Articles

RUSSELL'S NEUTRAL MONIST THEORY OF DESIRE

NICHOLAS GRIFFIN
Russell Research Centre / McMaster U.
Hamilton, ON, Canada L8S 4L6
NGRIFFIN@MCMASTER.CA

Russell's theory of desire in *The Analysis of* Mind is subject to a seemingly overwhelming objection, apparently stated first by Wittgenstein and subsequently elaborated even more compellingly by Anthony Kenny. The puzzle is that, before he became a neutral monist, Russell had used essentially the same argument as part of a critique of William James's theory of knowledge. Since Russell had already formulated the argument as part of his case against generally naturalistic, and specifically neutral monist, theories of propositional attitudes, why did he think his own neutral monist theory of desire was exempt? I canvass various suggestions, but argue that none of them are effective.

*B*y now it is pretty widely known that Gettier's famous counterexamples to the thesis that justified, true belief constitutes knowledge were to a large degree anticipated by Russell many years before. In *The Problems of Philosophy* there is the case of the man who in 1912 believes, truly, that the late prime minister's name begins with "B", but does so because he believes, falsely, that the late prime minister was Balfour, not Campbell-Bannerman (*PP*, p. 205) This case is repeated in *Human Knowledge* (p. 171), where it is joined by the case of the stopped clock: *S* forms a true belief as to the time of day by looking at a clock. The clock, unbeknownst to *S*, is stopped, but *S*

happens to look at it at the right time (*HK*, pp. 170–1).[1] Russell's examples are not strictly Gettier counter-examples, since Russell uses them only to refute the view that true belief constitutes knowledge and therefore makes no attempt to show that the true belief in question is also justified. However, it is not difficult (as Scheffler argues in the case of the clock) to elaborate the examples in such a way as to ensure that the justification condition is also met. The clock, for example, may have been just returned by the clock-repairer who assured *S* that it was in perfect working condition. In both cases, there is no knowledge because it is sheer luck that the belief is (justified and) true.[2]

While the two cases just mentioned have been quite widely noticed, there is a third, in *Theory of Knowledge*, which seems not to have been noticed at all and is part of Russell's critique in that book of James's neutral monism. Russell writes:

> Suppose, for example, that I wish to be with my dog, and start towards the next street in hopes of finding him there; but on the way I accidentally fall into a coal-cellar which he has also fallen into. Although I find him, it cannot be said that I knew where he was. (*TK*, p. 26)

Here of course it is sheer luck that Russell falls into the same cellar as did the dog and coming upon the dog thus luckily would be sufficient to defeat any claim to know where the dog was, even though one finds him expeditiously.

But Russell's example is puzzling in quite other ways than this, for where, one wants to ask, is the true belief in this example? The belief in question seems to be the belief that the dog was in the coal-cellar, but that was not a true belief that Russell entertained until he fell into

---

[1]   ISRAEL SCHEFFLER, *Conditions of Knowledge* (1965), p. 112, seems to have been the first to have noticed the stopped clock. But he makes no mention of the prime minister's name example, and neither, more surprisingly, does R. K. SHOPE in his otherwise apparently exhaustive (though now dated) *The Analysis of Knowing* (1983). J. E. LITTLEWOOD reported a conversation with Russell (*c.*1911) on the prime minister's name example in his *Mathematician's Miscellany* (1986), p. 128. I don't know who first drew attention to this case as a partial anticipation of Gettier.

[2]   "Luck" is used here without serious theoretical intent: it suits Russell's third example, which is my main concern in this article, and helps assimilate it to the two better-known examples. Some philosophers, of course, have given epistemic luck a much more robust theoretical role in dealing with Gettier examples, notably DUNCAN PRITCHARD, *Epistemic Luck* (2005). In invoking epistemic luck here, I am not taking sides on Pritchard's much more elaborate project.

the cellar himself—and then the belief *was* knowledge. It may seem as if Russell has muddled his example. It might also seem as if I'd done worse by trying to turn this into a Gettier example. Russell, after all, begins the example with "[s]uppose … I wish to be with my dog"; that is, with a case of desire rather than belief or knowledge. And yet, by the next sentence, Russell has turned it into a case of lack of knowledge ("it cannot be said that I knew where he was"). I shall come back to this apparent change of topic later, but for now I want to point out that Russell's main concern in this passage is with James's epistemology, not his theory of desire. Even so, it might be thought a stretch to claim that there is a Gettier example here, so let me defend that first.

To see what Russell's up to, we have to go back and consider how his example takes off from James's account of what it is for his idea of his dog to be (as he puts it) "cognitive of the real dog"[3]; that is, I take it, to constitute knowledge of the dog. James, of course, wants to put a practical construction on this and does so in the following characteristically idiosyncratic manner: if it is cognitive of the real dog, James says, the "idea is capable of leading to a chain of other experiences … that go from next to next and terminate at last in vivid sense-perceptions of a jumping, barking, hairy body" (*ibid.*). This condition is evidently met in the case Russell imagines, and yet, as Russell points out, he can hardly be said to have known where the dog was. James, of course, speaks of the "idea of his dog" being "cognitive", rather than about some belief about the dog being true, and this may seem insufficiently propositional for present purposes. But presumably James's idea of his dog could be suitably propositional, an idea of the dog's doing something or being thus-and-so or being some place. There seems, at any rate, nothing in his account which would preclude this, and I assume that, in setting up his counter-example, Russell has in mind the idea of his dog as being in the next street.

It might, nonetheless, be thought that the example still has a true belief deficit: the dog, after all, is in the coal-cellar, not the next street. We could, of course, change the example so that the coal-cellar was in the next street. That would still be effective: only one counter-example is needed. But really this is unnecessary even for a Gettier counter-example, if we take James's pragmatic conception of truth into account. According to James, "A conception is reckoned true … when it

---

[3]   *Essays in Radical Empiricism* (1912), p. 198.

can be made to lead to a sensation" (*op. cit.*, pp. 202–3), in this case, presumably, a sensation of the jumping, barking, hairy body. And that, in Russell's example, it did, albeit not in the way that was expected. The result, on pragmatist grounds, is a true belief, but not knowledge. Justification is easily added (suppose that the next street is where the butcher's shop is and the dog regularly makes its way there) to create a genuine Gettier counter-example for pragmatists. Moreover, James's criteria for knowledge are not only not sufficient for knowledge, they are not even necessary. Suppose Russell's idea is that the dog is gone forever. That idea is only "cognitive" if it *doesn't* lead to the jumping, barking, hairy body.

