbitrary and unfair examples of discrimination.

There are at least two possible ways in which the partialist could respond to this challenge. Firstly, they could attempt to formulate a blanket defence of all forms of partialism by claiming that as autonomous moral agents, we are within our rights to give preferential treatment to the members of any group we choose. The second approach would be to provide some way of differentiating between those partialisms which are acceptable (such as familism) from those which are not (such as racism). I will look at each strategy in turn.

The first approach seems unlikely to succeed. Indeed some commentators have suggested that it may be offensive to even discuss an approach which would justify, for example, racial partiality. I would argue, however, that it is possible to cite examples where at least *prima facie*, it appears permissible for an agent to favour the interests of a member of their own race. Let us consider the following situation. While walking down the street, an Afro-Caribbean man comes across two beggars, one Afro-Caribbean, one Caucasian. He has only one banknote, so he can only assist one of them, but although he can see

[8] John Cottingham (1986) p358

that the Caucasian is in most need, he wishes to aid the first beggar purely on the grounds that he is of his own race. Would such an act be morally permissible? Could one not argue, for example, that he is legitimately entitled to do what he wishes with his own money? After all, his act is supererogatory (he is not *obliged* to give to either) so arguably he does no wrong whoever he chooses to benefit.[9]

The above argument, which rests upon the issue of personal autonomy, does appear plausible. There is a strong presumption that the distribution of a person's own resources should be decided by the agent rather than imposed on them from outside. However it does not follow from this that one's choices should therefore be exempt from moral censure. While a benefactor may be *legally* entitled to donate their money to whatever causes they choose, no matter how frivolous or unworthy, from a moral point of view their actions are still open to criticism. As Cottingham concludes:

> *The upshot is that the autonomy argument, though creating a presumption in favour of people's being allowed to distribute their resources as they wish, is not strong enough to guarantee the partialist immunity from moral censure if his choices turn out to be based on arbitrary and capricious criteria.*[10]

The second approach, although more ambitious, seems to be a more correct path to follow. If some partialisms (e.g. familism) are to be defended, then our task will be to demonstrate how these acceptable partialisms differ from unacceptable partialisms such as racism.

One possible starting point for such an endeavour might be to examine the nature of moral judgements. A moral judgement cannot simply be an arbitrary prescription. If a particular directive is to count as a moral judgement, arguably it must be shown how it contributes to some overall

[9] This example introduces the much discussed gap between justification and motivation. A benefactor may or may not be *justified* in giving preferential treatment to a member of their own race, however this seems to be a separate matter from *why* they would wish to.
[10] Ibid. p363

view of a good life, i.e. the positive role which it plays in one's conception of how one's life should be lived if it is to be worthwhile. The authority for such a requirement can be seen in the works of Aristotle who made it clear that the object of ethics was *eudaimonia*, roughly translated as flourishing, or fulfilling one's potential. According to this theory, if partialism is to be a plausible ethical stance, it cannot be enough simply to state that certain types of agent-relative preference are either justified or permissible, it must be shown how giving greater weight to the interests of 'one's own' contributes to a fulfilled life. If we apply this requirement to the various forms of partialism in question, we should then be able to judge whether there are any relevant moral distinctions between them which would result in some being justified and others not.

Let us begin by considering familism. As we have already discussed, the principle is a simple one. It states that in deciding whether to give greater weight to the interests of A or B, it is morally permissible to assign a certain moral weight to the fact that A is a member of my family. But although this principle is based upon a non-eliminable reference to myself, it would be wrong to believe that familism is necessarily based upon selfish or narrowly self-interested motives. A parent who loves their child desires the child's happiness for their own sake – and in this sense the emotions involved are genuinely altruistic. One could argue that while all genuine love is altruistic it also inevitably comprises this non-eliminable agent-relative aspect. So, the reason that a parent shows partiality to their children is not that they possess some universalisable features which should merit special recognition; rather a parent gives extra weight to their children's interests precisely because they are *their* children. And clearly the same would also apply to other close family members. The crucial point is that in each case the partiality is exercised precisely because of the special relationship that the recipient has to the agent.

So it appears that familism cannot be seen simply as a case of arbitrary favouritism. The importance of genuine love within any conception of what it is to have a worthwhile life cannot be denied. The functioning of close familial relationships relies upon special concern being shown to family members not because of some universalisable features they may possess, but simply because they are agent-relative. If one were to give no extra weight to

the fact that a person was one's wife, father, or child, and instead simply assess their needs impartially (as a stranger might do) then that special connection which forms the foundation of familial love and friendship would be destroyed. Partiality to one's family seems to be an essential factor in one of the highest human goods and thus, to my mind, is justified on that basis.

Our next step then is to examine racial partiality to see whether it could also be justified by applying the life-plan argument. One would hope that such an approach would not sanction racism and at least one leading commentator has concluded that:

> *…there appears to be no remotely plausible case for arguing that it must find a place in all or most plausible blueprints for human welfare.*[11]

However, Cottingham also acknowledges that there are racists who would claim that the principle of favouritism towards members of their own race *is* part of their overall blueprint for the good life. While this may run contrary to many of our intuitions about what it is that makes a life worthwhile, we are not entitled to dismiss the argument on this basis. The 'argument from the life-plan' appears to leave open the possibility that two equally rational people may disagree about what makes a life fulfilling. Thus, if A's conception of *eudaimonia* involves a society based on racial fraternity and members of each race favouring 'their own', they may condemn moral universalism as unworkable and destabilising. B's life-plan, on the other hand, may see racial integration as essential to fulfilment and therefore extol the virtues of multiculturalism. There may be many different recipes for the good life. The life-plan doctrine does not, in itself, entail that a particular prescription cannot count as a moral judgement. It merely requires a proponent to demonstrate how and why it contributes to a fulfilled life. If we want to oppose a particular approach, we need to show convincingly why it doesn't.

Unfortunately, this appears more difficult that one might hope. In this case we cannot appeal to universalism. The racist's prescription that members of each race should favour the interests of their own is just as universalisable as

[11] Ibid. p370-371

the opposite approach. Neither can we defeat the racist by claiming that their position is arbitrary or capricious. A common objection to racism is that to favour a person's interests based on the colour of their skin is arbitrary and therefore unjustified: one could just as easily choose some other characteristic, such as height or hair colour. But I would argue that equating racism with the colour of a person's skin is an erroneous over-simplification. Racism has less to do with the visible morphological characteristics on the basis of which we make our informal classifications, and much more to do with the *culture* of different races. The racist could thus maintain that because the culture of each race is inherently linked with ethics and morality, the choice to favour one race over another is not necessarily an arbitrary choice.

Cottingham's solution is to appeal to empirical evidence. He claims that all of the evidence suggests that:

> *…abandoning racial and sexual partialities leads to richer, more fulfilling human relationships and institutions, an increase in respect for persons, a greater scope for self-development – in short, greater prospects for the achievement of eudaimonia.*[12]

However, such evidence is liable to be disputed by supporters of racial partiality and it appears unlikely that they could be convinced that they are in error. I would argue that productive discussions can only take place when interlocutors accept certain basic principles upon which the argument is based. If neither side can agree on such a fundamental principle as to what makes a life worth living, then I would suggest that the prospects for achieving some measure of resolution seem unlikely at best.

So it seems clear that there are obvious tensions between the principle of equality on the one hand and the notion of partiality towards one's family or race on the other. We have seen that if one wishes to defend morally acceptable partialities (such as familism) then the challenge is to find some way of distinguishing these from unacceptable partialities, such as racism. Unfortunately it appears that the argument from the life-plan cannot provide

[12] Ibid. p371

any way of resolving such a dispute. The fact that there may be many different recipes for the good life means that we cannot conclusively demonstrate that racial partiality is not a necessary factor for those who believe, for example, that racial fraternity is essential to human welfare. Thus we need to search for some other strategy for differentiating the forms of partiality, one which demonstrates conclusively why familism is morally acceptable but racial partiality is not. Such a task, however, may prove to be more difficult than we would like to believe.

Bibliography

John Cottingham: "Partiality, Favouritism and Morality", *Philosophical Quarterly* 36 (1986): 357-373

Charles Fried: *An Anatomy of Values,* Cambridge, Mass., 1970

Barbara Herman: "Integrity and Impartiality", *The Monist* 66 (1983): 233-50

John Kekes: "Morality and Impartiality", *American Philosophical Quarterly*, 18, no. 4 (1981): 295-303

Alasdair MacIntyre: "The Magic in the Pronoun 'My' ", *Ethics* 94 (1983): 113-25

Bernard Williams: "Persons, Character and Morality", in *Moral Luck: Philosophical Papers 1973-1980* (Cambridge, 1981), 18

Susan Wolf: "Moral Saints", *Journal of Philosophy* 79 (1982): 419-39

British Journal of Undergraduate Philosophy, 1(2), March 2006

How mythical is the 'myth of the given'?

Andrew Stephenson
Cardiff University
stephensonac@cf.ac.uk

This is a paper in epistemology (with some important overlaps in the philosophy of mind). I shall concentrate on the notion of immediate knowledge; that is, knowledge that need not be deduced or inferred, it need simply be there to be justified and useful, it is wholly independent, it is *given*. This notion of simply and surely *given* knowledge, of what shall henceforth be called 'the Given', is present in many theories of knowledge, it takes different forms, performs different functions and involves different nomenclature. The Given has been present as both *a priori* and *a posteriori* knowledge: in Descartes the Given is present as the clear and distinct idea of the cogito, which is *a priori*; in A. J. Ayers' sense-data theory the Given is present as the sense-datum, which is *a posteriori*; arguably in Kantian philosophy the Given is present as both the form and the matter of the manifold of sensations – *a priori* (the pure concepts and the pure intuitions of Space and Time) *and a posteriori* (the sensations or intuitions themselves). Yet, as we shall see, the Given always shares the same defining characteristics.

In his essay, *Empiricism and the Philosophy of Mind*, Wilfrid Sellars attempts to destroy what he sees as the 'Myth of the Given' with two separate arguments. First, he holds that it is a mistake to assimilate sensations with thoughts and thus to view sensing as knowing. He identifies two ways of utilizing the verb structure 'to know' in sense-data theories, and states that there is only knowledge of facts, not of particulars. Second, Sellars makes the original contribution of viewing concepts as abilities. For him, all thought and knowledge is conceptual and therefore cannot be simply given. In order to show this he shall use 'a myth to kill a myth.'[1] With what he calls the

[1] Sellars, p. 117.

