For *Tool Use and Causal Cognition*, ed. T. McCormack, C. Hoerl, and S. Butterfill (OUP, 2011).

Representing Causality

Christopher Peacocke

What is it to represent a relation as causal? The question has a wide interdisciplinary significance. The question itself has been at the centre of philosophical discussion from Hume onwards; it is one that is pivotal in human developmental psychology; disputes in ethology about animal tool use cannot be resolved without addressing the question. Here, in a very general form, are three additional questions about the representation of causality whose significance crosses the usual disciplinary boundaries:

(1) What would be good evidence that some creature is representing a relation as causal?

(2) What would *make* it good evidence? In particular, how is its status as good evidence grounded in a constitutive account of what it is to represent a relation as causal?

(3) What is the relation between the constitutive account of what it is to represent a relation as causal, and what makes a relation one of causality? What is the right model of the relation between the content of the representation and what is represented?

This is evidently an area in which issues of intentional content, the conditions for attribution of mental states, and the nature of the content's reference bear upon one another. The issues develop very quickly into more general ones about the relations between intentional content and metaphysics. That is part of their interest. But rather than start with such grander themes, I will work my way up to them from a starting point that is often adopted in the ethological literature. I begin with the question of the relation between tool use and the representation of causation.

1. Tool Use and Causal Understanding: Relying, Representing, Reasoning

A Capuchin monkey cracks open a hard nut by lifting a heavy stone above his head and then bringing it down to crack the nut against a hard surface. A Caledonian crow extracts maggots from inside a tree trunk by manipulating a thin stick. An infant obtains a toy that would otherwise be out of reach by pulling towards him the towel on which the toy is resting. Do these actions manifest a grasp of a causal connection between the use of the tool and desired outcome?

Some theorists are confident that the conditions in which these tools are used, or the means by which use of the tool is acquired, or in some cases the nature of the task itself, themselves establish that use of the tool is a manifestation of knowledge of the causal relations between the use of the tool and the outcome. Here are some sample statements from these theorists. In reviewing the literature, Marc Hauser writes of the one-year-old human infant pulling the towel to obtain the toy: "One-year-olds solve this problem immediately. This shows that they can represent the causal connection between an action on one object and its effect on another object. This comprehension is at the root of all tool use".[1] He also mentions David Premack's chimpanzee Sarah. Sarah paired the sequence consisting of first an apple, followed by two half-apples, with a knife. She paired a sequence that reversed that order (and so ending up with a whole apple, having started with two half-apples) not with a knife, but with tape. Hauser writes of this experimental result that it is "Perhaps the most stunning demonstration of cause-effect comprehension in the domain of tools" (p.38-9). Elisabetta Visalberghi and L. Limongelli acknowledge that in some cases tool use may result from "the formation of associations between success and actions with objects".[2] But, they argue, when the solution to a tool-using problem is either learned by imitation, or is sufficiently complex, or involves a suitably complex object, an "understanding of causal relations is necessary for success in

[1] M. Hauser, Wild Minds: What Animals Really Think (New York: Holt and Co., 2000), p.31.

[2] In 'Acting and understanding: Tool use revisited through the minds of capuchin monkeys', in Reaching into thought: The minds of the great apes (New York: Cambridge University Press, 1996), ed. A. Russon, K. Bard and S. Taylor Parker, p.72.

each of the above three problems (imitation, complex condition, and T-tube).” (p.73). More generally, Andrew Meltzoff writes, “In the developmental and animal psychology literature, one of the most celebrated examples of causal reasoning is the case of tool use”. He speaks of “the gold standard of using a stick to obtain an out-of-reach target”.[3]

What are the minimal conditions required for tool use to be a manifestation of causal understanding? And do these examples and the reasons offered meet those minimal conditions?

We have first to respect the distinction between *relying on* a relation and *representing* the relation. In walking, any animal relies on friction. It does not follow that the animal represents friction. Quite generally, from the fact that some action of an animal relies on a causal relation between A-events and G-events, it does not follow that the animal represents A-events as causing G-events. When an animal performs an action of kind A, and an event of kind G results, the animal may simply have grasped a rule with this content:

to get G, do A.

In many cases, the rule would take some more conditional form:

When C holds, to get G, do A.

Prima facie, to establish that an action, including a case of tool use, gives evidence for the representation of causality, we need reason for going beyond the more austere hypothesis that the animal has grasped such (possibly conditional) goal-action pairs. We can call this the *Causal Discrimination Challenge*.

The challenge really has two parts. The challenge is, initially, to cite in any particular case evidence that supports going beyond the more austere hypothesis. The challenge then unfolds into something more philosophical: to explain why such evidence

[3] ‘Infants’ Causal Learning’, in Causal Learning: Psychology, Philosophy and Computation (New York: Oxford University Press, 2007), ed. A. Gopnik and L. Schulz, both quotations from p.43.

is evidence for grasp of causal relations. Such an explanation must draw on a positive account of what it is to represent a relation as causal.

It is important not to underestimate the explanatory resources of the more austere hypothesis that tool use and other actions are explained by grasp of such conditional goal-action pairs, rather than by a representation of causal relations. Here are five brief observations on the range of cases that can be covered by the more austere hypothesis.

(i) Conditional goal-action pairs can explain some cases that are plausibly classified as examples of creativity and intelligence, without attributing to the animal any representation of causality. Goal-action pairs can be chained together in an animal's reasoning. Suppose an animal knows these two rules:

When C holds, to get G, do A. (Rule 1)
When G holds, to get H, do B. (Rule 2)

Suppose the animal perceives that C holds, and so knows:

To get G, do A;

so the animal knows

If I do A, then G holds.

Since our animal knows Rule 2, he can reason to the conclusion:

If I do A, then: to get H, do B.

A structure of this sort would suffice to explain Wolfgang Koehler's ape Sultan using a box to reach a banana that would otherwise be out of reach.[4] In this case, the two instances of the displayed goal-action pairs would be

> When there's a solid object with a surface on top, to get to a high point, climb on top of the object

and

> When you're at a high point, and there's food above you within a certain distance, reach up to it.

Such instrumental reasoning does not, as far as I can see, necessarily involve representation of relations as causal.

