Draft for David Braddon-Mitchell and Robert Nola, eds, Naturalistic Analysis, MIT Press, forthcoming

Physicalism without Pop-out

Philip Pettit

Imagine a very fine grid or graph on which dots are placed at various coordinates so that, as a consequence, this or that shape materializes there. Depending on the coordinates of the dots, different shapes will appear and for every shape there will be a pattern in the coordinates that guarantees its appearance. Take, for example, the diagonal line that slopes rightwards and upwards at an angle of 45 degrees from the origin. This line is bound to make an appearance so long as the coordinates satisfy the condition or pattern that as they move away from the origin, (0, 0), the coordinates are progressively larger pairs of equal numbers: (1, 1), (3, 3), and so on.

In the world of such dots and shapes, it is going to be in principle possible, for any array of dots that realizes a relevant shape, to derive the presence of the shape from the numerical coordinates of the dots. More particularly, it is going to be possible to derive that shape without reliance on anything other than geometrically attainable, a priori information: first, that the given array of coordinates instantiates this or that pattern; and second, that the pattern guarantees the presence of the shape in question. The nature of the shapes on any grid — if indeed there are any relevant shapes present — is going to be a priori derivable from the positions of the dots; it is going to be possible in principle to derive the one from the other.

The simplest and most appealing version of physicalism parallels this sort of doctrine about dots and shapes (Pettit 1994, 1995). It holds that just as the positions of the dots determine the nature of the shapes a priori, so the way the natural world is physically organized a priori determines the way it presents itself in psychological and other terms. The way things are physically configured entails the presence of psychological and other realities and it does this without reliance on anything other than what a priori analysis can in principle reveal. There is an a priori entailment from the way things are physically — however 'physical' is understood (Pettit 1993b; 1994; 1995) — to the ways they are in other respects (Chalmers 1996; Jackson 1998; Chalmers and Jackson 2001).

I am not going to defend my understanding of physicalism in this paper, nor my commitment to the truth of physicalism, so understood (for more see Pettit 2003; 2004). Rather I want to focus on a general problem it must confront. This is that even if such a physicalism is true, it is not going to be very satisfying in a range of important cases. I have discussed this

problem elsewhere for the case of consciousness in particular (Pettit 2005); here I look at it in a more general and systematic way.

The problem that I raise for the physicalist derivation of psychology is that while we may be in a position to believe that the psychological is a priori derivable from the physical, so that the presence of this or that psychological phenomenon is a priori derivable from how things physically are, we are very unlikely to be in a position actually to conduct a derivation or even to get a sense of how it would go. We suffer from a derivational deficiency that takes away from the satisfaction that derivations generally give us.[1]

Conducting a derivation means surveying and endorsing the premises, developing an insight into why they necessitate the conclusion, and being inclined on that basis — being thereby moved or pushed or forced — to assent to the conclusion. But it turns out that you may have excellent reason for believing that a conclusion can be derived from certain premises, and you may be able to survey and understand those premises, without the belief in the premises providing any insight-based inclination to believe the conclusion. You will believe the conclusion if you are rational, but the belief will not be the spontaneous product of understanding and embracing the premises. It will materialize without a motivating insight into why the premises make the embrace of the conclusion unavoidable.

I think that as things currently stand, we are in this position with regard to the derivation of many aspects of the psychological from the physical. Ronald Knox claimed to want an argument for god's existence that would bring him to his knees at the conclusion. Most physicalists of my ilk would like something of that kind with mental phenomena: a derivation from physical premises that would make salient and inescapable just how such phenomena can materialize in arrangements of purely physical stuff. The thesis of the paper, however, is that at least in many cases they are unlikely to feel anything like this level of satisfaction in the best accounts — the best, by my lights — that are currently available.

The paper is in three sections. In the first, I go back to the analogue of the grid with the dots and shapes and identify two cases where we might have good reason to believe in a priori derivability without actually having the corresponding derivational ability. And then in the following sections I look in the light of those two models at the physicalistic derivability of two sorts of psychological state. In the second section I consider psychological states that are representational but not recursively representational, as I shall put it. In the third I investigate the rather more complex and troublesome case of states that are recursively representational. These

are states such that not only do the subjects of the states represent things to be thus and so; they are also in a position to represent things as being so represented by them.

1. Two models of derivational deficiency

The diagonal case

Returning to the grid envisaged earlier, it is relatively easy to show how a derivation might be provided for the presence on the grid of an upward, rightward line that starts at the origin and slopes at an angle of 45 degrees. Call this line the diagonal slope and call the pattern that ensures the presence of such a line, pattern D; this is the numerical pattern under which the first coordinate is (0, 0) and others are progressively larger pairs of equal numbers: they take the form (1,1), (2,2)….(m,m).... The derivation will go like this:

1. A priori: any coordinates in pattern D generate the diagonal slope.
2. Empirically known: the coordinates plotted on the grid instantiate pattern D.

Conclusion: the coordinates plotted generate the diagonal slope.

The first premise in this argument gives us the design specifications for the diagonal slope, as we might put it, and the second premise asserts that they are satisfied. You can easily tell the design specifications from the definition of the diagonal slope, recognizing that only dots with coordinates in pattern D will lie on that slope. And you can tell that the coordinates conform to that pattern by ready inspection. Not only will you be able to see, then, that the conclusion is derivable from the empirical information, without reliance on anything other than a priori presupposition. You will be able to conduct that derivation with little or no trouble from within your own, internal perspective. As soon as you survey, understand and endorse the premises, you will be moved by the insight they provide to endorse the conclusion. There will be nothing to give you pause.

The inference may come to you so naturally that you can be said to see the diagonal slope in the sequence of number-pairs; you may even see it there as surely and determinately as in the visual appearance of the graph itself. The number-pairs will contain the information that the dots are on that slope, since the connection with the slope is a priori. And, even more strikingly, they will impart or convey that information in a way that will move you spontaneously to admit the conclusion.