'Myth of Jones' he questions the position of acts of self-awareness (the recognition of thought processes, beliefs and knowledge), claiming that they are not a *foundation* of consciousness, but a *consequence*. There are two readings of Sellars' Myth of Jones: a literal reading which gives us an historical account of knowledge and consciousness; and an allegorical reading that parallels the intellectual progression of each human. Common to both readings is the idea that there is a pre-theoretical stage at which thoughts and sensations are not only unidentifiable or un-named, but do not occur. Just as the human species had such a pre-theoretical stage, so does every infant. Or so the story goes. I shall divide Sellars' argument in two ways: by separating his criticism of the Given from the alternative programme of naturalistic behaviourism that he proposes; and by separating his criticism of the *a priori* form of the Given from that of the *a posteriori* form of the Given.

But first, I shall much more clearly define what is meant by the Given.

The Given is most readily defined as owning four primary characteristics, only two of which Sellars focuses on:

(1.i) The Given is a form of knowledge

So although we shall see that Sellars rejects both foundationalism and coherentism, the two dominant doctrines of counter-sceptical epistemology, he accepts that there can be forms of knowledge. He is a critic, not a sceptic.

(1.ii) The Given is a form of non-inferential knowledge

This could traditionally be seen as merely a necessary part of (2) below, if not completely sufficient for its explication. However, it is necessary to include (1.ii) as a distinct characteristic because Sellars accepts the possibility of self-intimating knowledge – in the sense that it can be had directly and non-inferentially – whereas he denies (2). At (1.ii), although the knowledge is immediate and non-inferential there are still, for Sellars, concepts in the background that enable us to have this knowledge of objects directly. For example, we can be aware that something is green non-inferentially simply by having a sensation of it, but only if we already have the concept of green (and

such is the case for most of our sense impressions). The manner in which this is possible shall be explained below. Nevertheless, in accepting (1.ii) Sellars is rejecting coherentism.

(2) The Given does not presuppose any other knowledge

(3) The Given's mere presence in the mind is sufficient for knowledge of it

At first review (2) and (3) may seem to be coextensive. However, (3) considered without (2) does not deny the possibility that this independent knowledge was initially acquired through other knowledge which may now be inconsequential. The independent knowledge, once acquired, can stand on its own. Therefore (2) is also a necessary condition of the Given as it asserts not simply that other knowledge is not needed for the present awareness of the independent knowledge, but also that it was in no way instrumental in initially acquiring the independent knowledge. To clarify, although knowledge is presently independent, this may not always have been the case. Therefore, in restricting this possibility, (2) adds something to a definition of the Given.

Indeed, (3) could now seem only a supplement to (2) because now it merely adds that we can have knowledge of the Given, a qualification that is arguably already implied by (1.i). However, attempts to amalgamate (2) and (3) are unsuccessful because (3) also adds something that the other characteristics do not – it denies any need for previous concepts in knowing the Given, or indeed the need of any thought process. (2) restricts the necessity of knowledge but accepts the possible necessity for mental faculties which may not strictly count as knowledge (concepts are here seen as abilities and not forms of knowledge themselves), a possibility that (3) denies. Thus we see that (2) and (3) are both separately necessary for the Given to be defined because *knowledge of the Given is not epistemically mediated in any way*. We do not need to have concepts to know the Given, so Sellars calls the Given a myth.

By denying the existence of the Given, by calling it a myth, Sellars is rejecting foundationalism. Although we can conceive of a form of knowledge that stands independently from other knowledge in that it neither references nor is the reference point of any other knowledge – it needs no justification nor gives any justification – it might be worth calling this knowledge *a* given, but it wouldn't play the role that has traditionally been assigned to *the* given. Consequently a belief in *a* Given does not necessarily entail foundationalism, but a theory of foundationalism necessarily entails a belief in *the* Given, and so a rejection of it is a rejection of foundationalism. This necessity for *the* Given to be the foundation, to be epistemically efficacious, will become more important below. (We shall see that Sellars denies that nonpropositional items can in any way justify beliefs and this is fundamental in his rejection of sense-data theories.) In understanding that *the* Given must be useful in justifying and acquiring other knowledge, we have its fourth and final defining characteristic:

(4) The Given must be epistemically efficacious

Sellars' initial disagreement with the empiricist epistemologies of Locke and Hume is concerned with their view of concept acquisition. The empiricist's idea is that we have a plethora (or manifold) of sensations that we are confronted with upon contact with the world. From these we gain more general concepts by virtue of having 'an innate ability to be aware of certain determinable sorts – *indeed,…we are aware of them simply by virtue of having sensations and images*.'[2] Sellars thinks that this idea is founded on a mistaken assimilation of sensations with thoughts. For Sellars thoughts are constituted of concepts and as such they can provide reasons for knowledge claims and can equally be the result of reasoning. Both Sellars and the empiricists agree that thinking can lead to knowing. However, the empiricists held that experiencing sensations can also lead to knowing, and therefore serve the same role as thoughts. Sellars maintains that this is a technical use of the term, merely denoting "knowledge" of particulars not facts, and is therefore not propositional. Remember, for Sellars propositionality is a necessary qualification of knowledge. This is inextricably bound with his linguistic

[2] Sellars, p. 62.

turn, which we shall come to, but briefly: 'knowledge' is 'knowing that' and 'knowing that' requires language. For Sellars sensations are mere feelings and cannot be reasons, on their own, for anything – they cannot independently give rise to thoughts or knowledge. The empiricists' misconception of knowledge is a result of the sensations/thoughts confusion and leads to the impossible dependence on the Given in sense-data theories. This dependence is impossible because sense-data are non-propositional (shown through more clearly distinguishing between sensations and thoughts) and therefore cannot serve as *the* Given, which must be epistemically efficacious. For Sellars non-propositional items cannot be *the* Given because that which is non-propositional (x) cannot be a reason for (y), or serve as a premise in an argument for (y), so is epistemically inefficacious.

Sellars' idea of concept acquisition is very different. It is here that he moves away from the traditional empiricist view of concepts as 'things' or 'ideas' that someone may be in possession of, to a view of concepts as abilities. In this sense Sellars' Myth of Jones can be read allegorically and is seen to parallel human growth. For example, a child will learn through being taught to say 'red' when there is a red thing there, and thus will learn to distinguish. But at this stage the child is merely repeating what it has been taught, albeit with an increasing level of ability. The ability will subsequently be acquired to conceive the incompatibility of, for example, red with green. Only now will a 'battery' of concepts (as abilities) gradually and inextricably be acquired, and conformation to the received semantic language structure will be attained. This is how Sellars allows for non-inferential knowledge while still presupposing other knowledge. He is criticizing any *a posteriori* form of the Given in his deconstruction of sense-data theory, and stopping short of any commitment to any *a priori* form of the Given in his reconstruction of a concept acquisition theory. Sellars denies (3), that knowledge can be *merely present* in the mind and blames this mistake on the traditional empiricist metaphor of sense impressions being 'pictures' in the mind. He also denies (2) in claiming that non-inferentially known propositions cannot be epistemically independent, because for them to be justified of a subject they must first be a reliable response in normal empirical conditions (they must have acquired reliable differential dispositions); and second the subject must

know that this is the case (this requirement of knowledge of normal empirical conditions renders such propositions dependent on other knowledge).

An externalist argument that might be formulated against Sellars might claim that such an early condition of merely being able to reliably distinguish is indeed adequate for knowledge, because externalism places no value on a subject's grasp of justificatory conditions. Yet this argument is based on the assumption that accounts of non-inferential knowledge, such as Sellars', must accord with the view that *knowing about* one's reliability is not required for 'first order' knowledge. But Sellars asserts conversely that this self-awareness *is* required for first order knowledge. While rejecting the above premises of externalism, Sellars does fully admit that an external connection is fundamental to the justification of non-inferential knowledge, which is clearly contrary to strict internalism. Again, as with foundationalism and coherentism, Sellars straddles the traditional division.

Through his destruction of the Given, Sellars formulates a new view that he calls psychological nominalism. Psychological nominalism specifically denies the doctrine that non-linguistical, non-conceptual awareness of particulars is the foundation of knowledge. In constructing a view 'according to which…all awareness even of particulars…is a linguistic affair'[3] Sellars is asserting that to be aware of particulars – numbers, people, chairs, sense-data sets and physical objects – a language structure has to be in place already, because facts constitute particulars, not vice versa.

Now we must place the above, allegorical reading of the Myth of Jones in conjunction with the more literal, historical reading of it. Here Sellars claims that self-awareness of cognition is one of the last acquired abilities of the human species' consciousness. Furthermore, he claims the same status, if not the same function, for sensations. Thoughts and sensations are thus theoretical entities based on observational entities. Sellars postulates the existence of human ('Rylean') ancestors, who are first able to distinguish, then to perceive as incompatible, and then gradually use *concepts* (which are not acquired singularly). Although after these three stages they have a semantic

[3] Sellars, p.63.

vocabulary, they as yet have no language concerning thoughts. But then a genius – Jones – appears and proposes a set of theoretical entities: thoughts. The language is then developed to enable them to talk about, and thus to conceive of, one's own and others' thoughts; the ability to construct the term 'I think' is attained. Thus Sellars postulates as to how our thoughts, as theoretical entities, were initially based on observational entities: behaviour patterns**.** Having rejected the traditional picture, he has offered an alternative.

That unobservables are mere theoreticals based on observables may initially seem a problematic claim, with the possible result of philosophical behaviourism. However, as Sellars is a scientific realist he is simultaneously claiming that thoughts do essential work in explanation and, at least to all intents and purposes, exist. He is at worst ambivalent with regard to ultimate ontology. The problem with this analogy with science is that theories in science are in a state of constant flux; meanings are changed and change each other. If this is the case in this epistemological context then thoughts as theoretical entities can only be justified by their coherence. In this sense it seems that Sellars is still, in some way, a coherentist.