(ii) Both Meltzoff in the human literature and Visalberghi and Limongelli in the animal literature seem to suggest that if what I would describe as a goal-means pair is learned by imitation, then that suggests that what is learned has a causal content.[5] It is indeed true that when a new pair is learned by imitation, the acquisition is not explained by what Visalberghi and Limongelli describe as learning "from the formation of associations between success and actions with objects" (p.72), if the associations are supposed to be with the learner's own actions. There seems to be no difficulty of principle with the possibility that which goal-means pairs are learned, and which sequences of event-types a subject learns through imitation, depends on the subject's perception of others as agents. It does not follow that *what* is learned has a causal content. In Meltzoff's well-known 1988 experiment, children learn from seeing others that touching a box with one's forehead is followed by the box's lighting up. Meltzoff

---

[4] W. Köhler, The Mentality of Apes (New York: Liveright, 1976), pp.39-40.

[5] Meltzoff, 'Infants' Causal Learning', p.45; Visalberghi and Limongelli, 'Acting and Understanding', p.73.

designed this experimental setup "to ensure that a new causal relation was being learned" (p.40). But the results of the experiment could more austerely be described as ones in which children learn from the actions of others the rule:

To get the box lighting up, touch it with your forehead.

(iii) An animal may have quite sophisticated representations of which events of a given type are followed by events of some other specific type, without conceiving of this relationship as causal. The recognition of a sequence-type may for instance involve a classification of the objects involved in the events as heavy or as light, as rigid or as flexible. An animal may draw on knowledge of these regularity-types in forming new rules of the form

If an object is rigid, to get it to meet G, do A to it.

Yet this still is not, apparently at least, something requiring the representation of causality. (I will return to deal with the philosophical position that this is all that grasp of causality ever involves.)

(iv) An appreciation of which regularities occur between events of a given type, and in which order, would be sufficient to explain the ability of David Premack's Sarah to see a knife rather than tape as the right thing to pair with the temporal sequence of an apple and then two half-apples. The two half-apples are present in the early part of a taping event, followed by a whole apple. In cutting the temporal order is reversed. More would be needed to show a representation of causal relations, rather than merely a sensitivity to temporal order.

(v) Attribution of the more austere goal-action pairs can also explain some patterns of errors in action. Westergaard and Suomi reported that while "a few capuchins cracked open walnuts with stones and probed with sticks to extract the nut meat", some

of them on some trials tried to use sticks on intact nuts.[6] This is what one would expect if the mental representation of which action-type goes with which goal became switched around. The hypothesis that the animals suddenly believe that sticks have different causal properties than they previously possessed would have consequences elsewhere, consequences that it is natural to conjecture need not be present when these errors in action are present.

One suggestion would be that what makes the difference between these goal-action pairs and the representation of causality is that in the latter case, there is representation of a connection between condition, outcome and event-kind that is known by the subject to apply beyond the case of his own actions and goals. Someone who is grasping the connection as causal, it may be said, is not merely representing the rule

When C holds, to get G, do A.

He is also representing some stronger generalization with the content

In circumstances C, any A-type event is followed by a G-type event.

The problem for this proposal is that while it may give a necessary condition on representing the connection as causal, it is certainly not sufficient. Even if the generalization needs to be understood as something more than a de facto truth about the actual world – because, for instance, the generalization is retained in the scope of reasoning about nonactual cases – there is nothing cited here that requires the generalization to be understood as involving causal connection. Its truth may simply require this: in nearby possible circumstances in which C holds, an A-type event is followed by a G-type event. This does not required G-type events to be caused by A-type events. They may, for instance, have a common cause consistently with this content. The thinker may also leave it open in his mental representation why in nearby circumstances the G-type events follow A-type events. So generalization beyond the case of one's own

[6] G.C. Westergaard and S.J. Suomi, 'Use of a tool-set by capuchin monkeys (Cebus apella)' Primates 34 (1993) 459-62.

actions does not yet amount to representation of a relation as causal – not even when there is projection to non-actual cases.

The upshot at this stage of our discussion is: we have not yet found any special connection between tool use and the representation of causality. If some representation of causality is involved in certain cases of tool use, it must involve something additional to the factors so far identified. What might it be?

One suggestion is that in cases in which tool use involves grasp of causality, when the agent does A to get G, he does so because he believes that if he does not do A, he will not get G. This is something that goes beyond the representation of a goal-action pair linking A with G. It involves acceptance of a conditional which, once the action is performed, the agent believes to have a false antecedent (that he doesn't do A at the time in question). So this thought could be developed into what we could call 'the believed-false antecedent' criterion. The proposal runs: when action is explained by belief in a conditional with an antecedent the subject believes to be false (once the action is performed), the action manifests some grasp, however primitive, of causality.

Someone tempted by the believed-false antecedent criterion could point to its application in other kinds of case. We want to distinguish the case in which a mother protects her offspring by luring a predator away from the nest and does so simply because she represents a goal-action pair, and the case in which she does so because she believes that if she does not move away from the nest, her offspring will be devoured.

I think the believed-false antecedent criterion is a step in the right direction. There is, however, no understanding or confirming counterfactuals without possession of some conception of what is involved in their truth. Though this is not the place to argue the matter in detail, I am in agreement with those who hold that the truth conditions and correspondingly the understanding conditions for counterfactuals have to be given in terms of laws and explanatory properties. This is the position of one who thinks that counterfactuals cannot be barely true, and that possible-worlds elucidations of their truth conditions, when correct, are so in virtue of other, more fundamental conditions. If this is so, the success of the believed-false antecedent criterion should point us to a more fundamental condition that is met by one whose tool use is a manifestation of grasp of

causality, a condition which relates to the materials in virtue of which counterfactuals are truth,

If an agent's doing A to get G really is a manifestation of a representation of something stronger, something involving causality, what would we expect, as a constitutive matter, to be the case? Before answering this question, I think we should identify a more primitive notion than causation *simpliciter*. We can conceive of a subject using a notion that does not distinguish, as we do, between causation and counterfactuals, and who uses this more primitive notion in representing relations between events. We can call it the relation of "yielding". The yielding relation holds between events when they are both causally and counterfactually related; it does not hold when they are neither causally nor counterfactually related; it is not determinate whether it holds when only one of the causal and counterfactual relations holds between the events. More explicitly:

> A particular event *a* yields a particular event *b* if it is the case both that *a* causes *b* and if *a* had not occurred, *b* would not have occurred.
> *a* does not yield *b* if *a* does not cause *b* and it is not the case that if *a* had not occurred, *b* would not have occurred.
> In the case in which the propositions *a causes b* and *if a had not occurred, b would not have occurred* diverge in truth-value, it is indeterminate whether *a* yields *b*.