Your position in this case may be so good, indeed, that you barely notice the numbers when you register the coordinates; you just register straightaway the diagonal slope that the numbers entail. Think by analogy of how the accomplished performer can look at the score of a

piece of music and hear the profile that that score determines. Or about the doctor who can look at the shadows on an X-ray and see the ulceration or the cyst that completely escapes the unpractised eye. Or about the sailor who can look at the water in the harbour and see where the various currents and rips are running.

I have been concentrating on the ease with which you can conduct the derivation of the diagonal slope in order to emphasize the element of skill — inferential habituation — that is needed for this exercise. But even if the ease of inference is not so great, we can still think of you as conducting the derivation and as seeing the diagonal slope in the coordinate pairs. It may take you some time to work out that a given array of coordinates satisfies a certain pattern and it may come to you only slowly, say as a result of going through various tests, that this is the pattern for the diagonal slope. But it still remains that at the end of the exercise you will be inclined on the basis of insight to assent to the presence of the slope; you will find yourself compelled to countenance the slope.

Most of us are familiar with the pop-out displays that look at first to be just a jumble of colors and shapes but that can come — perhaps immediately, perhaps only after attentional effort — to present a prominent gestalt to the eye. In the case of the diagonal line, something similar to this experience will routinely occur in someone with your know-how: something we might call inferential pop-out. Whether with greater or lesser ease, the presence of the diagonal slope in the array of coordinate pairs will come to be registered with a sense of 'Eureka': 'Now I see'.

Shallow derivational deficiency

Let us move now from the simple case of the diagonal slope to a figure that is not so easily definable and not so familiar: that of an S shape. Just as there will be design specifications for a diagonal slope that can be worked out a priori, so the same will be true with this shape. You will know that an S consists in a line that, within certain degrees of tolerance, moves down slightly and slowly to the right for a certain distance; that it then turns upwards and leftwards sharply for about twice that distance; and so on. And knowing this, you will be able to work out by analysis that only dots with coordinates in a certain pattern — call it pattern X — will constitute an S shape. You may have to do this by looking at the different shapes within a given scale that would conform to your sense of an S shape, answering to the demands of your perceptual concept, and then working out the constraints on the coordinates of any such S-type shape. Or you may be able to do it by finding an abstract, algebraic formula for constructing an S shape and then working out the limits on coordinates that will satisfy that formula.

Given that this is the case, you will be able to endorse an argument for the a priori derivability of the presence of an S shape of the following sort.

1. A priori: any coordinates in pattern X generate an S shape.
2. Empirical: the coordinates plotted on the grid are in pattern X.

Conclusion: the coordinates plotted generate an S shape.

Here, as in the other case, the number-pairs presented contain the information that the dots form an S shape. But unlike the other case they will not spontaneously impart or convey that information to you; they will not move you more or less spontaneously to admit the conclusion. The reason for this is that though the two premises are both within your epistemic reach, they are not at all obvious. It may not be obvious that any dots with coordinates in pattern X will make an S shape, however a priori that truth. And it certainly may not be obvious that the coordinates given conform to pattern X.

The reason for this lack of salience in the premises, quite simply, is that S shapes are much less commonly confronted in using geometrical displays than are diagonal slopes. No doubt, if you paid regular attention to constructing S shapes — the sort of attention you may pay to diagonal slopes — then you would develop a sense of what was required for an S shape, and you would get into the habit of seeing the required pattern X in sequences of number pairs. And in that case you might become as habituated in the inference from coordinates to shape as in the other case.

Given that you lack that sort of familiarity with constructing S shapes, however, you will not be derivationally moved in this case in the manner in which you were moved in the last. You may have reason to accept the soundness of the argument given, registering that the presence of the S slope is a priori derivable from the nature of the coordinates. But you will not be in a position to conduct the derivation: you will not be spontaneously susceptible to the reason the premises provide for endorsing the conclusion.

What this case shows, then, is that there can be a gap between knowing that a conclusion is a priori entailed — this is knowledge possessed here as in the diagonal case — and having the derivational skill or know-how whereby you can be moved by understanding and endorsing the premises into drawing the conclusion. In this case you are subject to a derivational deficiency that blocks the experience of an insight-based inclination to assent to the conclusion. It blocks the possibility of inferential perception and pop-out, to return to the metaphor introduced in the previous discussion.

The deficiency illustrated is shallow rather than deep, coming from a shortfall in the level and quality of attention you pay to how to construct S shapes. Thus it is a deficiency that can be more or less easily put right. You won't need any new sort of skill in order to develop a facility in deriving the S shape of the kind that you enjoy with the diagonal slope. You will only need to extend familiar skills into this new domain

There are many cases where lack of familiarity gives rise to the sort of derivational deficiency that we have been illustrating. Think of the beginner pilot who knows exactly what the instrument readings are telling her but who struggles to let them override her kinaesthetic sense: who struggles, say, to conclude that she is accelerating when her kinaesthetic intuition, without a view of the horizon, is that she is falling. Or think of the inveterate gambler who learns of the gambler's fallacy — the fallacy of thinking, say, that a run of heads with a fair coin makes tails more likely on the next throw — but who finds it very hard indeed to apply this knowledge in practice. In both of these cases, as in many more, there is a theoretical form of inferential knowledge but a lack of practical inferential skill (Pettit 1998; McGeer and Pettit 2002). In both cases there is a failure of the person to perceive or to be primed by the pattern in the data. The person believes that that pattern is there, and has excellent reason for believing it, but does not find it inferentially salient or stimulating.