But coherentism can have some strongly counterintuitive consequences if pushed to an extreme. For example, the Quine of *Two Dogmas of Empiricism* famously decides that, 'in point of epistemological footing the physical objects and the gods [of Homer] differ only in degree and not in kind.'[4] Here Quine sees no limit to which theoretical entities can be revised to remain within the interconnected web of knowledge. However, this particular manifestation of Quine's thought can indeed be seen as extreme, and elsewhere he is in agreement with Sellars on many points. Another result of Quine's view is a shift towards pragmatism but again Sellars moderates this doctrine by dichotomizing the image of the mind into the scientific (an embodied being subject to the study of science) and the manifest (a being with beliefs, desires and intentions). Thus Sellars, whilst holding in the Myth of Jones that thoughts are theoretical entities, does not succumb to relativism or revisionism, ideas which could have otherwise been the outcome of such a seemingly historical account.

[4] Quine, p. 44.

These accusations of coherentism and the externalist arguments are not arguments against Sellars' critique of the Given, but against his resulting theory of the mind. Moreover, they are arguments against his view of sensations, the function of which is '*built on* and *presupposes* their role in inter-subjective discourse,'[5] and they are arguments against his view of concepts and thoughts, the function of which is '*built on* and *presupposes* their inter-subjective status.'[6] Hence a new question can be formulated thus: Is Sellars' critique of the Given inextricable from his Myth of Jones? Triplett and deVries argue that the Myth of Jones cannot be simply a thought experiment, nor mere allegory, as its logical possibility is not enough to warrant Sellars' many resultant ideas. However, it seems to me that the essay *Empiricism and the Philosophy of Mind* separates neatly into two linked but distinct parts, a claim that can be clarified if we return to the bifurcation presented in the introduction, between the *a priori* and *a posteriori* forms of the Given. The arguments for the rejection of the *a posteriori* form of the Given in the first part stand on their own and are extremely effective and convincing. Here, Sellars has marked a turning point in analytic philosophy. It is only the *a priori* form of the Given whose rejection is based on the Myth of Jones. Furthermore, Sellars' new theory still holds some aspects, simply revised and qualified, of the traditional empiricist view. One major defence of the Given – that we do seem to have unique and privileged access to our own thoughts – cannot be held to contradict Sellars. Observation statements (Sellars' construal of sensations) and thoughts play a reporting role. This is a notion which simply reinterprets the meaning of privacy whilst retaining a certain amount of privilege in line with the tradition. Sellars' criticisms of the Myth of the Given remain incredibly persuasive, even if his Myth of Jones can itself be criticized.

I would like to thank my anonymous assessor…for more than just the title of this paper.

[5] Sellars, p. 115.
[6] Sellars, p. 107.

Bibliography

Garfield, Jay L., 'The Myth of Jones and the Mirror of Nature: Reflections on Introspection', *Philosophy and Phenomenological Research*, 50, 1 (1989), 1-26

Johnsen, Bredo L., 'The Given', *Philosophy and Phenomenological Research*, 46, 4 (1986), 597-613

Kukla, Rebecca, 'Myth, Memory and Misrecognition in Sellars' "Empiricism and the Philosophy of Mind"', *Philosophical Studies*, 101 (2-3), 2000, 161-211

Quine, W. V. O. *From a Logical Point of View* (Harvard: Harvard University Press, 1961)

Sellars, Wilfrid, *Empiricism and the Philosophy of Mind* (Cambridge: Harvard University Press, 1997)

Sellars, Wilfrid, *Science, Perception and Reality* (London: Routledge, 1963)

deVries, William A. and Triplett, Timm, *Knowledge, Mind, and the Given* (Indianapolis: Hackett, 2000)

Williams, Michael, *Problems of Knowledge: a critical introduction to epistemology* (Oxford: Oxford University Press, 2001)

British Journal of Undergraduate Philosophy, 1(2), March 2006

Self-overcoming and free will in Nietzsche

Ryan Dawson
Selwyn College, Cambridge
rd286@cam.ac.uk

Introduction

Nietzsche is often referred to as the philosopher of 'self-overcoming'. *Zarathustra* is littered with imperatives to overcome oneself – to remake one's character. As Zarathustra says, 'ready must thou be to burn thyself in thine own flame; how couldst thou become new if thou have not first become ashes?'[1] For Nietzsche, overcoming is to be achieved by the exertion of a strong will. So it must surprise readers to find that Nietzsche denies that human beings have freewill – the conscious experience of freewill is, for Nietzsche, an illusion.

The question, then, is how can self-overcoming be achieved without freewill? How do we will ourselves anew without a conscious will? This paradox is, in my view, too great for us to accept that Nietzsche simply contradicts himself. Its resolution is a serious interpretative problem.

In this essay, I shall begin by explaining the approaches to this question taken by two interpreters – Alexander Nehamas and Brian Leiter. Nehamas argues that the solution lies in a conception of a life as a set of actions in which later actions refer back to earlier ones in a way that resolves earlier conflicts. The acceptance of fate, on Nehamas' interpretation, is the creation of permanence in character. Leiter argues that we must accept that fatalism is the dominant theme in Nietzsche. Once we understand this, he says, we should accept that

[1] Nietzsche, *Thus Spake Zarathustra*, Thomas Common (trans.) (Herfordshire: Wordsworth Classics, 1997), p. 60

Nietzsche has an unusual conception of self-overcoming – a conception that presupposes fatalism.

In the final section I present an interpretation that assimilates and surpasses those of Nehamas and Leiter. I argue that each is grasping one level of a two-level picture. Hence my solution centres on the claim that Nietzsche has a two-level view of the self. These levels are best captured by the terms used in *Zarathustra*. There is an internal self as a set of drives located in the body, denoted by '*das Selbst*.'[2] There is also the self that is an external, social creation, most commonly referred to by Nietzsche as '*die Seele*.'[3] *Die Seele* is a creation of *das Selbst* – it arises out of the need for self-expression in a social setting. Leiter's account is adequate only for *das Selbst*. Nehamas gives an account that is useful only for *die Seele*. Understanding the interaction of these two levels is crucial to resolving the paradox.

Nehamas and self-overcoming

Nehamas sets out to interpret the phrase 'how one becomes what one is' (the subtitle to *Ecce Homo*) in his paper of the same name. In this section I shall explain his answer and how it bears on the paradox explained above.

Nehamas sees it as important that the concept of 'becoming' is used. He connects this with Nietzsche's claim that the world is becoming. For Nietzsche, the distinction between appearance and reality is a fiction. The distinction was created in a vain attempt by men to give substance to the ego. The reification of the ego is a desperate grasp at finding permanence ('being') in a world of constantly changing relations ('becoming'). The eternal recurrence figures as a psychological tool – it is a way for us to make our characters approximate to permanence without trying to escape the world of becoming. To see one's character as awash in becoming – merely a

[2] Most commonly translated as 'Self'.

[3] This is usually translated as 'soul'. Nietzsche sometimes uses the word in the Christian sense but this is not the use being discussed. The dominant use of '*Seele*' in Nietzsche is something close to 'character'. He deliberately hijacks the Christian word and puts it to a more 'worldly' use.

disconnected set of events – is to fall into nihilism. The recurrence gives us an approximation to being by forcing us to affirm our own lives.

To use the recurrence, we imagine that every action and every thing in the universe will repeat forever. Nietzsche asks us to do so, Nehamas tells us, not because this is a cosmological truth but because we need to affirm life in its entirety. By adopting a belief in recurrence we come to accept and affirm our past actions. By accepting the recurrence we become more responsible – every action is tied to our characters. In this way, one's character becomes a work of art – it acquires a greater permanence. Thus one's character is the closest approximation to being in a world of becoming. 'Becoming what one is' is a matter of affirming the actions that are already there – taking a disparate set of actions and making out of them a unified character.

Is this not in conflict with Nietzsche's denial of the will? Not according to how Nehamas interprets this denial. Nehamas interprets Nietzsche's attack as being an application of the doctrine of the will to power. The will to power, Nehamas claims, says that everything in the world is nothing apart from the sum of its effects on other things. The world is a big set of relations between things – these things being nothing other than props for relations. We might say that Nehamas attributes to Nietzsche a structuralist view of the world. It follows from this that there is no permanent subject behind actions.

I do not have space in this essay to assess whether Nehamas' view of the will to power is correct. But I agree with him that Nietzsche denies any Cartesian notion of the self. I will return to this towards the end of this essay.

So there is no permanent self behind our actions, creating them. Yet we must aspire to be creators of ourselves. How can we do so? By giving unity to our actions. Nehamas refers to Nietzsche's view of the self as 'soul as social structure of the drives and affects.'[4] The great men of history embrace the concerns of their age and find ways to unify them within their lives. They resolve apparent social contradictions and thus become motors for social

[4] Section 12 in Nietzsche, *Beyond Good and Evil,* in *Basic Writings of Nietzsche*, Kaufmann (trans.) (Modern Library, 1968)

change. This is how Nietzsche praises his great heroes – Goethe, Beethoven, Wagner, Nietzsche himself. In a model life, we will see that later actions refer back to earlier ones. The later actions will try to address problems posed by earlier actions. We will expect to see disparate and contradictory actions in early life that are brought together by later actions that, through their social import, resolve the earlier contradictions and give unity to the set of actions. Thus a character is formed. A set of actions becomes something that demands interpretation.

Perhaps the easiest way to see this is to look at Nietzsche himself. His books bring together disparate influences. They are rich with apparent inconsistencies that demand interpretation from the reader. *Ecce Homo* can be interpreted as Nietzsche's own attempt to bring the books together – to bring out the unified whole.

Nehamas presents a liberating vision of the Nietzschean self. He sees Nietzsche's self as an attack on 'antecedently set possibilities' for individuals.[5] He also sees it as embodying the truth that the most important thing that individuals can do is to construct a narrative for themselves. Nehamas conjectures that Nietzsche takes literary characters as the model for his ideal person. Literary characters are, after all, nothing more than the roles that they play in the story. This is why Nietzsche's biography centres on his books – his philosophy is an attempt to bring life and literature together. As Nehamas says, 'no one has brought literature closer to life than Nietzsche.'[6]

Let us return to our original question – 'how can self-overcoming be achieved without a willing subject?' The answer, as Nehamas sees it, should now be clear. Self-overcoming is not a matter of remaking a conscious, willing self that lies behind actions. It is a matter of having future actions refer back to, and resolve contradictions within, past actions. The man who has created himself is the man who has embodied and furthered his age.