I introduce the notion of yielding because if a subject does not draw a distinction between counterfactuals and causation, but does employ a notion of yielding, that certainly seems enough to credit the subject with a notion of an explanatory relation that goes well beyond statistical regularities in the actual and in nearby worlds. So now we can ask: if an agent's doing A to get G is a manifestation of a representation involving the notion of yielding, what else should we expect to be the case? Our agent will represent it as being the case that

> Doing A will yield a G-event.

For the displayed condition to hold, it must be the case that:

> There is some property P such that it's because A-events are P that they are followed by a G-event.

Commonly the agent will know what property of his action it is that verifies this existential quantification over properties. It is the heaviness of the object used to break the shell of the nut, or the rigidity of the stick used to pull the toy towards him. (The case in which the agent believes there is some such property, but does not know what it is, is something more sophisticated.) There are at least two kinds of discriminating evidence that would show, in context, that an agent's actions are based on such a representation of yielding or, in richer cases, causal relations.

First, when the agent knows what property of his actions explains the achievement of his goal, he will be surprised by being shown an apparently similar case in which the goal is met, following his action, yet is revealed to involve a visibly different effective property. Consider an agent who thinks, of a toy resting on a towel, that pulling the towel brings the toy close to him because the toy's movement is caused by the motion of the towel. He will be surprised if it is revealed that the toy's movement is really controlled by some pulling device beneath the surface on which the towel is resting, and the set-up is arranged so that the toy is moved whenever the experimenter sees the agent preparing to make a pulling motion.

By contrast, the agent who represents only goal-action pairs need not be surprised at all by this state of affairs. There were no commitments in the acceptance merely of goal-action pairs to any particular explanation of the movement of the toy. Correspondingly, a good experimental paradigm for detecting the difference between the two cases will fix on sensitivity to, and surprise at the absence of, the presumed causally explanatory property of the action in the case in which there is representation of yielding or of causality.[7] When the subject believes in the existence of such a property explaining

[7] The editors of this collection have drawn my attention to the fact that such a procedure for distinguishing appreciation of a causal connection, with a certain property as explanatory, as opposed to mere registration of a regularity, has actually been employed

why A-events are followed by G-events, the subject will of course also be in a position to hold that (other things equal) if the A-event had not occurred, then the G-event would not have occurred. So this account explains why the criterion that appealed to the believed-false antecedent was a promising move in the right direction.

A second kind of discriminating evidence relies on the fact that representation of the power of the relevant property P to explain G-events will not be restricted to the case of action. Moving cloths will move the objects on them even when the motion of the cloth is not caused by an intentional action. Weighty objects will crush fragile objects even when the event is not an intentional action. When there is representation of causality, the representations will not be cognitively isolated, by contrast with the goal-means pairs, which may well be so.

An intuitive mechanics is any system of mental representations and operations used by the subject, without conscious reasoning, either to predict or to explain those features of events and states affairs that involve not just their spatio-temporal properties and relations, but also such material properties and magnitudes as weight, solidity, resistance, momentum and force. In a wide range of cases, the properties the agent takes to be effective when there is representation of yielding or of causality will be properties which are attributed the same causal role in the agent's intuitive mechanics. It is at this point that action, the representation of causality and intuitive mechanics interact. By contrast, there need not be any involvement of an intuitive mechanics in the simpler case of mere representation of a goal-action pair.

In making these points, I do not imply that anyone employing the notions of yielding or causation, or using an intuitive mechanics, must explicitly know some analysis or link of these notions with that of a law. The connections may be merely tacitly known. The tacit knowledge is reflected in the commitments incurred and shown when the agent employs representations of yielding or causation.

---

by Merry Bullock, in the experiments reported in her paper 'Preschool children's understanding of causal connections', *British Journal of Developmental Psychology* 2 (1984) 139-48. See in particular her 'Surprise' condition and questions. Bullock is concerned with relations between two types of events, rather than an action type and an event type, but the rationale for the experimental paradigm is the same as in the text above.

An intuitive mechanics may also not make explicit use of the notion of a law. A very basic form of an intuitive mechanics may consist in a collection of stereotypical physically specified kinds of situation, with each kind associated with another, an instance of which follows an instance of the first kind, together with a specification of the properties or relations of the objects involved in virtue of which it so follows. This may all be done at the level of nonconceptual content. The intuitive mechanics may operate as constraining the agent's expectations of what kind of situation, specified in scenario terms, will follow a perceived situation specified in scenario terms.[8]

On the present account of what is involved in a representation of causality being involved in a subject's actions, a subject can represent causality without necessarily having the capacity to make attributions of agency to other subjects. This means that there are interesting and complex relations of this treatment to the highly illuminating discussion of the representation of causality in Susan Carey's chapter on the representation of causality in infants in her book *The Origin of Concepts*.[9] Carey argues that some of the most powerful evidence that infants represent causal notions comes from their differential sensitivity to states of affairs in which some event is explained by an agent in the situation, and those in which there is no apparent agent. The infants are not surprised when it is revealed that an agent is the source of the motion of some otherwise inert object (a bean bag). They are much more interested when the motion seems to come from a source that is known not to be capable of moving the inert object.[10] This compelling evidential state of affairs is consistent with the constitutive account of what it is to represent causality not mentioning representation of something else as an intentional agent.

---

[8] See my "Intuitive Mechanics, Psychological Reality and the Idea of a Material Object", in Spatial Representation ed. N. Eilan, R. McCarthy and B. Brewer (Oxford: Blackwell, 1993).