The fact that you can actively derive the presence of the diagonal slope from suitable coordinates — the fact that you can see the slope pop out from the coordinates — means that the claim about the derivability of the slope is going to be intuitively or phenomenologically satisfying. By contrast, the claim about the derivability of the S shape is going to be less satisfying; it asserts a relationship that you accept but cannot grasp in quite the same way. What will be available is only the proxy satisfaction of knowing from analogous cases like that of the diagonal slope what it would be like to be able to effect the derivation. You will be able to imagine the experience of being moved by insight into a pattern in the coordinates to acknowledge the presence of an S shape.

The position you are likely to be in with the S shape has parallels in the case of visual pop-out. You may know what it is for a gestalt to pop out of a visual display, and may know that there is a gestalt of, say, a cube to be seen in a certain pointilliste display. But you may just not be able to see the cube in the display — unlike, perhaps, others. You may have every reason to believe in the presence of the cube, and you may understand what this means from experience with other pop-outs, but you will still fall short of the full, phenomenological satisfaction that goes with seeing it there. Your position with the S shape will be exactly parallel. You will know

what it would be like to be able to derive the presence of the S shape, as a matter of inferential habit and compulsion, but you will not be able to enjoy that inferential experience yourself. You will be rationally required to acknowledge the presence of the S shape but you will not have any sense of inferential perception.

Deep derivational deficiency

Sticking with the general analogue of the grid and the dots, I now introduce a novelty that is designed to illustrate the possibility of a deeper derivational deficiency. In the cases considered so far, both plotted coordinates and the shapes they make are registered visually from a close-up consideration of the grid. Thus, by going back and forth between our visual tracking of the coordinates and our visual sense of what will count for us as a given shape — this, whether or not we have a formula for the shape — we can see that there is a necessary connection between the coordinates satisfying a certain patter and the shape being present; we can see the design specifications for the shape, expressing these as demands on the coordinates.

Consider now a case where the coordinates determine a shape, as before, but also determine something else — call it a profile — at the same time. The profile is a distally discernible property of the shape. It is something that can be seeen in the shape but only when you stand at a certain distance from the grid. The man-in-the-moon is a profile in this sense. And so is the figure that emerges in certain impressionist paintings, when you stand back from them. As the coordinates on a grid may fix the presence of a regular, proximally discernible shape, so clearly they may fix the presence of such a distally identifiable profile. The same argument will apply as in the two cases: the simple case shape and the profile-in-the-shape case; for sholrt, the shape case and the profile case.

The profile introduces a possible limitation that is very unlikely to strike in the other case. In the other case, whether it be that of the diagonal line or the S shape, it is possible to go back and forth between plotted coordinates and shape, since they are observable from the same viewpoint, and see how the pattern of determination or dependency goes. You will see this in salient exactitude with the diagonal line; you will see it only in broad outline — broad but scrutable outline — with the S shape. But in the profile case nothing like this may be possible, for you may not be able to go back and forth in the same way, tracking specific dependencies of profile on coordinates. Look close enough to detect the plotted coordinates and you won't see the profile. Go far enough away to see the profile and you won't be able to detect the coordinates plotted.

Imagine now that though you can change visual standpoint in relation to a grid, you can't rely on memory or notes to track the sorts of dependencies that obtain between a profile that appears at a distance — say, the profile of the pope's face — and the coordinates of the constituent dots. You can't go back and forth in the way you could with the ordinary shapes. A deep derivational deficiency will arise, if despite this limitation, you have good reason to believe that the coordinates on the grid — and any of a range of coordinate-sets satisfying a similar pattern, Y — allow the a priori derivation of the profile. And of course you will have reason to believe this, given the parallel between the profile case and the case with the regular shapes. You will have reason to think that there is an argument available of the usual kind:

1. A priori: any coordinates in pattern Y generate a papal profile.
2. Empirical: the coordinates plotted on the grid are in pattern Y.

Conclusion: the coordinates plotted generate a papal profile.

When you have access to the argument given, but only under the limitation mentioned, then you will suffer from a dual derivational problem. First, the shallow sort of derivational deficiency that goes with not having a ready grasp of the Y pattern and not being able to see when it is present. And second, the deep derivational deficiency that comes of not having any effective way of determining the design specifications for the papal profile, expressed in terms of coordinates: not having any way of working out pattern Y. There will be specific dependencies of the profile on the coordinates but these will not be salient, as in the diagonal case, nor scrutable, as in the case of the S shape; at most you will be able to venture only a broad hypothesis as to the form they take.

The effect of this deeper derivational deficiency on your sense of the a priori connection between the coordinates in pattern Y and the papal profile will be dramatic. It will mean that from your point of view there may be something quite inscrutable about the fact that the Y pattern guarantees the presence of the papal profile. You will lack the sense of understanding that would come with being able to see how in particular the profile depends on the coordinates.

This being so, we should notice, you are liable to have illusions about how the coordinates might be varied and the profile still be preserved. You will have a sense of the profile — an ability to recognise and imagine it, for example — that is not tied to registering the coordinates satisfying any pattern in particular; and vice versa. And so, although you believe in an abstract manner that there is a profile in the coordinates that guarantees the presence of the profile a priori — a priori and perhaps uniquely — it may seem that you can imagine the coordinates and the profile varying independently. You won't have the concrete, working sense

of the dependency of profile on coordinates that would banish such illusions of imaginability and conceivability.

The shallow derivational deficiency will block you, as we saw, from being able to experience anything like inferential pop-out, though you will at least have a sense of what such pop-out would be like. You will be in the position of someone who is familiar with visual pop-out, and knows there is a figure of a certain kind present in a certain visual display, but who just cannot get to see that figure. The deep derivational deficiency we have been discussing will make for worse trouble and more dissatisfaction. It will mean that not only do you not experience inferential pop-out, you will not have a sense even of what such pop-out would be like.