Of course, all this takes a harshly literal view of what James says, and, while I have no sympathy for philosophers who write vaguely or obscurely in the hope that their readers will formulate a defensible position on their behalf, it may be thought desirable to be more charitable, as Russell in fact is. He is not primarily concerned with poking holes in James's particular formulations of his position; he has a bigger, deeper objection in mind. Let truth and knowledge be defined in terms of "consequences" and "leading" in whatever way James fancies; Russell's example of the dog is intended to show that no such way will work, unless James can make a distinction between intended and unintended consequences. Thus Russell's finding the dog in the coal-cellar could count as knowing where the dog was *only* if Russell entered the coal-cellar intending to find the dog there. Moreover, it would be very difficult for James to make the necessary distinction between intended and unintended consequences. It would require a theory of desire, and that, Russell implies, would be difficult to achieve within the naturalistic, causal framework of James's neutral monism. This strikes me as a very powerful objection to James's position. It leads also to a problem for Russell scholarship.

Eight years after writing *Theory of Knowledge*, Russell published *The Analysis of Mind*. There he presented his own theory of desire. In it he rejected the idea that desires are directed to some object which the actions occasioned by the desire are intended jointly to achieve (*AMi*, p. 58).[4] Instead he proposed to explain desire behaviouristically by

---

[4] The same theory is presented in much the same words in "The Anatomy of Desire" (1919), *Papers* 15: 103–13; where, ironically, it precedes a review of Sir Ernest Shackleton's memoir of his last expedition which begins: "Man is distinguished from other

means of what he called "behaviour-cycles".[5]

> A "behaviour-cycle" is a series of voluntary or reflex movements of an animal, tending to cause a certain result, and continuing until that result is caused, unless they are interrupted by death, accident, or some new behaviour-cycle. (Here "accident" may be defined as the intervention of purely physical laws causing mechanical movements.)
>
> The "purpose" of a behaviour-cycle is the result which brings it to an end, normally by a condition of temporary quiescence—provided there is no interruption.
>
> An animal is said to "desire" the purpose of a behaviour-cycle while the behaviour-cycle is in progress.   (*AMi*, pp. 65–6)

Russell was led to this view by two sorts of considerations. Both are clearly and concisely stated in the 1918 prison notes and elaborated at greater length in *The Analysis of Mind* (pp. 59–63). Firstly, there is the need to explain human behaviour in terms "not too unlike" those we use in explaining animal behaviour:

> Mere external observation of animals leads the unscientific observer to say that they "desire" or "know" various things. This shows that desire and knowledge[6] can be *exemplified* in behaviour; therefore simplicity suggests that they might be *defined* by behaviour; i.e. the sort of behaviour which makes us say that an animal desires some end may be defined as *being* desire for that end.
>
>   ("Behaviourism and Knowledge", *Papers* 8: 257)

Secondly, there is the evidence of psychoanalysis, which shows that, since introspection is "difficult and uncertain", we may "assume

---

animals by intelligence and the power of persistent pursuit of ends" (*Papers* 15: 115). The first steps toward the theory are found in brief remarks in Russell's prison notes, "Behaviourism and Knowledge" (1918), *Papers* 8: 257. It also appears briefly in "On Propositions: What They Are and How They Mean" (1919), *Papers* 8: 288.

5   Russell's theory, however, is not, as we shall see, a purely behaviourist theory since he acknowledges that internal feelings (of discomfort) accompany the desire. Anthony Kenny, in an effective critique, argues that Russell's theory mixes elements of empiricist and behaviourist theories—neither of them satisfactory. *Cf. Action, Emotion, and Will* (1963), pp. 101–11.

6   Interestingly, Russell continues to treat desire and knowledge together, as he had in 1913. The basis for this is, no doubt, the propositional character of both.

desires of which the subject is unaware" (*ibid.*) on the basis of observed behaviour.

It may seem odd that Russell is simultaneously influenced by both behaviourism and psychoanalysis, which are typically thought to pull in opposite directions. But Russell's account reveals one respect in which they reinforce each other: psychoanalysis emphasized the unreliability of introspection, behaviourism dismissed it as unscientific. But of course Russell could only combine the two because he never entirely embraced behaviourism; he never rejected, as any fully fledged behaviourist must, the existence of internal states, including images and feelings. Russell admits these (quite sensibly) because he takes their occurrence to be too evident to be denied. Moreover, he also acknowledged "impractical wishes" which "suggest no activity directed towards their realization" (and likewise purely abstract knowledge which also lacks behavioural consequences).[7] "These", he said, "seem a difficulty for behaviourism" (*ibid.*), and he was evidently unwilling to dismiss them for the sake of doctrinal purity. In this, his empiricism trumped his behaviourism. *The Analysis of Mind* is perhaps Russell's most behaviourist work and his theory of desire is perhaps the most behaviourist part of it, but nonetheless both fall short of being fully behaviourist: the theory of desire because he acknowledged internal feelings of discomfort, and his general philosophy of mind because he acknowledged the existence of images.

In fact, it seems that psychoanalysis was a more important influence on his theory of desire than behaviourism (*AMi*, pp. 59–61). It was psychoanalysis that showed that the objects of desire which humans acknowledge (even to themselves) are not always the ones which motivate them. Observing what they do rather than what they profess is thus a better guide to what they really desire.[8] At all events, Russell's

---

[7] This was a fairly stable position of Russell's during the period when he was often supposed to be a behaviourist. For example, in a lecture on 10 December 1927 at Bryn Mawr he was reported to have said "I believe there is more in knowledge than behaviourism" (*The College News* [Bryn Mawr], 14 Dec. 1927, p. 4). I'm grateful to Ken Blackwell for bringing this text to my attention.

[8] It may also be noted that it was the First World War that persuaded Russell this sort of delusion could be both extreme and almost universal. Russell said that the war forced him to revise his view of human nature and that, even before he studied psychoanalysis, he came to "a view of human passions not unlike that of the psychoanalysts." "I had supposed … that it was quite common for parents to love their children, but the War persuaded me that it is a rare exception. I had supposed that most

theory of desire was not in any way suggested by the work of the behaviourists themselves, for behaviourists, then as now, tended to scorn the concept of desire as unscientific, part of what we would now call folk psychology, a concept to be replaced by more scientific talk of drives and motivation.[9] J. B. Watson was Russell's main source for behaviourist doctrine, but Watson, as Russell notes in "On Propositions" (*Papers* 8: 288), said nothing at all about desire. Russell's theory, therefore, seems very much his own creation—behaviouristically inclined but not yet behaviourism, like other aspects of his neutral monism.