[9] The Origin of Concepts (New York: Oxford University Press, 2009), chapter 6, 'Representations of Cause', see especially pp.234-242.

[10] See R. Saxe, J. Tenenbaum, and S. Carey, "Secret agents: 10 and 12-month-olds infer an unseen cause of the motion of an inanimate object", Psychological Science 16 (2005) 995-1001.

I close this section with three observations on the representation of causality as discussed so far. First, we have recently been considering the correct constitutive and evidential account of the case in which tool use does involve a representation of causality. But absolutely nothing in this discussion suggests that a constitutive account of what it is to represent causality should mention tool use. We did not invoke tool use in elucidating what it is to grasp causality. On the contrary, in discussing the case, we simply connected a grasped notion of causality with tool use. We may have developed an account of one way in which a representation of causality can be involved in the psychological explanation of behaviour; but nothing in these considerations implies that there cannot be other ways, that may not mention tool use at all. In fact, there are hints of the opposite in the above discussion. An intuitive mechanics may invoke a notion of causally explanatory properties, and it seems that such an intuitive mechanics could be grasped by a subject who does not engage in tool use at all (and may from congenital paralysis be incapable of doing so).

Second, the arguments I gave earlier in this section undermine the weaker claim that tool use, even of an ingenious and creative kind, is evidence for grasp of causality. Even creative and ingenious uses of tools can be explained without any attribution of grasp of causality. When use of tools does involve some grasp of causality, the account of grasp that need to be invoked does not itself have to mention tool use. It mentions properties involved in an intuitive mechanics.

Third, and by way of transition to some wider issues, in the case in which there is in fact representation of causality involved in a subject's tool use, the account I have given involves that representation treating causality as having the very features one would include in a metaphysical account of the relation itself. It mentions such things as the implication of a causally explanatory property and connections with counterfactuals. In other cases, one would distinguish sharply between representing a property or relation, and being right about its metaphysics. Yet we moved right to the metaphysics here in the account of the case in which the representation of causality is involved in the subject's tool use. Is there any good reason for this apparent involvement of the metaphysics? If so, why? Or could the representation of causality be founded in our ability to perceive causal relations? That is the topic for the next section.

2. Is the Representation of Causality either Perception or Action-Based?

It is in the nature of some concepts that they are available to a thinker only because the thinker stands in a certain relation to the subject matter of the concepts. The existence of this phenomenon is widely acknowledged in the case of perceptual demonstratives – such as the concept *that cat*, made available on a particular occasion by perception as of a cat. There is also wide acknowledgment of the point that concepts based on perceptual recognitional capacities for objects and kinds are available to a thinker only because of the thinker's relations to those object and kinds. The phenomenon is also often acknowledged for certain concepts of conscious events and states, concepts which are available only to those who know what it would be like to experience such events and states.[11] So the question arises: is there a way of representing the relation of causation that is similarly relation-based?

There are two obvious kinds of relation of a thinker to instances of causality that might be candidate relations for making available such a putative way of representing causation. They are the relations involved in perception, and the relations involved in action. I take perception first.

Mature human perception is shot through with contents that imply causality. Anyone who sees one thing as leaning on another, pushing another, pulling it, supporting it, blocking it, is having an experience whose correctness conditions require certain causal relations to hold between the perceived objects. So it might be suggested that for a subject to appreciate her own tool use as causal is simply for her to apply the same relation to the events involved as she perceives to be instantiated when she experiences other events as causally related. The question then arises: is causation an observational concept which, like any other observational concept, can be applied sometimes on the basis of perception, and is sometimes legitimately applied on other bases too?

---

[11] For a general discussion of a range of examples of relation-based thought, and their significance, see my paper 'Relation-Based Thought, Objectivity and Disagreement' in a Special Issue on Concepts, ed. E. Lalumera, Dialectica 64 (2010) 35-56.

I suggest that causation is not such a concept, on the following grounds. Consider something for which an observational model of understanding is correct, such as depth (distance ahead of the perceiver) or objective shape. The perception of depth, shape and other observationally understood notions all conform to the following principle:

> For an observationally understood property or relation, the instantiation of the property (or relation) explains the information from which perception of the property is computed.

Any computational mechanism that computes the objective shape of an object from the shape of its retinal image will meet this condition. The objective shape certainly explains, in combination with many other conditions, the shape of its retinal image(s). The computation of depth in stereopsis from disparity between the two retinal images conforms to this principle, since the depth of the object in question explains the disparity between the images.[12] The computations of shape from motion discussed early on in the development of computational models of vision also conform to this principle.[13] Could the perception of causality conform to the same model?

It is important to distinguish between: on the one hand, correctly representing certain properties of an event and computing a representation of causality from them; and on the other, being a case in which an instance of causation is causally explaining the information from which a representation of causality is computed. The first of these does not guarantee that the second will hold. Causal relations are often computed merely from spatial, material, and temporal information. This is one of the empirical lessons of Michotte's experiments.[14] In some cases, these computations also draw on information about whether the objects in the interaction are agents with attitudes. But in none of these cases does the holding of the causal relation between the events perceived explain the information from which the representations of causation are computed.

---

[12] D. Marr, Vision (San Francisco: Freeman, 1982).

[13] S. Ullman, The Interpretation of Visual Motion (Cambridge, MA: MIT Press, 1979).

[14] A. Michotte, The Perception of Causality (London: Methuen, 1963), tr. T. & E. Miles.

It might be replied: “It’s because the effect is caused by the prior event that it occurs at all (and has the properties it does); so the perceived properties of the effect are explained by the causal relation”. This is a confused response, considered as an attempted answer to the question at issue. The effect is not explained by the causal relation, but by the prior causing event, and the prior event’s properties and relations. If the effect’s existence and instantiation of certain properties were casually explained by the causal relation’s holding, it would be explained by something that already fully necessitates its existence. Explanation never involves that.

A rather different attempt to revive the idea of causation as observationally understood might run thus. “It will be adaptive to represent causal relations correctly by and large. So it will be part of the explanation of the representation of causality being computed from spatial, temporal and material information that the causal relation holds in certain cases.” That may well be true; but it is still not an instance of the perceptual model, which requires the causal relation on a particular occasion to explain the information from which the representation is computed. That requirement on the explanation of the information is occasion-specific. The requirement is distinct from, and stronger than, conditions concerning the adaptive explanation (if such exists) of the computation of causality from properties that are for the most part causally explanatory.