With our two versions of derivational deficiency distinguished, we have two models of how one may be in a position to believe in a priori derivability without having the ability actually to conduct a derivation. It is time now to apply those models to cases where physicalism claims to support the a priori derivability of psychological states. I shall argue that shallow derivational deficiency is the only problem in the first case but that the deep derivational deficiency also affects the second.

2. Non-recursively representational states

The information about the coordinates of the dots in our examples is parallel to the neuronal — better perhaps the neuro-environmental — information we might have about a psychological subject. If the sort of physicalism I accept is true, then that neuronal information will contain information about the psychological states of the subject, just as the information about the coordinates of the dots contains information about the shapes constituted by the dots. As there is an issue then about inferential insight in the diagrammatic case, so a similar question arises here. Can we expect as physicalists to be able to achieve — or to be able in principle to achieve — the sort of perception or insight that can prove elusive even in the diagrammatic case? Can we expect to be able to attain inferential pop-out?

I will discuss this question with regard to non-recursively representational states in this section and then recursively representational states in the next. I will argue that our position with states of the first kind is like our position with S shapes — not ideal, though not too bad — but that our position with states of the second kind is worse again; it is like the position with the profiles coded for, given we are denied access to how coordinates can vary and create varying profile effects.

The distinction between recursively and non-recursively representational states is relatively straightforward. Assume that representation can be naturalistically analyzed. Putting aside issues to do with explicit and implicit representation, and local or distributed realization, I propose that for current purposes we think of a representational state on the simple model that requires fulfillment of two conditions. First, that there is a generally robust connection — one that obtains in favorable conditions, however they are cast — between the form the state takes and the way the environment is configured; and, second, that the form taken by the state tends to lead the subject to behave in a manner that is intuitively appropriate for such an environment. However it is analyzed — whether in terms of these simple requirements or in some richer fashion — there is a more or less inescapable distinction between non-recursively and recursively representational states.

Non-recursively representational states will involve the representation of the environment without any representation of it as being so represented. For the subject whose only representational states are non-recursively representational, there will be no distinction between the environment as it is in itself, then, and the environment as it is for that creature — as it is, according to the creature's representations. The subject will be primed to respond in different ways to varying scenarios in the world, assuming something on the lines of the simple account of representation sketched above. But such variations in its environment will produce those different responses without the fact that they produce them being in any way represented by the creature. The varying scenarios will be represented <u>in</u> the subject, as we might say, but they will not be represented <u>for</u> the subject (Cummins 1983). They will be represented as the ways things really are, priming the subject to adjust appropriately; this must be the case with any effective representation. But they will not be represented as the ways things are represented by the subject as being, nor <u>a fortiori</u> will they be represented in any further, recursive manner: they will not be represented as the ways things are represented as represented as being, and so on.

Recursively representational states, by contrast, will be states such that the subject not only sees the environment as they represent it to be, being disposed to adjust and act appropriately; the subject will be in a position to see the environment as being represented in that manner. The environment will be represented as real in such states, assuming they are effective representations: that is, assuming their connection with adjustment and action is not suspended, as in the case of doubt. But it will be available to be represented as a represented and possibly not real environment. And indeed the recursive availability of such representation may be open-ended, with the subject having the capacity to form representations, at progressively higher levels,

of how things are according to lower-level representations. The further representation will become available in each case, we may assume, at the point where it is called for by some task on hand: say, by the task involved in checking the accuracy of a representation. It will be available on a need-to-know basis.

I shall assume that there is no particular difficulty for those in any camp about acknowledging the difference between recursively and non-recursively representational states; more on this in the next section. It is close to the distinction drawn by Ned Block, at least in his later formulations of the idea, between representational states to which the subject has access, being able to make judgments about the way things are represented to be in those states, and representational states to which the subject has no access (Block 2002). While there is certainly controversy as to whether consciousness involves more than access of this kind, no one appears to doubt that such access is possible with some states, and not possible with others.

For an example of non-recursively representational states, we might think of degrees of probability, and perhaps preference, as they are depicted in standard decision theory. Decision theory requires of those states only that they mutate in certain ways under various forms of evidential input and that they cohere with, and perhaps explain, the choices that the subject makes. The subject may assign a high degree of probability to a certain scenario without being aware of that scenario as something it represents as obtaining or as probably obtaining. It may assign a high degree of preference to a prospect without being aware of that prospect as something that it represents as attractive.[2] The subject may be just a well-engineered artifact that does not itself have beliefs about how the world is according to its beliefs, as distinct from how the world is, period; it may lack the very concept of representation.

Let us assume then that there are psychological states of a representational but non-recursively representational kind. And let us suppose that the way things are neuronally organized in a subject instantiates representational states of this kind. If the simple account of representation works, the way things are neuronally organized will dispose the subject to adjust to circumstances — to respond to evidence and to initiate action — in an appropriate manner; and if that account fails, then something different or extra will be required.

Might we be able to see the subject's neuronal configuration and, recognizing the pattern required for the representational states in question, just see in it the presence of those states? Seeing the subject's neuronal configuration will mean knowing how it is neuronally constructed and how it is integrated with environmental inputs, described in neuronally relevant terms. The question then is whether we could ever be able to move smoothly from a state of registering such

a neuronal or neuro-environmental pattern to a state of registering the subject's representational profile: registering its system of representational states.

In order to achieve this insight we will have to be able to have a sense of the design specifications on non-recursively representational states and we will have to be able to identify relevant neuronal configurations as satisfiers of those specifications.

Under the simple account of representation sketched — or indeed under many variations on that account — it will be relatively easy to grasp the design specifications. A representation will be present so far as there is a state that covaries in form with the putatively represented situation and that disposes the agent to act as is intuitively appropriate — given the agent's goals and collateral representations — to that situation. Thus the design specifications will impose limits on the causal sensitivity of the agent's states to the environment and on the causal productivity of those states in generating behavior.