Against his theory Wittgenstein objected: "if I wanted to eat an apple and someone punches me in the stomach so that I lose my appetite, then the punch was the thing I originally wanted."[10] Like Russell's example of falling into a coal-cellar, this also strikes me as a very powerful objection. Moreover, Russell's coal-cellar example can easily be revised to make it an effective counter-example to Russell's theory of desire. Suppose the dog is not to be found in the coal-cellar. Nonetheless, the fall no doubt brings Russell's dog-seeking behaviour-cycle to an end. So are we to conclude that what he really desired was to fall into the cellar?

---

people liked money better than almost anything else, but I discovered that they liked destruction even better. I had supposed that intellectuals frequently loved truth, but I found here again that not ten per cent of them prefer truth to popularity" (*Auto.* 2: 17). This cannot entirely be dismissed as exaggeration due to political fury, Russell's theory of desire would make most of these claims serious contenders for literal truth.

[9]   The concept of desire is conspicuous by its absence in much twentieth-century psychological literature, and not just that by behaviourists. For example, R. M. GOLDENSON, in his two-volume *The Encyclopedia of Human Behaviour* (1970), not only doesn't have an entry for desire, he doesn't even include the word in his index. The same was true of a number of other psychology texts I consulted.

[10]   LUDWIG WITTGENSTEIN, *Philosophical Remarks* (1975), p. 64. Wittgenstein doesn't indicate whose theory he is attacking, but there is little doubt it is Russell's. MONK, for example, assumes that that is the case (*Ludwig Wittgenstein: the Duty of Genius* [1990], p. 291), and GARTH HALLETT makes it clear how very closely Wittgenstein read *The Analysis of Mind* (*A Companion to Wittgenstein's "Philosophical Investigations"* [1977]). WITTGENSTEIN makes the same point rather less dramatically in "The Blue Book" where he talks about Russell's treatment of "expectation" and "wishing", this time with a rare citation of *Analysis of Mind*, Ch. 3: "We might … explain that a certain tension is said to be an expectation that *B* will come if it is relieved by *B*'s coming. If this is how we use the phrase then it is true to say that we don't know what we expect until our expectation has been fulfilled (*cf.* Russell)…. In Russell's way of using the word 'wishing' it makes no sense to say 'I wished for an apple but a pear has satisfied me'" (*The Blue and Brown Books* [1960], pp. 21–2).

The question which primarily concerns me is why did Russell, who deployed this argument to such good effect against James's theory of knowledge, not realize how telling it was against his own theory of desire? One suggestion is that in 1913 Russell was thinking about knowledge and not desire, so that when, a few years later, he came to think about desire he simply overlooked arguments which he had applied in an entirely different context. But this will not do. The prison notes make it quite clear that in 1918 Russell was thinking about knowledge and desire in tandem, just as he had five years earlier.

It seems clear, rather, that Russell must have thought that his theory was protected in some way against his own objection. His account of behaviour-cycles may be thought to give him some wiggle-room. In the first place, when he first starts to describe behaviour-cycles he notes that the actions are "appropriate" for a certain result (*AMi*, p. 63), and then, when he comes to define behaviour-cycles, he says that the actions "tend" to produce a certain result (*AMi*, p. 65).[11] Russell therefore seems to propose two different relations to connect a behaviour-cycle to its result: appropriateness, and what I take to be a probability relation, that the behaviour-cycle is likely to bring about the result. I want to argue that, within the context of Russell's own philosophy in 1921, only the second of these has any traction.

When behaviour is said to be "appropriate" for a certain result the claim is usually normative. It may mean that the behaviour was a morally or socially acceptable way to try to achieve the result, or that the means adopted were rationally adjusted to the end. None of these could be what Russell has in mind. There is also a more naturalistic sense in which we mean that a certain kind of behaviour would be *natural* for a certain type of creature to employ to achieve some end (e.g., it's appropriate for a bird, but not for a human, to regurgitate its food to feed its young). This naturalistic sense could have a place in Russell's theory of desire, but only if it can be cashed out in probabilistic terms. If it is understood in terms of behaviour which somehow fits the nature of the creature in question, and not probabilistically as behaviour the creature is likely to undertake, then it seems to invoke doctrines (like, e.g., Aristotelian essentialism) of natural kinds and/or natural necessity which would have been anathema to Russell. The probability relation is thus the only one which Russell can use to

---

[11]   I'm grateful to a referee for pressing this defence of Russell upon me.

explain the relation between a behaviour-cycle and its result.

So is Russell's theory, as thus understood, protected against the arguments we are considering because he defines a behaviour-cycle by means of the result that it *tends* to bring about? The first thing to note is that when Russell comes to define the *purpose* of a behaviour-cycle (as distinct from the behaviour-cycle itself), he defines it as "the result which brings [the behaviour-cycle] to an end" (*AMi*, p. 65), not as the result which *tends* to do so. But let's give him the benefit of the doubt on that, and adjust his definition of the purpose of a behaviour-cycle to make it accord with his definition of the behaviour cycle itself: the purpose of a behaviour-cycle is the result which *tends* to bring it to an end. For ease of reference, I shall call the resulting theory the statistical version of Russell's theory.

The statistical version does seem at first sight to be well protected against the objection as formulated by Wittgenstein against Russell or by Russell against James. It seems obvious that dog-seeking behaviour-cycles do not tend to end with falling into a coal-cellar; and apple-seeking behaviour-cycles do not tend to end with a punch in the stomach.[12] But this presupposes that we know what a dog-seeking or apple-seeking behaviour-cycle is. Only when we can identify the kind of behaviour-cycle involved are we in a position to determine its tendency to produce a certain result. The theoretical problem that Russell faces here is that of the reference class in probability theory: in order to determine the probability of a particular event's occurring one has to discover the frequency with which that sort of event occurs within some broader class of circumstances, the reference class, but there is no unequivocal way of deciding which reference class to use. Now it might seem unfair to tax Russell with solving the problem of the reference class before allowing him to use probability to formulate his theory of desire: the problem has no general, agreed solution.[13] But

---

[12] The converses, however, are true: falls into coal-cellars and punches in the stomach tend to bring dog-seeking and apple-seeking to an end.