A second objection to this attempt at revival starts from the point that misrepresentation is equally a form of representation. This attempt does not account for cases of perceptual misrepresentation – when, for instance, overgeneralization is adaptive, and causes representation of causal relations where there are none; or conversely fails to represent adaptively irrelevant instances of the causal relation. It is not always adaptive to represent causal relations correctly. It may be adaptive to represent a wider category of foods as dangerous than is in fact dangerous, if it is very difficult to characterize correctly the narrower, really dangerous class. We should always distinguish the constitutive question of what gives something its content from whether it is adaptive for an experience to have that content in given circumstances. Our question is the constitutive one, and whether in the case of causality the perceptual model applies.

I have been understating the case. It is not a merely contingent and a posteriori matter that the instantiation of a causal relation does not explain an experience or a

representation of causation. For the causal relation to be instantiated, there must be a law relating certain properties of the causally related events. We seem not to have any conception of how the existence of a law relating the properties of the token events in question could enter the causal explanation of a subject's impression on a particular occasion of causation.

I am inclined to conclude that although causation enters the content of many perceptual experiences, there is a deep sense in which it is not itself an observational notion, and we cannot apply to it the model of knowledge acquired by observation.

More generally, we can distinguish two directions of philosophical explanation that may hold between a concept and experiences whose content contains that concept. We can call the first direction 'the experience-to-concept direction'. When the explanation runs in the experience-to-concept direction, we individuate the concept in question in part by its role in perceptual experiences whose content contains it. For such a concept, a constitutive account of the concept's very nature involves its capacity to feature in the content of perceptual experiences. These are the genuinely observational notions.

When the explanation runs in this first, experience-to-concept, direction, we have an obligation to give a further philosophical account of what it is for an experience to have that content. We cannot just say that the experience represents a content as holding, as if the content involves some already individuated concept or content, precisely because in this first class of cases the concept has no identity independently of its ability to feature in the content of perception. The account of instance-individuation in *The Realm of Reason* was one attempt to say something about what it is for experiences to have spatial and temporal contents without presuming upon some independent account of the individuation of the concept (or content).[15]

In the second class of cases, we have philosophical explanation running in the concept-to-experience direction. In these cases, there is an independent account of the concept and its possession, and experiences represent the world as meeting a condition

[15] See my book The Realm of Reason (Oxford: Oxford University Press, 2004), pp.69-73, and T. Burge, 'Perceptual Entitlement', Philosophy and Phenomenological Research 67 (2003) 503-548.

that involves that concept. So, for instance, one may indeed see something as a telephone. But telephones do not have to look a particular way. It is also the case that one can individuate the concept *telephone*, and one can specify what it is to possess the concept, quite independently of its capacity to feature in the content of perceptual experiences.

If what I said earlier in this section about causation not fitting the observational model is correct, then the concept of causation falls into this second class, the case in which the direction of constitutive explanation runs from concept to experience. There is a widespread phenomenon of an experience absorbing the content of some concept whose nature is explained independently of its ability to feature in the content of experience. The phenomenon is present whenever we have perceptual experiences whose contents contain such concepts as *telephone*, *computer*, *bodyguard*, *food*, *metal* or myriad others.

If the nature of the concept of causation is to be elucidated by means of an explanation that runs in the concept-to-experience direction, we are thrown right back to the issue of what it is to represent a relation as causal. The fact that we can experience some events as causally related is not answering the question of the nature of our concept of causation, not even provisionally or partially.

The other salient possible resource I mentioned for a treatment of the concept of causation as relationally based is the set of relations to causal instances that are involved in action. The idea would be that, just as certain perceptual relations permit a perception-to-concept direction of explanation for observational concepts, so similarly some of the relations to causes involved in action permit what we could *pari passu* call an action-to-concept direction of explanation.

An attempt to develop such a position might start from the idea that in action you are commonly aware of that you are, for instance, raising your arm, pushing some object, rotating some objects, and so forth. In some of these cases the awareness is pure action-awareness. In others, it may involve proprioception. The relation-based theorist's initial idea may be that in all these cases, the state of affairs of which one is aware has causal implications. To think of a relation as causal, or an object as one whose properties is causally influential, is, this theorist says, to think of it as the same kind as is implicated in all of these cases of which one has an awareness, one way or another, in action.

There are two problems for such an action-oriented approach. It involves an action-to-concept direction of explanation of the nature the concept of causation. Correspondingly, it needs a constitutive account of what it is for the notion of causation to be in the content of action-awareness, just as a perceptual account of the concept of causation needs a constitutive account of what it is for the notion of causation to be in the content of perception. What, in the action case, could this account be? Here, as one might expect, the problems are structurally the mirror image of the problems for such perceptual accounts of the concept of causation. I argued that in the perceptual case, to have an observational representation of causation, it would have to be the case that the instantiation of the causal relation itself caused the information from which the representation of causation is computed. The parallel condition for action would be that an initiating motor event, or perhaps a trying, causes it to be the case that an event *a* causes an event *b*. But no initiating motor event, or trying, or any other event can cause it to be the case that *a* causes *b*. If some event is supposed to cause everything that is involved in *a*'s causing *b*, then it would have to cause the existence of a law relating some properties of *a* and *b*. That is not tenable.

Perhaps, more modestly, some event is supposed to count as causing *a*'s causing *b* simply by causing what is the difference between what holds when *a* causes *b*, and the case in which *a* does not occur – and since the existence of the law is common to these two states of affairs, no causation of a law is in question. That is indeed a less assuming position; but explaining the difference amounts then to no more than explaining the occurrence of *a*. That is fine, but it no longer gives a philosophical explanation of why the causal relation enters the content of the action awareness. It could at most explain why the event *a* enters the content of the action awareness. [16]

[16] Those who wish to pursue the fine structure of these arguments further can note that there is an argument for the case of action that mirrors the argument earlier in this section that the perceptual model of knowledge cannot be applied to the relation of causation. Just as the information from which a representation of causation is computed in the perceptual case is not itself causal, so what is produced by an initiating event in human action is not itself an instance of causation, as opposed to something that, once produced, stands in causal relations. What is computed from in the perceptual case is not causal; what is computed to in the action case is not causal.