But though we should have no problem in getting a grasp on the design specifications for a non-recursively representational state, we may well have a problem in identifying one or another neuronal array as a satisfier of those specifications: that is, in recognising the presence of the pattern that can guarantee the presence of the state. We are certainly likely to have this difficulty with an entity as complex as the human being, or perhaps any natural organism. But it may well be possible to overcome this problem in the parallel case of the simple piece of artificial intelligence, robotic or otherwise. We can imagine learning the input and output susceptibilities present in the electronic or mechanical make-up of a simple artifact and then, as we take those dispositions into account, recognising them as guarantors of the presence of representational states of this or that kind. We can imagine being attuned to how the physical structure of the artifice more or less manifestly implements the representational attitudes we ascribe. We would see the attitudes present in that structure as surely as the mechanic sees the different working parts of the car engine in the complexity beneath the hood.

We saw in the first section that, looking at the coordinates for a diagonal slope on a grid, we may enjoy a sort of inferential perception or pop-out. We may be able in the same way to have inferential perception of the presence of certain non-recursively representational states in the case of the simple artifact or at least we may be able to get close to that sort of experience. But it is unlikely that any one of us can achieve this insight with the complex, organic subject. In this case we will be under the same sort of limitation that besets most of us with the perception of the S shape in sequences of coordinates. We will suffer from the shallow form of derivational deficiency.

Assuming that some psychological states are non-recursively representational in the sense explained, we will able in principle to access the following sort of inferential knowledge in this more complex case.

1. A priori: any neuronal states in pattern Z realize such and such representational states.
2. Empirical: the neuronal states in the brain of this agent conform to pattern Z.

Conclusion: the neuronal states of the agent realize representational states of that kind.

The inferential knowledge that this argument gives us in the case of the complex subject — the knowledge of an a priori entailment that it yields — will not provide anything like inferential perception. It will not give us the ability to see the presence of the representations in the neuronal states.

This will be a source of dissatisfaction for any defender of my preferred variety of physicalism. But it need not be very frustrating. For just as there need be no deep puzzle involved in recognizing that a certain sequence of coordinates realizes an S shape — not at least for those of us who are capable of seeing simpler figures in such sequences — so there need be no particular puzzle associated with recognizing that a certain profile of neuronal states realizes a corresponding representational profile. In particular, there will be no puzzle involved for those of us who have an inferential sense of how a simpler electronic or mechanical profile can realize a simpler representational profile. Surveying such a neuronal configuration might give us no perception or insight into the representational profile that it implements but we can at least see what would be involved in achieving that sort of perception; we can draw on the parallel with the simpler case to give us a sense of this more complex counterpart.

What we can achieve in the case of relating non-recursively representational states to neuronal realizers is a sort of reduction that is familiar in philosophy. Take the reductive thesis in social ontology — an individualist as distinct from physicalist thesis — according to which the social entities and processes that exist in any social domain can be derived in principle from the dispositions and relationships of individuals (Pettit 1993a). With social entities and processes of any complexity — say, with banks and money and market exchange — there is going to be no possibility of achieving anything more than we can achieve in the derivation of S shapes. But that need not make for a serious problem, for there are lots of toy examples in this area where we can exercise that sort of derivational skill: we can see the presence of certain simple social realities in suitable specifications of individual attitudes and interactions; for example, we can see the presence of a convention in the behavior of people who hold certain attitudes towards the coordination of their activities (Lewis 1969). The existence of those examples makes accessible

the claim maintained in more complex cases, as the example of deriving the presence of a diagonal slope makes accessible the claim maintained about the derivability of the S shape.

Daniel Dennett (1979) speaks in the sort of psychological case we have been discussing of a difference between the physical stance in which we survey the electronic or neuronal construction of a subject and the intentional stance in which its representational or intentional profile become salient. Does the situation we have been describing merit this talk of a difference of stance? It does, so far as information is presented on the one side in a neuronal or electronic terms and on the other in terms of what is (non-recursively) represented. This scenario corresponds rather nicely, as it happens, to our model case where numerical information about the dots contrasts with figural information about the shapes.

But though there is a reason for talking here of a difference of stance, in another respect such talk is somewhat exaggerated. The reason is that numerical and figural information in the one case, and neuronal and (non-recursively) representational information in the other, are varieties of information that can be available simultaneously. There is no special difference of perspective or standpoint required for moving between the two. Just as I am, I can absorb either the numerical or the figural information about a geometrical diagram. Just as I am, I can absorb either the neuronal or the representational information about a psychological subject. This is the point, so it turns out, where recursively representational states are importantly different.

3. Recursively representational states

Recursively representational states are psychological states that enable the subject, not just to represent the environment after a certain fashion, but to represent it as an environment that is, precisely, represented. Recursively representational subjects may not be able to achieve recursion with every representational state they instantiate; some states, as it is said, may be subpersonal and unavailable to recursive representation. People will count as recursively representational subjects, on my usage, so far as they can achieve recursion with any representational states, however restricted the range of those states.

There ought to be nothing particularly controversial about positing the existence of recursively representational states and subjects. The recursion may obtain just in virtue of the subjects becoming able to form beliefs, not only about how the world is, but about how the world-as-represented is or, alternatively, about how the world appears or seems to be. And there is scarcely any denying that subjects like us do form such recursive representations. The recursion need not involve any experience of the initial representation; the recursive representation may be

purely a matter of belief, not of perception or sensation (see Carruthers 2000). And the belief involved may bear entirely on how the world is according to the original representation — how the world seems; it need not be an introspective or reflective belief about how I the subject see things, or about how they are represented within me. We may describe the recursive representation as a meta-representation or a higher-order representation but it is important to be clear that what it primarily represents is the world-according-to-the-relevant-representation — if you like, the content of that representation — and not the state of representation itself.