[13] The problem is not avoided by avoiding the frequency interpretation of probability, it affects, though in different forms, all major interpretations. See ALAN HÁJEK, "The Reference Class Problem Is Your Problem Too" (2007). Russell's use of "tendency" in this context may be thought to suggest that he favours a propensity interpretation of probability. I think this is unlikely, but in any case it makes no difference: the propensity interpretation faces the reference class problem, too. See DONALD GILLIES, "Varieties of Propensity" (2000).

Russell's theory faces the problem in a quite dramatically acute form.

The standard problem is this: we don't know with what probability seeking an apple will end with a punch in the stomach (alternatively: with the acquisition of an apple) unless we can identify the appropriate reference class of behaviours; intuitively, apple-seeking behaviours. The probability of success of such behaviour, of course, depends upon many things—where you are, what time of year it is, whether there's a famine—each of which yields different reference classes and different probabilities of success. But in the standard case we can at least start from the set of apple-seeking behaviours; the reference-class problem is then to determine which sub-set of such behaviours is the appropriate one to use in estimating the probability of success in the case at hand. The problem for the statistical version of Russell's theory is much more severe, for, on it, we cannot identify even the class of apple-seeking behaviours without being able to determine the probability that such behaviours will result in finding an apple. In the standard case, we need the reference class to fix the probability; in Russell's case, we need the probability to fix the reference class. Yet the probability can still only be given if we have the reference class. Identifying a behaviour-cycle by means of the probability of its outcome is therefore not a possibility that is open to Russell.

Perhaps dog-seeking and apple-seeking are both too specific. Maybe we should start with more generic types of behaviour and only when those are explained turn our attention to more specific behaviour-cycles. David Pears says that "Russell's theory achieves its best fit with hunger, sex and similar needs."[14] But for humans, at any rate, both sex and hunger are such variably configured cultural constructs that one would be hard pressed to identify any particular behaviour (independently of its purpose) as part of a food-seeking or a sex-seeking behaviour-cycle. Just think of all the things that people do to get laid: what chance is there of identifying all these as sex-seeking behaviours (especially when so many of them—brushing your teeth, talking pretentiously about postmodernism—are also used for other purposes). At best Russell's theory might work for limited, routinized, stereotypical actions:[15] for example, dogs turning round and round to flatten

---

[14] D. F. PEARS, "Russell's Theory of Desire" (1976), p. 223.

[15] I have in mind the sort of behaviour that ethologists, following Lorenz and Tinbergen, used to call "fixed action patterns", though without their assumption that such

the grass before lying down (something they do even on carpets). This, at least, is a fairly clearly identifiable piece of behaviour, which ends (unless it is interrupted) with the dog's lying down. Whether it has a tendency to flatten grass depends upon whether the dog spends its time inside or out. Inside, it tends to damage carpets, but I wouldn't attribute that purpose to my dog.

But even if we could completely solve the reference-class problem, the statistical version of the theory does little good. True, it would solve the two original problems as posed by Russell and Wittgenstein, for (with a suitable reference class) dog-seeking tends not to end in falling into a cellar and apple-seeking tends not to end with being punched in the stomach. But counter-examples are easily developed against the statistical version of the theory. Consider agents which are spectacularly bad at what they attempt, a police force which only occasionally manages to catch a criminal, a sports team that rarely wins a match. One wants to say, with folk-psychology, that the police force is attempting to catch criminals, it is just not very good at it: its behaviour-cycles tend not to the apprehension of criminals but to their escape. So on the statistical theory, Russell will be forced to say that the desire of this police force is to let the criminals escape. There are undoubtedly many police forces in the world where this is true (the criminals being well connected and powerful), but it seems unduly cynical to assume that the purpose of all ineffective police forces is to let criminals escape.

Of course, one could gerrymander the reference class. Presumably even the most incompetent police force catches some criminals, and choosing a reference class in which those cases predominate (as no doubt the force does in its annual report) might enable Russell to conclude that catching criminals was what they were attempting. But one shouldn't eliminate a tendency to failure by rigging the reference class. That policy cuts both ways, and more easily the other way. It takes

patterns are innate. (KENNY, p. 101, mentions the courtship rituals of gulls.) The stereotypical nature of the behaviour makes it possible to distinguish when the behaviour-cycle has been completed from when it has been interrupted. But in other respects fixed action patterns are not helpful to Russell. For example, the period of quiescence which follows their completion tends to occur whether or not the action pattern was successful and thus, on Russell's theory, would not correctly identify the purpose of the behaviour-cycle. Moreover, subsequent research showed that the action patterns were rarely so fixed or so easily identifiable as had previously been supposed.

some skill (as bureaucrats responsible for creating official statistics will insist, since their salaries depend upon it) to find a reference class which tends to success; it is easy to find ones which tend to failure. Every class of behaviour-cycles which tends to success in dog-finding can be subsumed in a wider class which has no such tendency. The fact is that unless we have some reasonable way of identifying the reference class, the assignment of purposes to behaviour-cycles becomes almost completely arbitrary: with a carefully chosen reference class, almost any purpose can be assigned to any behaviour-cycle. But not quite *any*, for there are some tendencies to failure so egregious that no choice of reference class can possibly help. Many behaviour-cycles have been undertaken in an attempt to find the elixir of life or a perpetual motion machine, but none has had *any* tendency to achieve the desired result. The failure here is worse than statistical, it is uniform and absolute. Many tendencies might be revealed by the whole class of such attempts, but none of them is remotely plausible as an account of what was desired. In all these cases, we are back to the original difficulty, which Russell urged against James, of making a plausible distinction between intended and unintended results.

Russell's theory does not handle incompetence well. It *helps* that the statistical version of the theory identifies the desired result as that which the behaviour-cycle tends to produce rather than the result it actually produces; that does eliminate the original objections due to Russell and Wittgenstein. But it does not help with pervasive incompetence: Russell cannot acknowledge the existence of a statistical tendency to fail in bringing about some result. The statistical theory redefines this as success in not bringing it about.