[5] Nehamas, Alexander, 'How One Becomes What One Is' in *Nietzsche*, Richardson and Leiter (eds.) (Oxford University Press, 2001), p. 261

[6] Leiter, Brian, 'The Paradox of Fatalism and Self-creation in Nietzsche' in *Nietzsche*, Richardson and Leiter (eds.) (Oxford University Press, 2001), p. 280

Leiter and fatalism

The problem takes on a very different character for Leiter. The denial of a substantial ego, for him, is the claim that the only causal forces in the mind are unconscious. Leiter heavily stresses, against Nehamas, that there are firmly set antecedent possibilities for individuals. We are bound by our physiologies. The problem that Leiter addresses in his 'The Paradox of Fatalism and Self-creating in Nietzsche' is how any kind of self-creation is possible given that we are bound by physiological facts.

Leiter captures the importance of Nietzsche's use of physiological language. Individuals, for Nietzsche, fall under physiological types. Nietzsche likes to explain people's beliefs in terms of their moral inclinations and their moral inclinations in terms of their physiological type. ('Assuming that one is a person, one necessarily has the philosophy that belongs to that person.'[7]) This materialist trend in Nietzsche, according to Leiter, comes from his age. Materialism was *en vogue* in Germany at the time. Schopenhauer, despite his metaphysical theory of the will, was also a committed materialist and was no doubt influential in Nietzsche's thinking about materialism.

Leiter formulates two conditions to articulate Nietzsche's position on what is necessary for self-creation:

Causal Condition: a person must be a necessary cause of what he becomes.

Autonomy Condition: the person's creation of self must be free. They must be more than a conduit for larger causal processes.

Leiter attributes to Nietzsche a fatalist position that accepts the first condition but not the second. He explains how Nietzsche's fatalism is entailed by Nietzsche's theory of the will.

[7] *The Gay Science*, Kaufmann (trans.) (New York: Vintage Books, 1974), Preface 2

Leiter correctly argues that Nietzsche views the conscious will as epiphenomenal. Nietzsche has two arguments for this. One is that 'a thought comes when it wishes, not when I wish.'[8] Phenomenologically, we do not control the course of consciousness. The second is that every action is unknowable. Ultimately, we cannot explain why actions are performed by conscious processes. Motives are simply after-the-fact fictions. Physiological explanations must replace motive-explanations.[9]

So how can self-creation fit into this? Leiter sees Nietzsche as having a largely political view of the self. The self is made up of various drives. One drive may be in control at one point, another may then take over – drives constantly vie for control of the body. These drives are present to consciousness. But consciousness does not control any of them. Consciousness is simply an epiphenomenal bystander.

Self-mastery is brought about by 'S-procedures'. This is a term that Leiter borrows from Galen Strawson.[10] S-procedures are ways to shape our behaviour by affecting the unconscious. Nietzsche's exhortations to a change in values can thus be read as exhortations to use S-procedures. This should not be seen as consciousness using strategies to influence the unconscious. The S-procedures can themselves only come about because of natural facts. It may feel like consciousness exerting control but S-procedures are, at a causal level, just unconscious processes controlling other unconscious processes.

Leiter quotes a passage in *Daybreak* which convincingly supports his claims about how Nietzsche views self-mastery.[11] Clearly, the autonomy condition is not satisfied in self-mastery. So, Leiter rightly infers, the autonomy condition is not satisfied in self-creation either. As I intend to show, Leiter wrongly concludes from this that Nietzsche's fatalism is dominant over his emphasis on self-creation.

---

[8] *Beyond Good and Evil* 17

[9] Knobe and Leiter, 'The Case for Nietzschean Moral Psychology', forthcoming in Leiter and Sinhababu (eds.) *Nietzsche and Morality* (Oxford University Press, 2007)

[10] Strawson, 'The Impossibility of Moral Responsibility', *Philosophical Studies*, 75 (1994), p. 5-24

[11] Richardson and Leiter, p. 318

The self-soul model

My aim in this section will be to give an interpretation of Nietzsche's view of the self that I have drawn primarily from *Zarathustra* and in particular 'The Despisers of the Body'. This interpretation depends upon a distinction between '*Selbst*', '*Seele*' and '*Ich*'. I will argue that there is no point of conflict between Nehamas and Leiter – both are incomplete because they have focused upon different levels of the self, expressed in *Zarathustra* by '*Selbst*' (Leiter's focus) and '*Seele*' (Nehamas' focus). For clarity I shall henceforth use 'Self' for '*Selbst*', 'soul' for '*Seele*' and 'ego' for '*Ich*'.[12]

The Self is a collection of drives vying for control of the body. I think it is well-captured by the political model of the self that is the focus of Leiter. This is revealed in 'The Despisers of the Body' when Nietzsche writes that:

> *behind thy thoughts and feelings, my brother, there is a mighty lord, an unknown sage – it is called the Self; it dwelleth in thy body, it is thy body.*[13]

The ego discussed by Leiter also features here. It is discussed as an epiphenomenal creation.[14] But what Leiter neglects, or simply lumps together with one of the above, is soul. Almost all of Nietzsche's rhetoric of self-overcoming in *Zarathustra* employs either 'soul' or the equivalent term 'spirit'.[15] 'Self', however, is barely mentioned outside of 'The Despisers of the Body'. I contend that there is a reason for this. 'Soul' is a public notion and is

[12] I believe these concepts are kept distinct in works other than *Zarathustra*. Although it is only in 'Richard Wagner in Bayreuth' that I can find an explicit use of the same *terms*. We read that Wagner 'remained loyal to his higher self, which demanded of him deeds in which his many-faceted nature participated as a whole and bade him suffer and learn so as to be capable of these deeds.' Found at end of III. 'Wagner in Bayreuth' in *Untimely Meditations*, Hollingdale (trans.) (Cambridge University Press, 1983)

[13] *Thus Spake Zarathustra*, p. 30

[14] Ibid., p. 30

[15] It should be noted that Nietzsche tends to use '*Seele*' and '*Geist*' equivalently. I do not mean that '*Seele*' is simply translated as 'spirit'. The use of two equivalent terms is in the original.

thus dealt with more commonly. Soul is the outward manifestation of the Self. It is Self's creation and expression.

> *The creating Self created for itself esteeming and despising, it created for itself joy and woe. The created body created for itself spirit, as a hand to its will.*[16]

The creating one creates his soul through his actions. The drives of the Self are expressed in actions. Nietzsche, however, rarely expresses this directly. He more commonly expresses it through two of his favourite metaphors. One is that he sees the Self as *pregnant* (German '*schwanger*') – the Self is mother and the soul is child. The other metaphor is the soul as a product of *work*. It is a creation of an artist. Both of these metaphors are employed together in this passage from 'The Higher Man':

> *In your self-seeking, you creators, there is the foresight and forseeing of the pregnant! What no one's eyes hath yet seen – namely, the fruit – this sheltereth and saveth and nourisheth your entire love.*
>
> *Where your entire love is – namely, with your child – there is also your entire virtue! Your work, your will is your 'neighbour': let no false values impose upon you!*[17]

We can act so as to express different sets of values. This is expressed in 'The Virtuous':

> *My friends! That your very Self be in your action, as the mother is in the child: let that be your formula of virtue!*[18]

Self-creation consists in finding a set of values that expresses the Self. Nietzsche also desires that these values be original. But in our originality we

[16] Ibid., p. 31
[17] Ibid., p. 281. It is noted that the German passage does not contain '*Selbst*'. 'Self-seeking' is Common's translation of *eigennutz,* a better translation would be 'self-interest'. Nonetheless, I maintain that the concept of the Self is operating in the background.
[18] Ibid., p. 93

must not adopt values that are beyond the limits of the Self. It is not authentic for a person short of attention to value careful study. This leads to self-deception and what Nietzsche would call an 'unhealthy' soul.

Further light is shed on the Self's creation of the soul by aphorism 19 in *Beyond Good and Evil*. Here the body is referred to as a 'commonwealth' composed of many 'under-souls'. Only some of these 'under-souls', or drives, are expressed in action. When a drive comes to expression there is a phenomenal feeling of identification with it.

> *What happens here is what happens in every well-constructed and happy commonwealth; namely, the governing class identifies itself with the success of the commonwealth.*[19]

This identification in willing is primarily discussed in connection with the ego. This is clearly distinguished from the soul. The ego is a phenomenal experience – it has no public presence or causal powers. But the identification of specific drives with the whole Self fits the model of drives being expressed in a soul that is constructed through action. For it is in action that the identification occurs. The existence and identifying function of the ego may be causally connected with the existence of a public soul.

The attack on freedom of the will, then, is an attack on the reification of the ego. Nietzsche stresses that it is not the ego that gives rise to the soul – it is the Self that creates the soul. But there is also a positive sense of freedom in Nietzsche. This freedom is spoken of in *Zarathustra* only in connection with the soul. The freedom of the soul is freedom from the values of the many. This is why a free life is only possible for 'great souls'[20]. But how is this achieved?

Interpretation is a key concept. To be free we must act in ways that challenge social conventions. Our actions will not admit of conventional interpretation, or will seem very poor by conventional judgment. Thus they will demand

[19] *Beyond Good and Evil* 19, p. 216
[20] *Zarathustra*, p. 47

unconventional interpretation. Challenging conventions in this way is a high expression of the will to power. I am taking control of the ways in which others look at me and understand me. It is in this sense that I am free.