The importance of representational recursion from our point of view is that whereas simple representational subjects will be lost in the world, as that world is represented by its states, recursively representational subjects will be able in principle to make a distinction between how the world is and how it is represented as being. They will do this just so far as they identify how the world is represented as being — form beliefs as to how it seems to be — and come to believe that it is not actually that way: the-represented-way the world is, so they can believe, is not the-real-way it is. There is something that the world will be like for such subjects so far as there is something — the represented-way-it-is — that is available to their representations and that is potentially different, according to their representations, from how the world is in itself.

There is now a well-established habit of glossing the notion of an experiential or phenomenally conscious state in terms of there being something it is like to be in that state (Nagel 1986). Recursively representational states, albeit they ensure that there is something that the world is going to be like for their bearers, need not be experiential states in this sense. There may be a close connection between being phenomenally conscious and being recursively representational (Pettit 2005); but nothing in the argument defended here depends on positing that connection, or on interpreting it in any particular way.

The most obvious candidates for recursively representational states are experiential perceptions, in particular those involving perceptual illusions that can survive conflicting beliefs. Consider the Mueller-Lyer illusion in which two lines of equal length look different because one line has an arrow head at each end, the other a reverse arrow head. Subjects who have the idea of representation can recognize that despite appearances the one line is not longer than the lower, and have beliefs about how the lines are represented for them — how they appear, what they are like — as distinct form how they are in themselves. There will be something very stable — something irritatingly stable — that the lines are like for the subject capable of recursive representation; there will be nothing that the lines are like, not at least in the same sense, for a creature without that capacity.

But recursively representational states include many representational states that are non-perceptual or at least not so purely perceptual. Consider the way things present themselves when I see them as requiring this or that response: this is the way to go on, I think, in using a certain word; this is the way to behave, I conclude, in determining my overall duty. Or consider the way the world is depicted by me when I see certain options as alternatives between which I can choose, and choose freely: these are directions in which I can move from the status quo. Or consider the manner in which I view someone who has apparently done me harm when I feel resentment and depict her as a responsible agent, fully deserving of blame. In all of these configurations of attitude — and clearly they are at the center of much of human life — the world assumes a certain form, according to my representations, and I am aware of this as a represented form that they have. I see the world in terms of certain patterns — those that go with concepts of obligation and freedom and responsibility — and I am aware of the world as represented, whether or not mistakenly, in that way.

It is probably clear where this line of thought is leading. The contrast between non-recursively and recursively representational states introduces a difference like that between the model involving the shapes and the model involving the profiles. Plotted coordinates and shapes are simultaneously available, being each available from the same visual standpoint, and so it is possible for someone who can recognise a given shape, having the concept of that shape, to work out the coordinates that will ensure its realization. Plotted coordinates and profiles are not simultaneously available, requiring different viewpoints, and under the model we described there is no possibility of moving between them in the same way.

Routinely representational states and neuronal configurations are simultaneously available in the manner of coordinates and shapes, at least under our simple account of representation or any of a number of variations on that account. But recursively representational states and neuronal configurations are not available in the same way and, as with coordinates and profiles, it is not possible in the ordinary run of things to move back and forth between the two.

The analogy between the profiles case and the case of recursively representational states becomes apparent once we ask after how we might gain a sense of the design specifications on a recursively representational state. It will not be enough for gaining a sense of these specifications to recognise the causal role that neurons will have to play, as this was enough with ordinary representational states. It will be necessary to gain access to the represented-ways-things-are that become available in recursive representation. But the only way of gaining this access will be by undergoing the recursive representation in question. And so the only way of achieving a sense of

the design specifications for recursively representational states will be by tracking the effect of variations in neuronal configuration on first-person experience of the represented-ways-things-are. This is directly analogous, of course, to the way we might hope to gain access to the design specifications for profile by tracking the effect of variations in plotted coordinates on how the grid looks from a distance.

As there was a deep derivational deficiency present in the profiles model, so there will be a deep derivational deficiency present in this case. We can certainly see how neuronal variations can contribute to variations in what is non-recursively represented, changing the causal susceptibilities and propensities of the subject. This is because we can correlate changes on the one side with changes on the other. But how are we to correlate changes on the neuronal side with changes in how things are recursively represented? The only possibility will be to see the effect of such changes on one's own recursive representations. I can see how a third party — say, a simple organism — changes its representations as neuronal changes occur, just through seeing its altered dispostions to action. But I can see changes in recursive representation only in my own case, identifying variations in the represented-ways-things-are.

In the profiles model, clearly, it would be possible for someone to gain a sense of the design specifications for the papal profile by being able to go back and forth between the plotted coordinates that determine the profile and the profile itself. What holds here in parallel is that a deep derivational deficiency will be avoidable so far as people can go back and forth between neuronal observation and first-person experience, coming to learn, however roughly, of the way in which neuronal variation makes for variation in their own recursive representation.

But there's the rub. The deep derivational deficiency that threatens here looks much more inescapable that the corresponding deficiency in the profiles case. For there is no easy analogue to the recipe for repair that we offered in that case. You can overcome the difficulty with profiles just by being given access to the connections between variations in the plotted coordinates you see and variations in the profiles you see. But there is nothing to give you access to the connections between neuronal variations and variations in how the world is recursively represented.

Or at least there is nothing to give you this access at present. It is entirely possible that future technology will enable us to gain a sense of how recursive representation depends on neuronal variation, letting us see patterns that our neuronal configurations must satisfy if they are to sustain this or that form of recursive representation. But as things currently stand, this is something that is effectively denied to us. We may have very good reason to believe, as

physicalists believe, in the existence of patterns that provide an a priori demonstrable guarantee — a guarantee demonstrable in principle — of the presence of certain recursively representational states. We may even have well-argued hypotheses as to the general form that those patterns are likely to take. But we may still suffer from a deep derivational deficiency.