The reference-class problem, however, haunts Russell's theory in a more profound way that we haven't yet considered. On Russell's theory we can identify the purpose of a behaviour-cycle by observing the tendency of a class of behaviour-cycles to which the one in question belongs. The tendency has to be observed in *completed* behaviour-cycles; that is, ones which end in success or failure (failure occurring when the animal gives up, i.e., stops the behaviour-cycle without achieving the state which it is the statistical tendency of that class of behaviour-cycles to produce). But a behaviour-cycle can end in more ways than just success or failure. In addition to being completed, Russell notes that behaviour-cycles can be interrupted, by death, accident or another behaviour-cycle. So a behaviour-cycle can end in one of

four ways: in death, in accident, in being interrupted by another behaviour-cycle or in completion (either with success or failure). The purpose of a behaviour-cycle can only be determined by means of a reference class of *completed* behaviour-cycles. But we cannot possibly identify which behaviour-cycles are completed unless we can identify which are interrupted.

It seems probable that Russell thought his theory was protected against Wittgenstein's objection by his provisos that death, accident or another behaviour-cycle might interrupt a behaviour-cycle short of achieving its purpose. Russell would surely want to argue that the punch in the stomach, like the fall into the coal cellar, was an accident that interrupted the behaviour-cycle and thus cannot be identified with its purpose. For this defence to be effective it has to be clear when a behaviour-cycle is interrupted and this is only possible if it is clear when death or an accident occurs or when another behaviour-cycle begins. Of the three, only death is clear enough to be helpful. Since death ends all behaviour-cycles with a prolonged, indeed permanent, period of quiescence, one might conclude, on Russell's theory, that it was the constant desire of every living thing. But let us assume that this is not the case and that death is always (except, of course, in the case of suicide) an interruption to a behaviour-cycle rather than its satisfaction. And, most importantly, that it is always clear (or at least clear enough) when death interrupts a behaviour-cycle.[16] Neither of Russell's other two provisos have this advantage and I shall argue that, on Russell's theory, it is impossible to tell with sufficient clarity when a behaviour-cycle is interrupted by an accident or another behaviour-cycle. We can do this, I shall argue, only if we have already identified the behaviour-cycle which is interrupted. On Russell's theory, we can do this only by observing the achievement of its purpose, and we can know that the purpose has been achieved only if we can know that it has *not* been interrupted. So let's consider, first, what Russell says about accidents.

Russell defines an accident for present purposes "as the intervention of purely physical laws causing mechanical movements" (*AMi,* p. 65). But these laws intervene all the time; indeed, their intervention is

---

[16] As with all empirical concepts, there will be uncertainty about the borderline. But this will coincide with potential uncertainty about desire. Was that sound air escaping *post-mortem* from the lungs, or a *pre-mortem* attempt to say something?

a precondition of our being able to undertake any purposive behaviour at all. Russell is aware of the problem. He considers the case of an animal which falls over a cliff: "if the animal is killed at the end of the fall, we have, at first sight, just the characteristics of a cycle of actions embodying desire, namely, restless motion until the ground is reached, and then quiescence" (*AMi*, p. 64). What rules this out, he says, is "the obviously mechanical nature of the whole occurrence" and the fact that "when an animal survives a fall, it tends not to repeat the experience" (*ibid.*).

Neither of these conditions, alone or together, will do the work required. It would be nice to think that having once fallen, an animal will avoid falling again, but this is often not the case. Indeed, the animal after a bad fall might have difficulty staying on its feet because it is stunned or injured. After a period of quiescence, the animal rises to its feet only to fall again, after another period of quiescence it repeats the performance, perhaps several times. Of course, we naturally think that the animal is trying not to fall, but how can the neutral monist know this? How is the neutral monist to avoid seeing this as a series of successful attempts to fall? There seems to be no way without being able to make the distinction between what the animal intends and what it does not (as Russell himself clearly realized when dealing with James). And this distinction seems to require thinking of desire as an attitude taken to anticipated results, exactly the sort of theory that Russell wanted to replace (*AMi*, pp. 58–9).

Moreover, repetition is not only not a sufficient condition for desire, it is not a necessary one either. Some actions are repeated even when not desired, and many desired actions are not repeated (and some cannot be). A castaway driven by hunger to eat his companion is not likely, one hopes, to repeat the experience. A mountaineer who wishes to be the first to climb a certain peak, if successful, cannot repeat the achievement. A desire to climb it again may result in a behaviour-cycle substantially similar to that of the first, but it will, for all that, be a substantially different desire.

The "obviously mechanical" nature of the occurrence seems a bit more promising. Obviously, we do actually make (and often correctly) a distinction between purposive behaviour and accidents. But we typically do this by first identifying the desire and then identifying as accidents the events which prevent its satisfaction. Thus my desire to have a drink is frustrated when the bottle I am holding slips from my

fingers and smashes on the floor. But Russell, who has to make this distinction on the basis of behaviour alone and without prior knowledge of my desire, seems to be in no position to decide whether I have a frustrated desire to have a drink or merely a desire to drop a bottle on the floor. Once again, we need a distinction between intended and unintended motions, analogous to the distinction between intended and unintended consequences, the absence of which was regarded as fatal to James's theory of knowledge in 1913. Simply calling accidental movements "mechanical" merely labels one side of the distinction, it does not show how the distinction is to be made.

What Russell seems to have in mind is that, in the case of accidents, an animal's behaviour will be predictable by the laws of physics alone. He says in connection with the falling animal that "ordinary physical forces operate upon the animal's body almost as if it were dead matter" (*AMi*, p. 64). This harks back to the general distinction he makes in his neutral monism between events which are governed by the laws of physics, those which are governed by the laws of psychology, and those which are governed by both. Let us suppose that intentional behaviour falls into the third class and accidental behaviour into the first. Even this requires some massaging because of the falling animal's "desperate struggles while it is in the air", to which Russell replies that the animal's "centre of gravity will move exactly as it would if the animal were dead" (*ibid.*). This is true, but only in a vacuum.[17] In reality, the path of the animal's centre of gravity will be the combined result of its actions and the operative physical forces (in this case, gravity and wind resistance), just as it was before the fall. One might say that, only in the case of the fall, do physical forces predominate. But even this is not true: the motion of the animal's centre of gravity when it is walking is predominantly a matter of gravity and the mechanics of levers. What is true is that the mechanical forces (as distinct from gravity and wind resistance) are largely under the control of the animal and have a much bigger influence on the movement of the animal's centre of gravity when it is walking than when it is falling. But this brings us back, once again, to the distinction between intended and unintended movements—from which there seems to be no possible escape.