The Self must interpret its own actions. In doing so, the Self creates an understanding of its soul. The art of self-creation lies in interpreting past actions in radical ways – ways that accept and affirm these actions in accordance with the eternal recurrence. Our new actions will then be in accordance with our radical scheme of interpretation. These actions can thus be used to depart so radically from conventional schemes as to force others to think further about our actions and evaluate our interpretative scheme.[21]

Nehamas' emphasis on the interpretation of past actions and their unification through the eternal recurrence captures some of the importance for Nietzsche of self-understanding. What Nehamas does not explicitly state, and cannot do without a distinction between Self and soul, is that we change ourselves by changing our *self-image*. It is a striking psychological fact that we can change our characters by finding a new view of ourselves. This can only be explained when we have a distinction between an external, public soul and an internal, private Self. The self-image is the view that the Self forms of itself, based upon its own construction – its soul. This picture also explains the importance of solitude for Nietzsche. It is in solitude that we cast off the shackles of conventional understanding and contemplate our past actions in a new light.

We now have an answer to Leiter's claim that Nietzsche has an odd sense of 'self-creation'. I understand him to mean that the sense is odd because there is no way to see both 'author' and 'work'. For Leiter, all we have is the Self that reshapes itself through S-procedures. He is not wrong to say that the Self reshapes through changing values, and from a certain perspective it can be helpful to talk of this reshaping in terms of S-procedures. But he misses that the Self creates a soul as a manifestation of itself. The Self is the author and the soul is the work. In a sense, Leiter is not wrong as the soul is nothing over

[21] Here I am regarding interpretation as inextricable from valuation. To interpret an action positively is to see it as exemplifying some positive value.

and above the Self. But it is vital that the soul is the Self's only means of public expression. Leiter, rather reductively, sees only a Self remaking itself because he considers the situation only from an internal perspective. When also seen from the outside, we see a Self and a soul that are constantly being remade in response to one another.

So Leiter is wrong that we must see Nietzsche's fatalism as dominant over his emphasis on self-overcoming. The two are not in genuine conflict.

What of the point that Nehamas emphasises about there being no 'doer' behind the deed? It seems as though the Self would be just such a doer. Is this an objection to the Self-soul model?

My reply is that it is the 'popular mind' that doubles the deed. I interpret section 13 of essay 1 of the *Genealogy* as an attack on the popular fiction of the ego. The ego is posited as an all-powerful force behind action. This fiction is useful for the weak-willed – it means that they do not have to create themselves. The reality behind action, though not mentioned in I.13 of the *Genealogy*, is the Self. Why is this not mentioned? Because Nietzsche is describing how the popular mind adds a fiction behind deeds to make the deeds seem more intelligible. The Self does not make deeds more intelligible. We can only glean a vague understanding of the Self from its actions. The Self is of use to the psychologist – not the popular mind. There is no conflict with the remark that 'there is no "being" behind doing, effecting, becoming'. The Self remains within the realm of becoming. It is a multiplicity of conflicting drives. The fictive 'doer' of the popular mind belongs to a non-existent realm of being.

There remains a further important issue to be resolved. Does this position allow us to agree with Nehamas that Nietzsche attacks antecedent possibilities for persons? And also with Leiter that persons are fundamentally limited by their biology?

I don't think these two claims are fully compatible – we have to meet them half-way. We can agree with Nehamas insofar as the creating one is able to carve apart the realm of social possibility. The creating ones depart from the

cultural logic and Nietzsche's heroes (Beethoven, Goethe, Nietzsche himself) even change that logic. Nietzsche can be seen as attacking antecedent social possibilities insofar as he sees such possibilities as always susceptible of being remade.

But there is also a sense in which Leiter is right. For only Nietzsche's nobles are capable of interpreting themselves radically and departing from the current sphere of social possibility. Nietzsche's herd can only follow pre-existing paths – it is a type-fact about them that they lack the ability to defy social conventions. Further, there are limits on how even particular nobles are able to remake the cultural logic. If the noble's soul is to be healthy then he must create and follow rules that he is capable of following.

Conclusion

Nietzsche's attack on antecedent possibilities is to be interpreted as a celebration of human individuality. We need not live by a moral code that does not fit us – we can make our own. Nietzsche celebrates individuality *above* morality. Nietzschean freedom is both negative, in that it is freedom from the values of others, and positive, in that it is freedom to create an authentic soul. Nietzschean freedom is the authentic interaction of the Self and the soul.

Special thanks to Manuel Dries and to Nicholas Cunild.

Bibliography

Knobe and Leiter, 'The Case for Nietzschean Moral Psychology', forthcoming in Leiter and Sinhababu (eds.) *Nietzsche and Morality* (Oxford University Press, 2007)

Leiter, Brian, 'The Paradox of Fatalism and Self-creation in Nietzsche' in *Nietzsche*, Richardson and Leiter (eds.) (Oxford University Press, 2001)

——, *Nietzsche on Morality* (London: Routledge, 2002)

Nehamas, Alexander, 'How One Becomes What One Is' in *Nietzsche*, Richardson and Leiter (eds.) (Oxford University Press, 2001)

——, *Life as Literature* (Cambridge: Harvard University Press, 1985)

Nietzsche, Friedrich, 'On Truth and Lying in a Non-moral Sense', in *The Birth of Tragedy And Other Writings,* Guess and Speirs (eds.) (Cambridge University Press, 1999)

——, 'Wagner in Bayreuth' in *Untimely Meditations*, Hollingdale (trans.) (Cambridge University Press, 1983)

——, *The Gay Science,* Kaufmann (trans.) (New York: Vintage Books, 1974)

——, *Thus Spake Zarathustra*, Thomas Common (trans.) (Herfordshire: Wordsworth Classics, 1997)

——, *Beyond Good and Evil,* in *Basic Writings of Nietzsche*, Kaufmann (trans.) (Modern Library, 1968)

——, *Genealogy of Morals*, in *Basic Writings of Nietzsche*, Kaufmann (trans.) (Modern Library, 1968)

——, *Ecce Homo,* in *Basic Writings of Nietzsche*, Kaufmann (trans.) (New York: Modern Library, 1968)

——, *Will to Power,* Kaufmann and Hollingdale (trans.) (New York: Vintage Books, 1967)

Strawson, Galen, 'The Impossibility of Moral Responsibility', *Philosophical Studies*, 75 (1994), p. 5-24

British Journal of Undergraduate Philosophy, 1(2), March 2006

Do liberals have an unrealistic view of the self?

Catherine Ruffell
University of Bristol
cr4782@bristol.ac.uk

Rawls' 'A Theory of Justice', one of the most important modern liberal texts, demands that individuals should choose the principles of distributive justice by which we should live from behind a veil of ignorance. This means that people should determine these principles with no knowledge of their own social circumstances or natural abilities. Rawls call this state the original position. According to Rawls, this will result in principles that are truly just, as they are chosen in a situation of true equality by rational beings. However, this argument has come up against much criticism, particularly from communitarian writers, because, they claim, it presupposes an unrealistic view of the self. According to communitarians, the self, detached from its social context, is a nonsensical concept. As Taylor puts it, "the free individual of the West is only what he is by virtue of the whole society and civilisation which brought him to be and which nourishes him," (Taylor, 1992, 45). In this essay, I shall examine the liberal conception of the self, and the communitarian criticism of it. Ultimately, I shall argue that the communitarian view is indeed fatal to liberalism, and that the self, as it is conceived of by liberals, is not an entity that should, or in fact could, determine principles of distributive justice.

To examine the communitarian criticism of the liberal conception of the self, it is first necessary fully to understand this conception, and see how it forms the foundation for modern liberal thinking. It is a conception that has its origins in Kantian moral theory. Traditionally, theories of both justice and ethics aimed towards a particular conception of the good life. For example, utilitarianism sees maximised utility as the aim of society. However, Kant disagreed with this way of doing ethics. Instead, he claimed that everyone has different ideas of the good life. Thus, any ethical principles based on a

particular conception of the good life can only ever be contingent. Furthermore, our ends and desires are given to us by our social or historical circumstances. If we base our principles on these factors, then we are not self-governing, but are in fact giving up our liberty; it is a "capitulation to determinations given outside of us," (Sandel, 1992, 15). Instead, Kant argued, our principles should be based on reason alone, without the interference of personal ends. This led to the claims firstly that the subject is prior to its ends, and secondly that the right is prior to the good (Sandel, 1992, 17) where the right is "derived entirely from the concept of freedom in the external relationships of human beings, and has nothing to do with the end which all men have by nature or with the recognised means of attaining this end," (Kant, quoted in Sandel, 1992, 15).

How does this relate to liberal politics? From the idea that the subject is prior to its ends, Rawls and other liberals have developed the central doctrine of modern liberalism: the primacy of individual rights. Liberals believe that individuals in a political society have certain rights that cannot be infringed. These rights take priority over any duty or obligation that the individuals might have to their society (Taylor, 1992, 30). The primacy of individual rights seems in many ways an attractive concept. If each individual's rights are respected, then there is no chance that we will have to submit to someone else's will. Even in a democracy, which many regard as the fairest political system, if the primacy of individual rights is not observed then there is a strong chance that the minority will be oppressed by the majority view. Liberalism aims to avoid this (Wolff, 1996, 115).

However, despite its intuitive appeal I would argue that liberalism, and particularly the doctrine of the primacy of individual rights, have some significant flaws. Just as Locke and Kant were criticised by Hegel, communitarian critics have accused Rawls' brand of modern liberalism of taking an overly 'abstract and individualistic approach' (Kymlicka, 2002, 209). While liberals see people as "isolated individuals who, in their own little protected sphere, pursue their own good in what they take to be their own way" (Wolff, 1996, 144), communitarians argue that people are "thoroughly social beings, [whose] identities and self-understandings are bound up with the communities in which we are placed" (Wolff, 1996, 144).

A key aspect of communitarian criticism prominent in the writings of both Sandel and Taylor is that the liberal self is not a realistic concept. Rawls wants us to choose our principles of distributive justice from behind the veil of ignorance. This suggests that he sees people as being separate from their position in society, their desires and preferences, and their personal history. Rawls is claiming that his unencumbered self can be completely removed from these things, and yet still essentially be this thing called 'the self'. Communitarians disagree with this conception. As Sandel argues, it rules out the possibility of what he calls constitutive ends – ends which are in some way an intrinsic part of who we are. Under the liberal view "no role or commitment could define me so completely that I could not understand myself without it. No project could be so essential that turning away from it would call into question the person I am," (Sandel 1992, 19).