The second problem is that simple attempts to give an action-based account of the concept of causation will not at all uniquely determine that concept. When one tries to refine simple accounts to make them more discriminating, the reference to action becomes redundant. For example, an initial problem is that all the cases in which, in action, one is aware of some property or relation with causal implications are also cases in which there is an intentional agent producing some result. Since the concept of causation applies, even in very simple everyday cases, when there is no intentional agent, the relation-based theorist has to expand the range of cases that are captured by his phrase "same kind of relation as in all these action-awareness cases".

He might try to do this by fiat, by saying instead "same kind of relation but without necessarily involving an agent". That is better, but still not wide enough. There are also everyday instances of the causal relation that do not involve in any way something that is of a sort that we can do or bring about. So how is the relation-based theorist to specify the more relaxed kind of sameness relation that, according to him, is involved in our grasp of causation? This is, incidentally, equally a challenge for someone aiming to employ the obscure notion of agent causation in an account of the concept of causation.

The only specification that I can find that would meet the need of our imagined theorist would be to something like this: "sameness in respect of the existence of a property P such that it's because F-events are P that they are followed by G-events, just as there is such a property similarly linking (say) action-types and goals". (One could also consider a variant proposal, replacing "action types and goals" by "tryings and some component of actions".) The problem of redundancy is then very apparent. This bridge from the action cases to arbitrary cases of causation is so richly structured that it already contains an explanation of what it is for events or states to stand in a causal relation. The appeal to action has become redundant. There are structurally analogous problems for the perceptual attempts.

To say that the reference to action is redundant in the account of the general, full concept is not of course to say that action, or indeed perception, may not be an important stepping stone on the way to acquisition of the general concept. We should sharply

distinguish between perception and action playing that role in acquisition and their being mentioned in a constitutive account of possession of the concept itself.

The redundancy we noted two paragraphs back is also (and relatedly to the distinction just drawn) in sharp contrast structurally with views that aim to explain the general concept *pain* in terms of a relation of identity of subjective kind of events in other subjects with the thinker's own pain-events. In that case, the relevant bridging relation to the non-local cases (the pains of others, or oneself at other times) by no means makes the reference to one's own pains explanatorily redundant in the constitutive account of the concept. It is those pains that fix the relevant subjective type, and that is far from redundant in the constitutive account. The same applies to the role of perceptual experience in the individuation of observational concepts, such as observational concepts of shape. The type of experience one enjoys when one perceives something as oval is an ineliminable element of the constitutive account of the observational shape concept *oval*. It is not merely a stepping stone in the process of acquisition of some concept whose nature can be specified without any allusion to perceptual experience.

None of this is at all to deny that the states of affairs of which one is conscious in normal action do genuinely have causal implications. We have to remember that this fact is consistent with two further points.

First, to be aware of something as F, where F has causal implications, is not the same as being aware of its implications as causal implications. Just as in perception, we distinguish perceiving objective states of affairs from having a representation of the more sophisticated notion of objectivity, so we must do something parallel for action. We must distinguish between being aware, in action, of states of affairs that have causal implications, and having some representation of the notion of causation. The most primitive kind of action awareness that involves awareness of some of one's relations to things in the world need not involve a conception of those relations as causal, even if they are causal. An agent may have an action awareness of turning a doorknob. Turning is here a causal notion. It does not follow that this agent must possess the general notion of causation – any more than one who sees something as red must have the general concept of colour.

Second, even for an agent who possesses a representation of causality, and for whom causation even enters his awareness in action, it may still be the case that the direction of philosophical explanation we discussed earlier runs concept-to-action, rather than the reverse. There is no denial here of the presence of causality in the phenomenology of action in the case of more sophisticated creatures and mature humans. The issue is the direction of explanation of what it is for the content to be there. That it is there in these cases is not in dispute.

There are some concepts plausibly individuated by their relations to action or to perception, beyond the familiar observational concepts. Amongst such concepts are are, arguably, concepts of particular action types (grasping, jumping, reaching for...), the concept of action in general, and the concept of an agent. The properties picked out by these concepts involve causation in various ways. It takes philosophical reflection, of a modest sort, to appreciate the presence of these causal elements. Everyday nonphilosophical users of the concepts of grasping, jumping, reaching for... and the rest need not be representing them consciously as causal. The everyday user grasps these notions because he thinks of a jumping as a kind of event he can perceive as a jumping, and which falls under a general category of being an action. He thinks of actions in general as events of the same relevant sort as his own actions, of which he has a distinctive awareness. He thinks of an agent as a subject of the same kind as he himself instantiates in virtue of his being able to act; and so forth. Analogous points may be made on the perceptual side about the concepts of seeing, hearing,...., the concept of perception in general, and the concept of a perceiver.[17] This is the briefest of overviews - there is of course much more to be said about the fine structure of each of these cases. The important point for present purposes is that my position is not at all that interesting concepts or notions cannot be individuated by their relations to perception and action. My position is only that the general concept of causation is not one of them. The individuation of the concepts just mentioned by their relations to action and perception works without redundancy precisely because these concepts, unlike the concept of causation, do not apply beyond the psychological domain. In considering possible bridges from action to grasp of causation in general, we saw the threat of redundancy when we

[17] See my Truly Understood (Oxford: Oxford University Press, 2008), chapters in Part II.

needed to add a clause "sameness in respect of the existence of a property P such that it's because F-events are P that they are followed by G-events..." (as above). When we are treating a concept such as that of a particular action type, of that of an agent, that does not extend (in the relevant sense) to non-psychological cases, we do not need to move into the territory where redundancy threatens when we try to build the bridge that would reach such cases.

If we think about the causal relation not by virtue of any special relation we bear to its instances in perception or action, or anything else, then it is natural to speculate that we must think of causation in a way much more closely tied to what makes it the relation it is – to the relation's own metaphysics.