In any case, the behaviour of an animal's centre of gravity seems a

---

[17] I'm grateful to Andrew Brennan for pointing this out.

tenuous way to make the necessary distinction. In zip lining and bungee jumping one's centre of gravity moves exactly as it would if one were dead, but the behaviour is intentional. Nor does it capture the case of my falling beer bottle: I simply loosen my grip on it and it falls. The issue is whether I loosen it intentionally or not. My centre of gravity, and that of the bottle, move in exactly the same way in each case. We may concede that if I lose my grip unintentionally the event is properly described by the laws of physics alone, whereas if I do so intentionally it requires both physical and psychological laws to explain the event. But overt behaviour does not reveal this difference. The only way to grasp that is to take into account what I intended to do.

Clearly Russell's falling into a coal-cellar was an accident. Or was it? The psychoanalytic literature is replete with examples of apparent accidents which (supposedly) served some hidden purpose and were not accidental at all. Psychoanalytic considerations may change the case from one in which a behaviour-cycle is interrupted by an accident into one in which it is interrupted by another behaviour-cycle (or subsumed under a larger behaviour-cycle). Moreover, Russell's account does not even satisfactorily account for the psychoanalytic examples he is trying to accommodate. If Russell's analyst concludes that his fall into the coal-cellar was no accident, he is unlikely to conclude (as Russell, on his theory of desire, must[18]) that Russell unconsciously desired to fall into a coal-cellar. He is far more likely to conclude that it served some other purpose than to bring about the event that brought the behaviour-cycle to an end, perhaps to avoid some stressful appointment later in the day. And how, of all the subsequent appointments Russell has to cancel as a result of his fall, is either he or his analyst, using only the resources of Russell's theory of desire, going to identify the one the cancelling of which was the purpose of Russell's fall into the coal-cellar? All this, it seems, can only be explained in terms of "intentions", "ends", and "objectives", albeit ones of which Russell himself was unaware and which require the skills of the analyst to reveal, but still ones of the very sort that Russell's theory of desire was intended to avoid.

It is not at all clear that Russell can make an adequate distinction

---

[18] The fall, we suppose, brings a period of quiescence which marks the end of a behaviour-cycle.

between completed behaviour-cycles and those which end in accidents. But the problem becomes even worse when we consider behaviour-cycles that are interrupted by other behaviour-cycles. This, as Kenny says (*op. cit.*, p. 107), makes Russell's definition of a behaviour-cycle "wholly unusable". Where one behaviour-cycle interrupts another, some very simple situations are problematic for the theory. Consider a person who goes to the fruit bowl, picks up an apple, and eats it. One could hardly hope for a simpler or more straightforward example of an apple-obtaining behaviour-cycle. And yet they actually went to the fruit bowl hoping to eat a banana, but finding all the bananas gone, decided to have an apple instead. In this case, we want to say that a banana-obtaining behaviour-cycle is interrupted by an apple-obtaining one. But to say that is to identify the behaviour-cycle by its goal, not the other way round as Russell requires. It's hard to see how Russell's theory could accommodate this circumstance: the two behaviour-cycles are indistinguishable until the very end. We could admit possible behaviour-cycles: in the nearest world in which there are bananas in the bowl the person eats a banana. But this would be anathema to Russell for it runs against the whole naturalistic tendency of his post-1918 philosophy. But that is just the problem: when dealing with intentionality neither behaviourism nor empiricism[19] work well, something which Russell appreciated in 1913 but seems to have forgotten in 1918.

The statistical theory does not help here. No doubt most banana-seeking behaviour-cycles do not end this way, they tend to end in the consumption of bananas. But plenty of behaviour-cycles, especially those where success is difficult, are replaced by others, where the goal is easier to attain. I might wish to do 75 push-ups but, once the behaviour-cycle has started, be driven, time and again, by pain and exhaustion to adopt a more modest goal: say 50. The statistical theory will imply that my purpose all along was to do 50. Students frequently find that behaviour-cycles which, in Russell's terms, constitute a desire to study are interrupted by ones constituting desires to party, play games, or sleep in. Only a cynical professor would conclude that they never had a desire to study.

How can we tell when one behaviour-cycle is replaced by another?

---

[19] It is worth repeating KENNY's point (*loc. cit.*) that Russell's theory of desire is not *purely* behaviourist but includes empiricist elements (namely, feelings) as well.

Intuitively, we identify a behaviour-cycle by means of its purpose. The purpose binds all the component behaviours together to form the behaviour-cycle. But this intuitive understanding is not available to Russell. For Russell we have to divine the purpose from the behaviour-cycle, and the behaviour-cycle has to be identified from its component actions. What binds a sequence of actions into a single behaviour-cycle on Russell's view? Russell does not offer an explicit account, but I suspect that he would argue that the unity of a behaviour-cycle is to be explained by the operation of causal psychological laws of the kind that his neutral monism admits. These will include mnemic causal laws (for a creature will typically adjust its behaviour in the cycle in accordance with the results of previous behaviour), but will not be confined to these. Physical causal laws will be involved as well, but will be insufficient to explain a behaviour-cycle on their own. Quite how this would work out is, of course, a more difficult matter, but perhaps it can be done. The point at which one behaviour-cycle is interrupted by another could then be identified by some discontinuity in the causal chain.

But behaviour-cycles do not just interrupt each other. One behaviour-cycle may form part of a larger one; or one may lead to another: in both cases with statistically significant frequency. In such cases there will be no causal discontinuity in the chain of behaviours. A couple go on a date, they have dinner, flirt, go home, have sex, smoke a cigarette, and fall asleep. If we assume that their falling asleep is the quiescence which normally follows the conclusion of a behaviour-cycle (rather than the purpose of the whole enterprise), then we will likely conclude on Russell's theory that everything that preceded it arose from their desire for a cigarette. It's not clear what basis Russell might have for breaking it up into smaller behaviour-cycles. Is having dinner a behaviour-cycle which constitutes a desire to eat and gets interrupted by the flirting behaviour-cycle (or, alternatively, was the eating behaviour-cycle completed and followed by the flirting)? Or is the dinner part of the flirting behaviour-cycle?[20] Then again, is flirting

[20] And this raises yet another kind of problem. Through any time period there is only a single sequence of behaviours that a creature can exhibit. (Behaviour is individuated extensionally, as Russell's naturalism requires.) But it is entirely possible that several desires are encompassed at the same time by that sequence of behaviours, perhaps all of them satisfied. (Desires are individuated intensionally.) Again, folk-psychology can make the necessary distinction by appealing to possible behaviour (or directly to

a behaviour-cycle in its own right or is it part of the behaviour-cycle which constitutes a desire for sex? On Russell's theory, it depends upon the frequency with which flirting leads to bonking.