But is this conception of the self a reasonable standpoint for liberals? Sandel says it is not, and his argument for this is based around the kind of community which Rawls' self is able to join. If the self is prior to its ends, and we therefore believe in the primacy of individual rights, then an individual can only be a member of a political community on a voluntary or cooperative basis. There can be no moral tie to the community (the individual would then be a member on a constitutive basis) (Sandel, 1992, 19). However, this type of membership of a political community is, according to Sandel, incompatible with the very project of determining principles of distributive justice. Rawlsian liberals claim that our natural attributes, and other aspects of our nature that develop from our social and historical background, are only arbitrarily ours. Therefore we have no claim over them. This is why Rawls is concerned with *distributing* the social goods which emerge from these attributes throughout a society, via his principles of distributive justice. However, Sandel argues, rightly, I think, that if we are not constitutive members of a society, then there is no acceptable step that we can take that allows us to transfer our attributes to our society (Sandel, 1992, 21-22). If we concern ourselves with principles of distributive justice at all, then we have to assume a constitutive link between individuals and the societies to which they belong. But such a link does not allow for the primacy of individual rights; the obligation to belong to a community has to take priority. In other words,

ends become prior to the subject, and the good becomes prior to the right. It is impossible to be concerned with distributive justice *and* take the individualist stance on the self that Rawls' liberalism demands (Sandel, 1992, 23).

Taylor also reaches the conclusion that we need to accept a moral obligation to our community, although he arrives there using a slightly different argument. Taylor focuses on the notion of what it is to make free choices. As I have tried to explain, liberalism attempts to protect the freedom of individuals by ensuring that no one else's will is imposed upon them. Liberals would argue that to be free and autonomous, it is important that one's individual rights are given priority over everything else. However, Taylor argues that this is to misunderstand the idea of freedom, and specifically, what it is to make free choices. He claims that our capacity to make choices can only be developed within societies. We need some experience on which to base our choices (Taylor, 1992, 34). The capacity to make free choices is also one that is, according to Taylor, a quality deserving of respect. It is a quality that we *should* develop. Therefore, like Sandel, Taylor claims that we have an obligation of some kind to belong to a community, in order to develop the characteristics which make us rational humans (Taylor, 1992, 35). But again, this obligation to the community rules out the primacy of individual rights that is so essential to the liberal view, and again rules out Rawls' idea of the unencumbered self. As Taylor points out, "the free individual of the West is only what he is by virtue of the whole society and civilisation which brought him to be and which nourishes him… And I want to claim finally that all this creates a significant obligation to belong for whoever would affirm the value of this freedom" (Taylor, 1992, 45).

The claim that the liberal conception of the unencumbered self is nonsensical is a damning one, as it forms the very basis of liberal theory. Is there any way in which liberals can successfully refute these arguments? Rawls has made some attempt to do this. He claims that communitarians such as Sandel have misunderstood the purpose of the unencumbered self and the original position. In suggesting that principles of distributive justice should be made from behind the veil of ignorance, Rawls is not, he argues, making a metaphysical claim about the nature of the self. He is not questioning the idea

that our ends are constitutive parts of us. Instead, he is merely asking us to imagine what principles of justice we would come up with if we were unaware of our personal desires and ends. The original position is to be taken as a hypothetical thought experiment: "The description of the parties may seem to presuppose some metaphysical conception of the person, for example, that the essential nature of persons is independent of and prior to their contingent attributes, including their final character as a whole. But this is an illusion caused by not seeing the original position as a device of representation" (Rawls, 1992, 203). However, I do not think that this argument helps Rawls to escape the criticisms of Sandel and Taylor. It is, I think, fairly obvious that Rawls is not asking anyone to really enter the original position. Whatever your conception of the self, this is clearly impossible. So Rawls' theory must be taken as hypothetical. What Sandel and Taylor are arguing is not that Rawls' unencumbered self is an *actual* impossibility, but that it makes no sense to even *think* about such an entity as being capable of making moral or political decisions. Rawls thinks that principles of justice can be determined independently of conceptions of the good. The communitarians have argued that they are not, and their objections are equally valid whether Rawls' claims are metaphysical or otherwise.

I would, however, argue that there is a possible response to communitarian philosophers. Although communitarian criticisms of liberalism seem successful, they are not able to offer a viable alternative political theory – their ideas are at best incomplete. Perhaps it is not rational to advocate the primacy of individual rights, but this is not to say that using conceptions of the good as the aim of our principles of justice is without its own problems. We still need to ask how we are to choose between the many different conceptions of the good. Walzer has attempted to resolve this issue with his theory of complex equality. He argued that what we need is a plurality of distributive principles, each based on different conceptions of the good, and each applicable to different social goods in different societies at different times. He claims that "to search for unity is to misunderstand the subject matter of distributive justice" (Walzer, 1983, 4). However, there are more issues that can be raised concerning this theory. It is, as Kymlicka points out, "a form of cultural relativism," (Kymlicka, 2002, 211) and I am sure that many

philosophers would not be happy with the idea that there is no final conception of the good on which to base principles of justice.

Nonetheless, the problems that communitarianism faces are not relevant to the success of its criticisms of liberalism. Communitarian critics have successfully highlighted a fundamental flaw in Rawls' modern liberalism, which is, it seems, inescapable. Furthermore, it is a flaw in the very foundation of liberal theory, and is therefore *fatal* to Rawlsian liberalism.

Bibliography

Kymlicka, W. 2002. Contemporary Political Philosophy, An Introduction. (OUP : Oxford)

Rawls, J. 1992. *Justice as Fairness: Political not Metaphysical* in Avineri, S. and de-Shalit, A. (eds.) Communitarianism and Individualism. (OUP : New York)

Sandel, M. 1992. *The Procedural Republic* in Avineri, S. and de-Shalit, A. (eds.) Communitarianism and Individualism. (OUP : New York)

Taylor, C. 1992. *Atomism* in Avineri, S. and de-Shalit, A. (eds.) Communitarianism and Individualism. (OUP : New York)

Walzer, M. 1983. Spheres of Justice. (Blackwell : Oxford)

Wolff, J. 1966. An Introduction to Political Philosophy. (OUP : Oxford)

.

British Journal of Undergraduate Philosophy, 1(2), March 2006

Book reviews

The mechanical mind: a philosophical introduction to minds, machines and mental representation

Tim Crane, *Penguin Books*

Edward Grefenstette
University of Sheffield
pha04eg@shef.ac.uk

From my experience of studying the rather 'new' topic of Philosophy of Mind (well, *newer* than most), there seem to be two major types of texts which lie on the opposing extremes of the academic balance of detail. On one hand, you have excellent initiatory 'textbooks' such as Smith & Jones' *The Philosophy of Mind* or Chalmers' book carrying the same title, while on the other hand you have a plethora of fundamental papers, articles and books by authors ranging from Descartes to Putnam via Turing, Searle, the Churchlands and various other illustrious mathematicians and philosophers from the past few centuries. The former give you a wide view of the 'big picture', providing you with suggestions for further reading and solid enough foundations to twiddle your arguments cogently into a discussion with a specialist, but can be slightly frustrating in that they generally only scratch the surface of arguments. *A contrario,* the latter give you the impression of being a specialist, until some well-read student comes along with "Well as [some author] points out, your point is not valid because..." and makes you regret that you haven't read more textbooks.

Isn't there a middle ground? Aren't there any 'focused textbooks' that provide you with a comfortable view of a particular line of argument and its ramifications, but also delve into the discursive details to a satisfying depth? Fear not, fellow aspiring cognitive scientists, for such books exist and Tim

Crane's *The Mechanical Mind* is one of them. In this short review of Crane's work, I will provide an outline of what to expect in and from this book before offering a few comments on Crane's arguments and presentation.

Despite the title, *The Mechanical Mind* is not just about, or even centred around, problems of artificial intelligence or thought associated with mechanical systems (in the colloquial sense of mechanical, grinding gears or processors and whatnot), but rather focuses on arguments supporting a computational theory of the mind. While questions of thinking computers are present in the later chapters, Crane describes these lines of thought as a detour of arguments complementary to the main cluster of cognitive theories he details in the first and last set of chapters.

The first chapter, "The Puzzle of Representation" introduces us to the basic aspects of representation. There is nothing about a slab of wood and four legs that makes me say "That's a chair. Why don't I have a seat?" However when I look at it, I seem to know it is a chair. I present it to myself. Representation is how the mind presents the world to the conscious self. It is the key, Crane says, to associating meaning to the symbols we find in the images our eye receives, the feelings our body perceives, and the words and concepts we come to understand. But what is it, exactly? You cannot really draw the relation 'P→Q', so representation cannot just be mental images. However we also have pre-linguistic elements, so representation cannot simply be language either. What is it for an object to be representational, or for an entity to have representations?

While keeping the problems of representation raised in chapter one on the table, Crane shines the spotlight more specifically onto the human mind in the second chapter suitably entitled "Understanding Thinkers and their Thoughts". Crane first discusses some of the classic issues related to the nature of mind (other minds, mind-body problem) before raising the questions he considers to be most problematic: How does commonsense psychology actually relate to how the mind works – are there such things in the mind as desire or intention? How do these notions relate to processes actually occurring in the brain? With some reservations, Crane argues that most notions of commonsense psychology do correspond to processes and states

described by scientific psychology, insofar as commonsense psychology is, as Adam Morton calls it, a 'Theory [of a] Theory". Crane states that this implies that a computational view of the mind is therefore not counterintuitive to a practitioner of commonsense psychology, and might therefore have some grounding in reality.

Having touched upon the hypothesis that the mind is computational, Crane posits that the question "Is the mind a computer?" cannot be avoided, which also raises the less necessary but equally interesting question "Can computers therefore think [have minds]?" In the third chapter, "Computers and Thought", Crane explains the basic ideas behind digital computers: how they are designed to follow instructions to the letter in an ordered, step-by-step manner. Such instructions are statements (corresponding to related machine states) which are linked to other states through conjuncts or conditional disjuncts, forming an algorithm which determines how the computer will produce output data based on input data. The idea that the brain has such algorithms seems counter-intuitive, as we are arguably not aware of any processes of the sort. Crane replies that we can actually think of many discreet processes of the sort. For example, electromagnetic waves of light being transformed into images by the eye and brain. Our body takes stimuli as input and produces behaviour as output, and both commonsense psychology and scientific psychology state that there is a certain level of predictability associated with visible patterns of behaviour, thus making this line of argumentation quite strong. Crane argues that the existence of such algorithmic elements in the functional structure of our brain explains the process of rationalization we go through when we adapt our behaviour to our situation, or to fulfil our desires and match our beliefs.