3. The Elements of Metaphysics in Understanding; and Actualist Issues

If thought about causation involves some representation of the elements that make it causation, it should be possible to point to components in our understanding that correspond to those elements in the metaphysics. Without attempting a full-dress treatment of causation, I indicate here how two such elements are reflected in understanding.

Causation by a particular event *a* of a particular event *b* requires that there be some property (possibly relational) of *a* that causally explains the occurrence of *b*, plausibly by the property's featuring in some law relating it to some property of *b* (again, possibly a relational property). The recognition, in our understanding, of the existence of such a causally explanatory property is implicit in much of our causal-explanatory vocabulary. Nancy Cartwright has noted that many of our causal explanations employ 'thick' vocabulary that implies the existence of causal relations.[18] The verbs 'compresses' and 'smothers' are among her examples. I would add that in many such cases, applicability of the causal notion implies that a quite specific property or magnitude is causally explanatory. It is implied that it is the force exerted by what is doing the compressing that explains the compression. It is similarly the blocking of access to air or

[18] N. Cartwright, Hunting Causes and Using Them (Cambridge: Cambridge University Press, 2007), section 2.3.

gas in the case of smothering. Sometimes a causal expression implies that there is some such explanatory property without specifying what it is. The verb 'attracts' may be an example of this. But in all such cases, understanding the expression involves some tacit knowledge that its applicability involves the presence of a causally explanatory property or magnitude.

The second element of the metaphysics reflected in our thought is something of interest in its own right. It has proved hard to characterize accurately. The element can be introduced by noting an apparent tension in our thought about causal relations.

On the one hand, causal notions seem to be actual-world relations, in the sense that their applicability depends only upon the way the actual world is. This is a powerful general intuition that is backed in the literature by consideration of cases. On the other hand, there are examples that seem to suggest that the applicability of causal notions must depend not only on how things actually are, but also on how things are in some class of non-actual worlds. This is not just an apparent tension in our metaphysics of causation. It is also correspondingly a tension in our account of understanding, if understanding is built from materials that are in the metaphysics of the relation.

To understand the tension better, I state in turn first the case for what we can call the actualist intuitions, and then state the case for the non-actualist intuitions.

The actualist position starts from the core idea that what causes an event, and what causally explains an event, depends only on how things actually are. This actualist position readily agrees that, in many ordinary cases, when such causal relations hold, then certain counterfactuals involving a causing event, or an explaining condition, will also hold. In many ordinary cases, if the causing event had not occurred, the effect would not have occurred either. But the holding of the counterfactual is not what makes the case on of causation. Nor does the counterfactual hold in all cases of causation. Perhaps, if the causing event had not occurred, some new event would have been triggered that would equally have caused the effect in question. This by no means undermines the claim that the original event caused the effect in the actual world. Correspondingly, we do not need to know how, counterfactually, things would have been in order to know what causes

what, and what causally explains what.[19] These core intuitions are one of the factors underlying some of the detailed objections in the literature to counterfactual theories of causation.

Conversely, as one would expect if the actualist position is correct, the holding of various conditions in non-actual worlds never seems to ensure the holding of causal relations in the actual world. Suppose there is someone who intends to assassinate the President, and you frustrate his intentions by locking him in a room. If you had not locked in the would-be assassin, the President would not be alive. It does not follow that the event of your locking him in is the cause of the President's continuing to live. The causes of the President's continuing to live – the presence of oxygen, his normal bodily temperature regulation and the rest – are nothing so special. These causes of the President's continuing to live are the same in kind as for any other human not in receipt of medical attention. They seem to have nothing to do with the events involving the locking-in of the assassin.

Finally, to bolster his case the actualist may cite notions that are plausibly explained in part in terms of causal relations, and for which the actualist intuitions equally apply – as they should if his view is correct. Perception and intentional action are plausibly explicated in part in terms of causal relations. Many examples in the literature on perception and action support the view that whether someone is perceiving an object, and whether an event is an intentional action, depends only on what is actually the case, be the counterfactuals as they may.[20]

The best case for the rival position of the non-actualist rests on examples. However compelling the case for actualism just presented may sound, there are examples in which we seem to make true explanatory claims, true causal explanatory claims, and even sometimes straight causal claims - and in which counterfactuals seem to be an

---

[19] This case has been pressed by several writers, but is particularly well argued by T. Maudlin in The Metaphysics within Physics (Oxford: Oxford University Press, 2007) section 5.1.

[20] See my Holistic Explanation: Action, Space, Interpretation (Oxford: Oxford University Press, 1979), in the chapter entitled 'Deviant Causal Chains'. I think it is also the correct lesson to draw from the examples in H. Frankfurt, 'Alternate Possibilities and Moral Responsibility', Journal of Philosophy 66 (1969) 829-839.

essential element in elucidating why they are true. Jonathan Schaffer has a raft of vivid examples of this sort.[21] A detonator may work in this way: when the plunger is pushed down, a shield is removed that was previously preventing a triggering liquid from reaching the explosive. As Schaffer says, guns in fact work structurally in the same way. Yet we certainly say that pushing the plunger explains, causally explains and causes the explosion. Similarly, we say that pulling the trigger causally explains the bullet coming out of the barrel of the gun. Some theorists have, plausibly, distinguished a notion of influence as tied to causation, and separated it from counterfactuals. However helpful that distinction may be for other issues, it does not seem to help here at an intuitive level. We would normally say that pushing the plunger of the detonator influences whether there is an explosion or not. The device may be of such a kind that the way the plunger is pushed affects the size or force of the explosion. Attributions of responsibility and negligence also seem to go in step with the counterfactuals in these cases.[22]

How are these conflicting intuitions to be reconciled? Maudlin seems to do so by saying that causation is relative to a taxonomy. He says that "*Anything* that regularly results in a gun firing….counts as a cause that changes the inertial state" (164). He also says that this judgement is not reversed even if at the microlevel "the right thing to say" is that pulling the trigger does not produce the firing. Since the firing just is a sum of events at the microlevel, we seem on this view to have no outright, non-relative answer to the question "Does pulling the trigger cause the firing or not?". I want instead to offer a reconciliation that finds something right in both the actualist thesis and the apparently non-actualist examples; but what is right is very different in the two cases.