Russell's theory of desire simply doesn't have the resources to deal with these questions, and many more like them—the previous points barely scratch the surface. Russell is not in a position to say where one behaviour-cycle ends and another begins, and until he can do that he is not in a position to identify the purpose of a behaviour cycle. The causal connectedness of a behaviour-cycle will not help here because the entire sequence of events is causally connected: each part is causally connected to the next and the same mixture of psychological and physical causal laws operates throughout.

If we allow behaviour-cycles to interrupt one another, there is no way of telling, on Russell's theory, whether a behaviour-cycle is completed or not; nor, if it was completed, whether it ended in success or failure (though Russell's account will tend to reclassify systematic failure as a different kind of success). Wittgenstein was right in "The Blue Book" when he noted the absurdity of a theory on which we do not know what we expect or wish for until the behaviour-cycle comes to an end. But the actual situation seems somewhat worse: even when the behaviour-cycle has ended, we may still not know what we wished for or expected.

Since, on Russell's theory, to identify a behaviour-cycle we have to see how it tends to end when completed, and since it may end by being interrupted by a different behaviour-cycle, Kenny argues (*op. cit.*, pp. 107–8) that we would have to identify the second behaviour-cycle before we could identify the first, thereby resulting in an infinite regress. This would give us a nice a priori argument against the theory, but the assumption on which it is based is too strong. To identify a behaviour-cycle we need to know where it tends to end. But to identify when it is interrupted by another behaviour-cycle, we don't need to be able to identify the second behaviour-cycle, merely where it starts. We don't need full behavioural identity criteria for behaviour-cycles (which would make it impossible to identify any behaviour-cycle), we merely need a partial distinctness condition: we need to be able to tell when an action is part of a new behaviour-cycle and not part of the old one.

Can this be done? Russell might appeal to the period of quiescence

purposes), something which is not open to Russell.

which, he says, typically follows the successful conclusion of a behaviour-cycle, so that, in the absence of a period of quiescence, we can infer that the behaviour-cycle was interrupted rather than completed. But quiescence is neither a sufficient nor a necessary condition for the completion of a behaviour-cycle. Long and complex behaviour-cycles often include periods of quiescence as creatures rest and recuperate the better to pursue their goal. And creatures who live exciting lives will frequently find that one behaviour-cycle succeeds another with no chance of quiescence.

But if we can't tell whether a behaviour-cycle ends in success or interruption, perhaps we can tell nonetheless when a new one begins. Here the prospects for success seem somewhat better because, as already noted, Russell's theory of desire admits of non-behaviourist elements in the form of feeling. Russell holds that a behaviour-cycle is started by "some sensation of the sort we call disagreeable" (*AMi*, p. 67). So we might say that one behaviour-cycle is completed and succeeded by another when one disagreeable feeling is brought to an end and replaced by another; while one behaviour-cycle is interrupted by another when the disagreeable feeling that led to the first continues, but is eclipsed by a different and stronger disagreeable feeling. There are phenomenological problems here. A hungry animal seeking food which finds itself stalked by a predator is unlikely to feel hungry as it tries to escape. No doubt the hunger will return once it is safe from the predator, but wouldn't this be a *new* disagreeable feeling leading to a new behaviour-cycle? Since Russell identifies these feelings as sensations it is hard to conclude otherwise. It would be difficult (especially for an empiricist) to maintain that, while being chased by the predator, the creature continues to have an unsensed sensation of hunger.

However these difficulties are to be sorted out, the theory faces still other ones because of the lessons Russell took from psychoanalysis. People will, in general, be poor judges of their disagreeable feelings, at least as regards the behaviour that will bring them to an end. Russell builds part of his case for his theory of desire on the fact that we can have desires which we do not recognize. It follows, therefore, that we may not correctly identify the disagreeable feelings which give rise to them. The amorous couple described above may think that they desired food and then, once they had eaten, sex. But, on Russell's theory, they may mistake sexual frustration for hunger, or nicotine addiction

for both. If disagreeable feelings are to enable us to distinguish between completed and interrupted behaviour-cycles, then they must carry their specification on their face (as Kenny says, *op. cit.*, p. 104) and in fine enough detail so that we will be able to tell phenomenologically when the disagreeableness that produces banana-obtaining behaviour is replaced by a qualitatively different disagreeableness that produces apple-obtaining behaviour. This seems frankly incredible. Moreover, as Wittgenstein points out ("Blue Book", p. 21), on such a theory explaining the meaning of "wanting an apple" does nothing to explain the meaning of "wanting a banana", for both are complex names of different types of disagreeable feelings. What is needed is a theory that treats "an apple" and "a banana" as different arguments of the function "wanting *x*".

The second respect in which Russell's philosophy of mind in 1921 avoids pure behaviourism is in its admission of images. Russell says little about images in connection with desire, but we might consider a theory, very roughly along the following lines: a behaviour-cycle is initiated by the image of something absent but desired, and ceases when the object which is imagined is present. By admitting the image as an image of the object desired, we avoid the problem of discriminating desires by means of feelings. Discriminating them by means of images looks much more promising. Images are cognitive in ways that feelings are not, and there is no shortage of them. One might plausibly postulate that, for everything a creature desires, it has an image of the thing desired. The account needs, of course, a good deal more elaboration. We need an account of how the image initiates the behaviour-cycle and of how the presence of the object ends it. And of the relation of the image, which may be quite generic, to the object, which is inevitably specific (and not necessarily even of the same genera as the image—as in the banana and apple example). And, of course, it is not clear that Russell (or anyone else) knows how to explain what it is for an image to be *of* something. But I shall ignore all these matters, because, as I shall argue, it doesn't really matter how these things are settled: the ideas+behaviour-cycle theory is defective, however the details are filled out.