The idea of thinking computers encounters some more specific problems. After all, although psychology states that the mind is computational, it doesn't necessarily imply that a strict monolithic ruleset akin to that which runs a computer system is present. Consequently, can a computational system that follows rules *ad verbatim* actually think? After all, even if the output is convincing, does the computer actually understand the information being processed? The further we delve into them, the more these two questions seem

to merge into one: "What is nature of the language of thought?"

These problems are addressed in more detail as Crane searches for an answer to this last question in chapter four, "The Mechanisms of Thought". Surely, he argues, our mind has its own language, *mentalese*, that it uses to rule and describe our representational states. Crane discusses questions surrounding the nature of this language: does it provide further grounding for the computational theory of mind? Does it answer the problems facing our hopes of one day developing thinking computers? Connectionist theory is also brought to the readers' attention as an updated approach to AI and an alternative (but possibly complementary) theory to mentalese, arguing that while the objection 'rules cannot produce thought' holds for the classical 'expert system' approach of AI, one could consider a complex web of interconnected computing units (which Crane refers to as 'layers'). Thus Crane puts forward the idea that representation is an inherent consequence of the complexity of a computational system. However, he states that both mentalese and connectionist theory face problems which he believes may not be solvable without empirical evidence.

The final chapter, "The Mechanisms of Thought", brings the focus back to problems with representation itself. The past few chapters have attempted to explain how representation can be present in formal computational systems, and how agents capable of representation (i.e. humans or more specifically, the brain) could be reduced to computational systems. But to what, Crane asks, can representation therefore be reduced? Not many psychologists would argue that biochemical signals in the brain are representations. And many philosophers would have problems with the idea of a set algorithm ruling representation, as our representations do not always match reality, and one would expect a causally deterministic model of representation to be consistently right or wrong. Crane concludes that reductionism fails when dealing with matters of representation, but that we cannot dismiss the advances made in previous chapters so easily. He argues that non-reductionism does not deny that the mind is computational, but rather indicates that it is our views on what cognition *is* that must be revised.

The Epilogue leaves philosophers with a few open questions about what has been said, should they wish to reflect upon the mechanical model proposed in the book. It discusses some of the problems indirectly associated with computational theories of mind, such as issues of consciousness and phenomenology: even if we accept that the mind has a computational nature, how do we account for the *way* the mind *acts* upon our perceptions? Is such a concept even compatible with the 'input/output' model mechanical thought proposes? Crane suggests that such concerns do not necessarily conflict with the idea that mind has some sort of computational (or connectionist) aspect, but ultimately leaves their resolution to the reader.

In order to explain why I believe this book to be worthwhile, I refer you back to my initial statement about this book being halfway between textbook and personal argument. Crane peppers his chapters with sections I'd call 'anecdotic', in that they are not necessary to Crane's line of argument, but offer some keen insight into more technical aspects of the question. In other places, Crane sometimes encourages readers to re-refer to the section in question once they have understood the general structure of that part of the argument. Academics preferring works of a more traditional literary nature may dislike this, but I believe Crane's method to be beneficially pedagogical. It is definitely more in line with scientific study practices, leading it to be a popular book amongst psychology students reading cognitive psychology.

Students of metaphysics, however, may have certain qualms with the way Crane leads the arguments. On one hand, problems with each stance Crane describes are explained and commented upon, sometimes even argued in great detail. For example, in chapter four, Crane gives a fairly good account of the debate sparked by Dreyfus' and Searle's (separate) arguments against machine thought, and attempts to show how both retorts do not actually deny the possibility of artificial intelligence. However in other areas, especially in the earlier chapters, some might feel Crane passes over or even ignores potentially interesting areas of philosophy of mind. I, for example, regret not reading anything relating to property dualism or emergent materialism, and how compatible these theories are with the computational view of the mind. Also, while Crane does refer to the problem of images not being able to describe elements of language such as logical relations and notions of consequence, I

believe he doesn't do the issue of pre-linguistic elements of representation full justice. Although I don't believe it to be particularly problematic for the main positions Crane outlines, insofar as this purports to be a bit of an introductory textbook to the field of computational theories of the mind, I believe this part of the debate deserved mention in the later chapters.

However, my reply to the argument that Crane has an obvious physicalist bias which causes him to sometimes skip over portions of the debate would be that this is one of the strengths of the book, and that Crane makes an effort to take this into account. We are constantly reminded of the hypothetical nature of the assumptions Crane makes, which allows him to move along the line of argumentation rapidly and concisely while retaining a decent level of philosophical prudence. Although he obviously supports many aspects of the computational theory of mind, this book is more of a fleshed-out description of the theory and its supporting arguments, rather than an argument for it in itself. In fact, the pedagogical textbook-like nature I mentioned earlier encouraged the practice of returning to problematic aspects of each theory mentioned along the way, so as to think about and get involved in the ongoing debate surrounding the many aspects of this theory.

In conclusion, I think this book fulfils its purpose well. It provides the reader with a healthy overview of problems of representation, introduces arguments suggesting that the mind has an underlying systemic nature, discusses the implications of the mind being computational (namely the possibility of other mechanical systems being capable of thought), gives us some idea concerning the nature of the underlying systematic structure of the mind, and finally comes back to problems of representation, exploring how they apply to the theories outlined in the previous chapters. Each chapter is complemented by an annotated bibliography for academics wanting to challenge or further discuss points raised along the way. I would not recommend this as an introductory textbook to philosophy of mind, as it assumes some knowledge of the field, but taken with a grain of salt and an open mind, I believe this book to be an excellent introduction to the philosophy behind computational psychology. It is easy to read, reasonably simple to understand, written to be referred back to, and provides an excellent support for newcomers and those involved in the field alike.

The prayers and tears of Jacques Derrida: religion without religion

John D. Caputo, *Indiana University Press*

Andrew Stephenson
Cardiff University
stephensonac@cf.ac.uk

Jacques Derrida, it would be fair to say, is a big name in continental philosophy. Indeed, he is one of the biggest. How fitting then that a book about him – and it is, quintessentially, about him and not about deconstruction – should typify so perfectly all the stereotypes, true and false, that are graphed onto this tradition. The arguments, where they exist, are by no means explicit. Metaphors and puns are common and either superfluous or laden with too much responsibility. From at least three different languages words that needn't be left untranslated are left frustratingly untranslated, most often without any explanatory note. So called 'Edifying Divertissements' occasion the text in lengthy swathes, and although Caputo describes them as 'quasi-theological musings and amusements', their relevance, meaning, and importance are left far from clear. Biographical details and tenuous links to historical characters such as Saint Augustine are imbued, via a method of pseudo-psychoanalysis, with crucial consequence. These are the negative aspects of stereotypical continental philosophy: literary invention without merit. But the book also has the positive aspects, supposedly entirely lacking from stereotypical analytic philosophy, of being interesting, relative, thought provoking, and compelling. So much for stereotypes…

This will be a largely negative review. This book has a lot of problems. Aesthetically the substantial quotes that scatter the pages, starting or ending a section, often from Derrida or Kierkegaard, far surpass in beauty and insight any of the metaphors or puns that are authored by Caputo. (The 'tears' in the title, by the way, can be read as either salty droplets of water or rips and fissures.) Philosophically the problems are far more serious. When I started

working on this review I had two positive things I wanted to say about this book. First was the rather patronizing but nevertheless complimentary point that this is a nice book: it is easy and enjoyable to read, and it does not take itself too seriously. Almost as a matter of high principle, I took its self-deprecating tone to override the many philosophically reprehensible aspects of the book – to render it at least 'worth reading'. However, I have now changed my mind and must rescind that concession. This is not a nice book. This is a nasty book that is arrogantly elitist and subtly dogmatic in such a way as to produce merely the appearance of self-deprecation.

The second positive thing about the book, however, stands by philosophical merit alone. Uniquely in the literature on this topic, it has a most focused, informed, and elucidating discussion regarding the identification of Derrida's most famous neologism *différance* with negative theology's God. This section is crucial to any student who is hoping to make sense of the apophatic language that is used to define *différance*, often as non-full or non-simple or non-present. Caputo rightly shows us why *différance* is not, and cannot be, theological. Unlike the negative theologians' denials, which are intended to gesture towards the inexplicable nature of God, *différance* destabilizes precisely the transcendence of language that any theology seeks to achieve.

But it is here that Caputo makes the concession that ultimately undermines his central thesis. Deconstruction, in a very Kantian way, rejects the notion of transcendence in all its forms (resorting instead to quasi-transcendent*al* philosophy in search of the conditions of possibility and impossibility for language, meaning, and truth). *Différance* has its origins in Saussurian structuralism where words are given meaning only within an inescapable matrix of difference by which A is A because it is not B. Thus meaning is fully constituted by difference. Talk of God seeks the transcendent signified that resides outside or beyond this all-encompassing matrix. Talk of God, whether masked with apophatic or kataphatic language, seeks a *prima causa* or a *causa finalis* that has inherent meaning in virtue of nothing but itself. The remnant structuralist roots of poststructuralist deconstruction reject the possibility of this absolutely. In the end it comes down to the old Kantian chestnut of our inability to know the *noumena*, our inability to step outside our culture/society/language/morality/head. Caputo concedes that the meaning of

Derrida's (in)famous claim in *Of Grammatology* that 'there is nothing outside of the text' is simply an affirmation of this Copernican revolution: it carries no relativist ontological commitment. In turn he quickly concedes that Derrida's is a religion without religion, a religion without God – and here, I think, he has problems.