The effect of the deficiency, as in the profiles case, may be to deprive us a sense of what is imaginable and conceivable that answers to our beliefs about what is a priori necessary. The sense of the conceivable and the inconceivable — the intuition into what is possible and impossible — will be responsive, plausibly, to our sense of specific dependencies. And that sense of dependencies will be missing in this case, as with the profiles. Thus, regardless of our beliefs as to what is a priori guaranteed by certain neuronal configurations, we may think we can imagine things running counter to those beliefs. We may find ourselves with an intuition that a given configuration might be present without the corresponding represented-way-thing-are, and the other way around.

Take a recursively representational state such as seeing red, or following a rule, or identifying a behavioral option, or endorsing a feeling of resentment towards someone. It may be possible for us to try to work out the sorts of functional and physical conditions — ultimately, the neuronal patterns — that would suffice for such recursive representations. But even as we try to assure ourselves that such and such conditions would be bound to make things present themselves the way that they intuitively present themselves in these experiences, we have to recognize that all we can do is argue and assert. We will lack the spontaneous intuitions to back up these hypotheses.

Thus, we may analyze a conscious experience such as that of seeing red in terms of information-processing (Pettit 2003) but notwithstanding the case for that hypothesis, there remains the intuition that the physically determined processing might remain fixed while the color appearance changes or fades. Again, we may analyse the experience of rule-following in terms of how things are bound to seem from the perspective of an agent with certain extrapolative and regulative dispositions (Pettit 2002), but there still remains the intuition that these dispositions could be in place without the subject entering normative space and having a sense of the right and the wrong, the appropriate and the inappropriate. We may analyze the notion of responsibility in terms of the capacity we attribute to those with whom we think it is worth conversing, a capacity that we take them to possess even when they fail to exercise it (Pettit and Smith 1996; Pettit, Philip 2001 a; 2001 b). But there still remains the intuition that the capacity we attribute requires nothing more than complexity — in particular, a complex susceptibility to

conversational influence — and that it could well be in place without the freedom of the will that responsibility strictly presupposes.

Were we able to have a concrete sense of dependencies in these cases, rather than abstract hypotheses as to the form they take, then we would be able to experience something like inferential pop-out. Or at least we would be able to have a sense of what such pop-out would be like. We would be able to simulate the fulfilment of the antecedent conditions given in the analyses, and find ourselves driven to acknowledge that under those conditions, the represented-ways-things-are would be bound to fall in line. But in all these cases effective simulation will be denied us, lacking as we do the opportunity to track the specific dependencies of recursive representation on neuronal pattern.

Hume is well known, at least under some interpretations, for arguing that the experience of having one sort of event more or less invariably precede another leads us, mistakenly, to posit a necessary connection — by his account, a causal linkage — between the first and the second types of occurrence. The derivational or projective fallacy that he imputes to us is the inverse of the derivational or projective deficiency alleged here. In his case, there is no necessary connection between the distinct events envisaged and the problem is that there is an experience of being led by the occurrence of events in the first category to the expectation of the occurrence of events in the second. In our case, there is a necessary connection between physical and psychological phenomena — an a priori entailment — and the problem is that that there is no experience of being led by variations in the first domain to the expectation of variations in the second. It is this that puts the experience of pop-out beyond our reach.

Hume's problem is that our psychological habits induce us to posit among things connections that do not exist there, while my problem is that there are connections among things that our psychological habits do not induce us to posit. He rails against the siren songs of nature and custom; I lament the failure of nature and custom to sing. There is an allegation on both sides that mind and world are misaligned at the level of epistemic impulse but where he indicts epistemic impulse as a source of error — it indicates the presence of a non-existent relationship — I indict it as a source of ignorance: it fails to signal the presence of a relationship that does by our account obtain.

One final, consoling thought. The derivational deficiency that afflicts us according to the story told here may eventually prove to be remediable. Take the physicalistic claim that the way colors recursively present themselves is determined by how subjects are primed to register and process color-related information — information to do with illuminance, contrast, constancy and

the like (Pettit 2003) — using it as a basis for accommodating to their environment. At least one experiment shows — and, in the relevant way, shows the subjects of the experiment — that when the informational process is disturbed, then color-perception is disturbed too and that when the informational process is restored so as to ensure smooth functioning, then color-perception is restored at the same time. This amounts to showing an effect in how the world is presented visually to a subject — an effect in how colors are recursively represented — that is consequent on physical changes in his or her make-up.

The changes induced in the subjects of this experiment were brought about externally. The researcher had them wear glasses in which the lenses were colored in different ways — red at the top and green at the bottom, or blue on the right edges and yellow on the left (Kohler 1961; 1964). The idea was to see whether the way colors presented themselves would change — whether the distortions would disappear — as the brain adjusted and leached them out: that is, as it extracted information from the environment, consistent with the background, perhaps hardwired assumption that things do not change color when one moves one's head. The finding, exactly as a physicalist would predict, was that as soon as the visual system got back to processing color information consistently with that assumption, and in a manner suited to smooth color-based discrimination and behavior, the way things looked for the subject did indeed change; it returned to normal (Pettit 2003).

This adjustment took a long time to occur — a matter of several weeks — and involved a good deal of inconvenience. But imagine that one could induce the changes in minutes or even hours and see for oneself how the normal color appearance goes with normal adjustment in the registering and processing of related information. In that case one would surely be within reach of pop-out. One would be able to bring oneself to the point of vindicating the physicalist analysis of color appearance in terms of an ability on the part of the subject of the appearance to resolve and make use of certain color-related information. And one would be able to get a vivid inferential sense of how, in the presence of the appropriate, physically based ability, it is impossible that the subject not enjoy color appearances; and not enjoy precisely these or those appearances in particular. I see no reason, then, to despair. The frustrations of physicalism may not be with us forever.