The first thing to note about such an ideas+behaviour-cycle account of desire is that it goes flatly against the account that Russell actually gives in *The Analysis of Mind*, not just in the sense that Russell makes no use of images in explaining desire, but in that he expressly

excludes them. Such a theory as the one I have just hinted at is exactly the kind of "pull" theory of desire which Russell wants to replace with a "push" theory: "The primitive non-cognitive element in desire", he says, "seems to be a push, not a pull, an impulsion away from the actual, rather than an attraction towards the ideal" (*AMi*, p. 68). As such, it is part of the "ordinary unreflecting opinion" about desire which it is necessary to "almost complete[ly] revers[e]" in order to arrive at "true views" (*AMi*, p. 58). He continues:

> It is natural to regard desire as in its essence an attitude towards something which is imagined, not actual; this something is called the *end* or *object* of the desire, and is said to be the *purpose* of any action resulting from the desire. We think of the content of the desire as being just like the content of a belief, while the attitude taken up towards the content is different. According to this theory, when we say: "I hope it will rain," or "I expect it will rain," we express, in the first case, a desire, and in the second, a belief, with an identical content, namely, the image of rain.
>
> (*AMi*, p. 58)

This view, which Russell attributes to "common sense", is one which he thinks is "radically mistaken" (*ibid*., p. 59) and which it is the express purpose of his new theory to replace.

So, if we propose giving ideas a role in Russell's theory of desire, we are not actually defending his theory from the crucial objection, but rather presenting him with an entirely new theory—one, moreover, which he has already expressly rejected—and then defending that. So we still have no explanation as to why he thought that his own, actual theory of desire was immune to the criticism he brought against James. But, worse than this, the defence is ineffective anyway, for a very similar objection can be raised against the ideas + behaviour-cycle theory as was raised against the theory that understands desire in terms of behaviour-cycles on their own. Suppose I have an image of a policeman. It initiates a certain behaviour-cycle—furtive behaviour, hiding in shadows, wearing disguises, and driving fast through red lights—all of which comes to an end in the presence of the actual policeman. It doesn't follow from this that I desired to be arrested. Yet again, the distinction between intended and unintended consequences is crucial.

Russell's objection to James seems to me an extremely powerful one.

In addition, it seems that it can be easily adapted to apply to Russell's early neutral monist theory of desire and moreover to various variants of that theory which he might have proposed but didn't. It is not at all clear to me why he did not recognize this. My best guess would be that he thought that, having allowed for death, accident and new behaviour-cycles, he had protected the theory against the objection. But if that was his view, he was grievously mistaken, and for fairly simple reasons. If a behaviour-cycle can end in four ways—death, accident, interruption by another behaviour-cycle, and completion—we can tell when it ends in completion only if we can eliminate the other three endings. Arguably we can do this well enough in the case of death, and perhaps we could manage it in the case of accident, but there is no way it can be done in the case of interruption by another behaviour-cycle. There is thus no way to tell when a behaviour-cycle is completed, and thus no way to tell what its purpose was. Unless we have some way of independently identifying behaviour-cycles, we are simply in no position to judge whether one interrupts another or whether the second replaces the first once it is complete. Russell may have supposed that further scientific work would clarify the notion, that perhaps behaviourists could find a workable, operational account of what would constitute a single behaviour cycle. But in this, too, he was mistaken. The closest the behaviourists got to it was the notion of a single experimental trial (e.g. the running of a rat through a maze), but this was an artificial laboratory concept, with start- and end-points prescribed by the experimenter. No amount of experimental sophistication could convert this into the sort of notion that would serve Russell's purposes.

One important criterion of the success of a theory of desire is that it must explain desire at least as well as folk-psychology does. This is not to say that folk-psychology is always correct. A very good theory should be able to explain where folk-psychology is not. But no satisfactory theory can get wrong huge numbers of cases which it gets right. Folk-psychology in some sense provides the data that a theory of desire has to account for. By this criterion, Russell's theory behaves surprisingly badly. But what is, perhaps, most surprising is that Russell had anticipated its basic failure several years before he put it forward.[21]

---

[21] Parts of this paper were read at the Annual Meeting of the Bertrand Russell Society

WORKS CITED

GILLIES, DONALD. "Varieties of Propensity". *British Journal for the Philosophy of Science* 51 (2000): 807–35.

GOLDENSON, R. M. *The Encyclopedia of Human Behaviour.* 2 vols. New York: Doubleday, 1970.

HÁJEK, ALAN. "The Reference Class Problem Is Your Problem Too". *Synthese* 156 (2007): 563–85.

HALLETT, GARTH. *A Companion to Wittgenstein's "Philosophical Investigations".* Ithaca, NY: Cornell U. P., 1977.

JAMES, WILLIAM. *Essays in Radical Empiricism.* London: Longmans, Green, 1912.

KENNY, ANTHONY. *Action, Emotion, and Will.* London: Routledge and Kegan Paul, 1963.

LITTLEWOOD, J. E. *Mathematician's Miscellany.* Ed. Béla Bollobás. Cambridge: Cambridge U. P., 1986.

MONK, RAY. *Ludwig Wittgenstein: the Duty of Genius.* London: Cape, 1990.

PEARS, D. F. "Russell's Theory of Desire". In J. E. Thomas and Kenneth Blackwell, eds., *Russell in Review.* Toronto: Samuel Stevens and Hakkert, 1976.

PRITCHARD, DUNCAN. *Epistemic Luck.* Oxford: Clarendon P., 2005.

RUSSELL, BERTRAND. *PP.*

—. *TK* (*Papers* 7).

—. "Behaviourism and Knowledge" (1918). *Papers* 8: **18b**.

—. "The Anatomy of Desire" (1919). *Papers* 15: **23**.

—. "On Propositions: What They Are and How They Mean" (1919). *Papers* 8: **20**.

—. "Heroic Adventure in the Arctic" (review of E. Shackleton, *South*) (1920). *Papers* 15: **24**.

—. *AMi.*

—. *HK.*

—. *Auto.* 2.

SCHEFFLER, ISRAEL. *Conditions of Knowledge.* Chicago: Scott, Foresman, 1965.

SHOPE, R. K. *The Analysis of Knowing: A Decade of Research.* Princeton: Princeton U. P., 1983.

WITTGENSTEIN, LUDWIG. *The Blue and Brown Books.* 2nd edn. Oxford: Blackwell, 1960.

—. *Philosophical Remarks.* Ed. Rush Rhees; transl. R. Hargreaves and R. White. Oxford: Blackwell, 1975.