Caputo, in perfect stereotypical continentalist style, fails to address the necessary or sufficient conditions required for a definition of religion. Although there are issues about the definition of religion that go beyond the remit of this review, I shall tentatively suggest that a necessary condition for religion is a belief in the existence of an Other World. This Other World is of course itself ill defined, and it may also be argued that belief in at least one deity is also a necessary condition of religion, but I do not think it obligatory to go this far (and so, incidentally, to exclude Buddhism). We have seen that Caputo concedes that deconstruction strictly denies the possibility of transcendence, and that religion necessarily relies on transcendence. Therefore deconstruction can have no religion; it cannot be religious. This is my contention all too simply put, and needless to say Caputo realises this tension and attempts to manoeuvre his way around it (by creating an *ad hoc* definition of religion as a passion for the impossible), but along the way I think he makes several even more highly questionable claims.

Caputo characterizes deconstruction too as a passion for the impossible, and in this I would acquiesce. Most notably, in his exchange with John Searle, Derrida held that it is impossible to ignore the marginal cases that contravened J. L. Austin's 'ordinary language' philosophy, and that rather it was precisely these impossibly problematic cases on which rested the possibility of language itself. (Incidentally, this exchange left analytic and continental philosophers sharply divided on the issue of who came off best). Deconstruction is concerned with the margins of philosophy, the bits that need to be ignored for traditional theories to function; deconstruction is concerned with the impossible, true enough. Another example of this is found in Derrida's encounter with Emmanuel Levinas and his ethics of the Other, which plays a central role in Caputo's book. Levinas wanted to decentre the subject and make the wholly Other the centre of a new ethics. Derrida is sympathetic towards such an ethics, but he points out that the concept of

wholly Other is impossible, because to name any such thing as wholly Other is to trespass ever so slightly – but ever so crucially – on its otherness.

Caputo begins his misreading here by further naming the wholly Other as God. Now, if the ethics of deconstruction revolve around the impossible concept of the wholly Other, and the wholly Other is God, then deconstruction is most certainly religious. Thus Caputo speaks throughout the book of the prophetic bent of deconstruction, and even goes so far as to call Derrida a new Messiah. Apart from this being a most ridiculously flagrant example of the much-lamented hero worship in continental philosophy, in doing this he inverts all the deconstructive work that Derrida did in *Spectres of Marx*. Here and elsewhere Derrida begins to outline a structure of Messianicity that is utterly devoid of all spatially and temporally locatable Messianisms (such as those manifested in Moses, Jesus, and Mohammed). This allows for a deconstructive ethics that is not nihilistic or postmodern but rather represents an openness to the future and announces an always already-deferred call for justice. The details of this must be left unfurnished, but the point remains that by calling Derrida prophetic, by announcing the coming, indeed the having-come of a new Messiah, Caputo has not only failed to remain true to the deconstructive method and ethic, he has fully regressed to the dogmas and restrictions of traditional religions.

Caputo ends his book like this:

> *'…the passion for the impossible is…the passion for God, the passion of God. Whether or not one rightly passes for an atheist. If there is one.'*

If there is one?! There certainly *is* at least one: myself for one; Derrida for another; Marx for a third, Sartre for a forth and Ayer for a fifth; and countless others for countless more.

But there's even worse to come. Just as *any* faithful religious person – however tolerant and accepting – *must*; and every religious Fundamentalist *does*; so Caputo *also* brings every other (that is, every wholly Other) under his own religion, suspending – crucially, suppressively, explicitly and dangerously – that other's otherness:

'Deconstruction is a certain faith. Indeed, what is not?'

What is not?! Deconstruction for one; Marxism for another; Existentialism for a third and Positivism for a fourth. Yet not only does Caputo make the Quinian claim that theories and ontologies are relative, and each involves a certain 'faith' in an arbitrary choice and an *ad hoc* halt to the infinite regression of explanation. This would be odd (and entirely unargued for in this book) but perhaps acceptable. But he *also* simultaneously and hypocritically makes the claim that this 'faith' is theological in the restricted Western sense that implies religion. Perhaps this is nothing more than the logico-linguistic fallacy of equivocation. Put most simply: everything requires faith; faith is theological; therefore everything is theological. Put most commonly: a ham sandwich is better than nothing; nothing is better than eternal happiness; therefore a ham sandwich is better than eternal happiness.

Perhaps this is all it is, but there seems to me something more menacing than ham in Caputo's conclusions. For one thing, ham is not irreconcilable with deconstruction in the way that fundamentalism is. And so this book, I am surprised and regretful to inform you, is a nasty book…but it is still, for very different reasons than I initially thought, worth reading.

Upcoming BUPS events

Philosophy is of course much, much better if you're with people who are passionate about the subject and know what they're talking about. BUPS exists to bring together undergrads who love philosophy. Our events offer opportunities to give or discuss really great papers, to meet and mix with other undergrads who think worrying about ethics or the fundamental structure of mind and world is kinda cool. To build an understanding of how philosophy is done across the country. To meet other students who like this stuff as much as you do, have done their reading and want to talk.

We also organise the UK's only big, annual national undergraduate philosophy conference. Last year in Durham there were 50 of us from 20 universities across three countries, giving and discussing 16 carefully-selected papers over three days. This year we're going even further: five conferences over the next 10 months, publication of the best papers in the country, and for the end-of-year British Undergraduate Philosophy Conference in September 2006 a total delegation of about 80 undergraduate philosophers.

Interested?

Good. You should be at the events listed over the page then! You can see a typical programme or download the BUPC 2005 conference report at our website – **www.bups.org**. If you're not already on the BUPS-L mailing list for announcements, you can also subscribe through the site. Don't worry – BUPS membership is free and our conferences are all tailored to fit a student budget. Submit a paper or come along when you can - we'd love to meet you!

Latest details of all our activities, profiles of the committee and an updated list of upcoming events are always available at: www.bups.org

Any enquiries can be addressed to: info@bups.org

BUPS events & conferences 2006

Saturday 4th February 2006
BJUP day conference, University of Sheffield
Keynote: Professor Robert Hopkins (Sheffield)
Day conference to celebrate the launch of the BJUP

Friday 7th - Sunday 9th April
BUPS Philosophy Skills Conference, University of Nottingham
Keynote: Professor Michael Clark (Nottingham)
A combination of great papers and workshops on improving your key philosophy skills

Friday 30th June - Sunday 2nd July
Delegation to NPAPA 2006, University of Warwick
Keynote: Professor Bill Brewer (Warwick)
We have a limited number of places for undergrads to attend the UK's best postgraduate conference

Friday 8th - Sunday 10th September
British Undergraduate Philosophy Conference 2006, St.John's College, University of Durham
Keynote: Professor AC Grayling (Birkbeck)
This is the big, end-of-year BUPS conference. We've booked the large Leech Hall in St.John's and will be expecting quite a few attendees. Priority for places will be given to people who have been to at least one other BUPS event during the year!

BUPS also hosts a series of online discussions, accessible at any time of day or night via email – check the site for details!

Subscribing and submitting papers to the BJUP

BJUP Subscriptions

The BJUP is the English-speaking world's only national undergraduate philosophy journal. We publish the best papers from BUPS' conferences, but also accept high-quality essays by direct submission.

Our non-profit status keeps the cost of subscription to our print version down, and all BUPS members receive the electronic version of the journal for free. New issues go out quarterly. We offer three levels of subscription:

BUPS Member Subscription (Electronic)

Becoming a member of BUPS is really, really easy – all you need to do is join the BUPS-L mailing list. The electronic version of the journal is distributed to all BUPS members. We hope you enjoy it!

Individual Subscription (Print)

An annual subscription to the print version of the journal costs £40 in the UK, and a little more for international postage. Printed in A5 size on 80gsm paper with a 250gsm card cover.

Institutional Subscription (Print + Electronic)

Institutions (libraries, schools, universities) wishing to subscribe to the journal receive both a print copy and a personalised electronic copy licensed for unlimited distribution to, and printing by, current students of the institution. This package costs £60 per year for UK delivery, slightly more for overseas postage.

Subscriptions run for a single academic year, a current subscription covering the print version of issues 1(1)–1(4). Full details of how to subscribe, and methods of payment we accept are available at the journal's webpage:

www.bups.org/BJUP

Submitting a paper to the BJUP

Most papers we publish will be 2,000 – 2,500 words in length. However we will consider papers of any length. We would suggest that you limit your submission to a maximum of 5,000 words, though, since papers longer than this are often better dealt with as a series of shorter, tighter, more focused essays.

What we're looking for in papers that we publish is actually quite simple. We like work that is:

- carefully structured
- argumentative rather than merely descriptive
- clearly written
- knowledgeable about a given subject area
- offering a new argument or point of view
- not just written for area specialists

As a general tip, don't write with 'This is for a journal, I must be technical, formal and use lots of jargon to show I know my subject...' running through your mind. Explanation to others who may not have read the same authors as you, clear laying out of thoughts and a good, well-worked-out and -offered argument that says something a bit different and interesting: these are the key characteristics of the best papers we've received. Don't be afraid to tackle difficult or technical subjects – we're all keen philosophers here – but do so as carefully and clearly as possible and you have a much better chance of being published.

Most of our papers are analytic, but we are delighted to accept and publish good papers in both the analytic and continental traditions.

We accept papers electronically as Microsoft Word .DOC or Adobe Acrobat .PDF files. If you have problems sending in these formats, please contact us and we will try to find another mutually acceptable file format.

Papers should be submitted via email to **bjup@bups.org** and should be prepared for blind review with a separate cover sheet giving name, affiliation, contact details and paper title.

Don't worry about following the journal's house style before submission. The only requirement we have in advance is that you follow English spelling conventions. Any other requirements will be made clear if your paper is accepted for publication.

Please do not submit papers for a BUPS conference and the journal at the same time. We'll make suggestions for rewriting or restructuring papers we think could be publishable with a bit of work. Please do not re-submit a particular paper if it has been rejected for a BUPS conference or the BJUP.

Reviewing papers fairly is a difficult and time-consuming job – please give us a couple of weeks and do not submit your paper elsewhere in the meantime.

We run the journal on the minimum copyright requirements possible. By submitting work you license BUPS and the BJUP to publish your work in the print and electronic versions of our journal, and agree to credit the journal as the original point of publication if the paper is later published as part of a collection or book. That's all – you are not giving us copyright over your work, or granting a licence to reprint your work in the future. We're philosophers not lawyers, so we hope that's pretty clear and fair...

(courgettes not asparagus…)