Conclusion

Physicalists of my preferred stripe have to think that there is no room for conceiving of a gap — not at least under full information and understanding — between things being physically thus and so and this or that psychological state materializing. They have to think that once the

physical is suitably fixed, then as an a priori discernible matter the psychological will be fixed as well. But that commitment is hard to sustain because of a variety of intutions to the effect that certain recursively representational phenomena can vary independently of the physical.

I think that there are good arguments for analyzing many such phenomena — I have mentioned conscious perception, rule-following, free will — in physically unexceptional terms and for believing that if the physical terms on which the world operates are fixed, then the psychological will have been fixed as well. But I do admit that there are many sneaking intuitions that suggest it cannot be so. My response is to suggest that those backsliding intuitions simply come of the fact that in the cases in question we lack the capacity to habituate ourselves in the connections alleged and to recognize their effects. Habituation in these cases — habituation in deriving recursively representational states of any kind — can develop only by virtue of being able to experiment with the effects on the represented-ways-things-are of variations in the physical conditions of the brain. And as science and technology currently lie, that sort of tutoring is denied us.

We can all sympathize with the person who knows that there is the gestalt of a cube present in a pointilliste drawing but who just cannot get it to pop out perceptually or, worse, who does not even know what pop-out is like. The lesson of this paper is that if sympathy is due in that case, it is also due on a very much wider front. Those of us who believe in the a priori derivability of the psychological from the physical know, or think we know, that all psychological profiles are present in purely physicalistic configurations. But we are condemned to a position in which we just cannot make that fact salient and irresistible for recursively representational states. We are deeply deprived, at least for the moment, of any sense of inferential pop-out.[3]

ppettit@princeton.edu

References

Block, N. (2002). Concepts of Consciousness. Philosophy of Mind: Classic and Contemporary Readings. D. Chalmers. Oxford, Oxford University Press.

Carruthers, P. (2000). Phenomenal Consciousness: A Naturalistic Theory. Cambridge, Cambridge University Press.

Chalmers, D. (1996). The Conscious Mind: In Search of a Fundamental Theory. New York, Oxford University Press.

Chalmers, D. and F. Jackson (2001). "Conceptual Analysis and Reductive Explanation." Philosophical Review **110**: 315-60.

Cummins, R. (1983). The Nature of Psychological Explanation. Cambridge, Mass., MIT Press.

Dennett, D. (1979). Brainstorms. Brighton, Harvester Press.

Jackson, F. (1998). From Metaphysics to Ethics: A Defence of Conceptual Analysis. Oxford, Oxford University Press.
Kohler, I. (1961). "Experiments with Goggles." Scientific American **206**(May): 62-72.
Kohler, I. (1964). "The formation and transformation of the perceptual world." Psychological Issues **3**(4, Monography 12): 1-173.
Levine, J. (1993). On Leaving Out What It's Like. Consciousness: Psychological and Philosophical Essays. M. Davies and G. Humphreys. Oxford, Blackwell.
Lewis, D. (1969). Convention. Cambridge, Mass., Harvard University Press.
McGeer, V. and P. Pettit (2002). "The Self-regulating Mind." Language and Communication **22**: 281-99.
Nagel, T. (1986). The View from Nowhere. Oxford, Oxford University Press.
Pettit, P. (1993a). The Common Mind: An Essay on Psychology, Society and Politics, paperback edition 1996,. New York, Oxford University Press.
Pettit, P. (1993b). "A Definition of Physicalism." Analysis **53**: 213-23.
Pettit, P. (1994). "Microphysicalism without Contingent Micro-macro Laws." Analysis **54**: 253-57.
Pettit, P. (1995). "Microphysicalism, Dottism, and Reduction." Analysis **55**: 141-46.
Pettit, P. (1998). "Practical Belief and Philosophical Theory." Australasian Journal of Philosophy **76**: 15-33.
Pettit, P. (2001). The Capacity to Have Done Otherwise. Relating to Responsibility: Essays in Honour of Tony Honore on his 80th Birthday. P. C. a. J. Gardner. Oxford, Hart**:** 21-35; reprinted in P.Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.
Pettit, P. (2001). A Theory of Freedom: From the Psychology to the Politics of Agency. Cambridge and New York, Polity and Oxford University Press.
Pettit, P. (2002). Rules, Reasons, and Norms: Selected Essays. Oxford, Oxford University Press.
Pettit, P. (2003). "Looks as Powers." Philosophical Issues (supp. to Nous) **13**.
Pettit, P. (2004). Motion Blindness and the Knowledge Argument. The Knowledge Argument. P. Ludlow, Y. Nagasawa and D. Stoljar. Cambridge, Mass., M.I.T. Press.
Pettit, P. (2005). Consciousness and the Frustrations of Physicalism. Essays in Honour of Frank Jackson. I. Ravenscroft. Oxford, Oxford University Press.
Pettit, P. and M. Smith (1996). "Freedom in Belief and Desire." Journal of Philosophy **93**: 429-49; reprinted in F.Jackson, P.Pettit and M.Smith, 2004, Mind, Morality and Explanation, Oxford, Oxford University Press.

---

[1] The problem is related in a variety of ways to the problem of the explanatory gap that Joseph Levine (1993) made famous. I look at the deficiency that he identified with a view to derivation in particular, rather than to explanation more generally, and I attempt to identify sources of the difficulty.

[2] For all that decision theory supposes, indeed, the subject need not even be aware of the prospect as attractive; it may be attracted without having any belief to the effect that the prospect is attractive or attracting.

[3] My thanks for comments on earlier drafts to Colin Klein, Victoria McGeer, Susanna Siegel and, in particular, Jessica Wilson; she wrote an excellent commentary for the presentation of an early version to the Kline Seminar at the University of Missouri, Columbia, November 2004. I am also

grateful for the many comments received from the participants in that seminar and for useful conversations with David Chalmers and Daniel Stoljar.