What is right in the actualist position is this: in any case in which some particular event or state of affairs can be causally explained at all, there is some causal explanation

---

[21] In his paper 'Causes need not be Physically Connected to their Effects: The Case for Negative Causation', in Contemporary Debates in Philosophy of Science ed. C. Hitchcock (Malden, MA: Blackwell, 2004). For clarity, I should emphasize that Schaffer says explicitly that he leaves open "whether causation is a purely counterfactual affair"; his position is rather that "causation has a counterfactual aspect" (both quotations from p.214).

[22] The case of negligence is particularly clear: see the discussion by H. Hart and A. Honoré, in their Causation in the Law (Oxford: Oxford University Press, 1985), p. 195.

of it that cites only actually existing conditions. You do not need to look beyond the actual world to find the explaining conditions. The explanation will involve a law, which may be outright or probabilistic, may be from a special science or from some fundamental discipline.

This claim about what is right in the actualist position should not of course be restricted, as if it were a contingent claim, to the world that is in fact actual. What is asserted here about the actual world holds for explanation in an arbitrary world. For any world *w*, in any case in which some particular event or state of affairs in *w* can be causally explained in *w* at all, there is some causal explanation of it that cites only particular conditions that hold in *w*. The explanation will involve the laws of *w*. I call this *intraworld* explanation.

What is right in the non-actualist position concerns by contrast what I will dub *comparative* explanation. If what I have said so far is right, for each event, under a given description, and state of affairs in a given world that is causally explicable at all in that world, there will be an explanation that could in principle be set out, mentioning only conditions and events in that world, an explanation of why it meets that description. It may be helpful here to imagine a page. The page sets out, for an event under a given description, the conditions concerning a given world, that explain the event, under that description, in that world. The conditions will be those that are connected, by law, with the explained event's falling under the description in question. Some have proposed richer demands that might be applied to intraworld explanation. Aronson and Fair, amongst others, proposed that genuine explanation always involves transfer of energy.[23] I am sceptical that such richer conditions are always met, or if met are philosophically explanatory, particularly when we consider the laws of such special sciences as economics. Such richer requirements need not be part of an account of intraworld explanation. The account in terms of causal laws may be enough.

Often, however, what we seek in an explanation is something comparative, something that speaks to the difference or similarity between two worlds, in particular the difference or similarity between the respective intraworld explanations of events and

[23] J. Aronson, 'On the grammar of "cause"', Synthese 22 (1971) 414-30; D. Fair,'Causation and the flow of energy' Ekenntnis 14 (1979) 219-50.

states of affairs within each of the two or more worlds in question. For example, we may wonder why the President is still alive even though there was an assassin around for a time. The explanation of why the actual world is one in which the President is alive despite the would-be assassin's intentions is that you locked the would-be assassin in a room. This is not a causal explanation of why the President continues to be alive. It is rather an explanation of the difference between the actual world and the possible world that would otherwise exist when an assassin is on the loose, when there is no one locking him up.

Comparative explanation is not causal explanation. In terms of our image of pages: comparative explanations compare pages, rather than setting out a new page. An illuminating comparison between pages should not be confused with writing out a new page. The tasks of writing out a new page correctly, and comparing different pages, are distinct.

Similarly, I suggest, in the detonator case we seek an explanation of the difference between the actual world, where there is an explosion, and the possible world (or worlds) in which the plunger is not lowered. The difference between the two cases is the plunging; but this does not imply that there is anything wrong with the explanation that mentions only the actual-world interaction between the explosive substance and its triggering substance (or electric shock, or whatever), after a shield is removed.

In comparative explanation, we explain the difference or similarity between two histories or salient possible states of the world. Often the comparative difference is difference from one of Maudlin's 'inertial states', which is why his remarks on divergence from inertial states do plausibly cover so many cases. While this comparative explanation is not itself explanation by laws within a single world, nonetheless when done correctly, it is founded in facts about intraworld causal explanation.

It follows that we have to be very careful in the use of the widely exploited phrase 'difference-making'. In citing a difference between two causal histories in two worlds, each of which cites a cause that is wholly a matter of how things are in its respective world, we are making a comparison. It does not follow that the difference we cite implies that causation in any one world is to be analyzed in terms of counterfactuals, or that actualism, or its generalization to an arbitrary world, is false.

I do not want to claim that the line drawn by ordinary language between the causal and the non-causal coincides with the line I have drawn between intraworld causal explanation and comparative, interworld explanation. It does not. Everyone would say that pressing the plunger on the detonator causes the explosion. My thesis is rather that we need to recognize the distinction between the two kinds of explanation if we are going to articulate what is right in the actualist intuitions, and if we are to articulate the methodology, goals and rationale of empirical explanation.

In making these relatively a priori points about the distinction between causal explanation and comparative explanation, I have not relied on any special epistemic relation to causation available in our thought. I have relied only reflection on the concept of causation and the conditions of its correct application in hypothetical examples. I think that although comparative explanations are important to us, and sometime sources of great insight, they must always rest on genuine intraworld explanations that respect the actualist thesis. We can never fully understand the world without having intraworld explanations that respect the actualist thesis. We are in a position to know that we need to seek them if we want that understanding. Our practice with the notion of explanation reflects this element of the metaphysics. But it is equally the case that the science of a domain, if we are to have a full understanding, must answer both questions of intraworld explanation and questions of comparative explanation.

* * * * * * * * *

If we step back and consider the issues from the standpoint of the theory of concepts in general, the preceding discussion can be seen as a case study of one particular concept, causation, for which some implicit grasp of the metaphysics is – unlike observational and other relation-based concepts – involved in possessing the concept itself. Rational practices of thought involving the concept manifest this grasp. For such concepts, there is in the nature of the case extensive overlap between the task of giving the metaphysics of the property picked out by the concept, and the task of giving an account of grasp of the concept. It should be on our agenda in philosophy to identify in

more detail the range of concepts which display such overlap, and to consider its many ramifications. [24]

[24] I started pulling together some thoughts on these issues in introducing the general discussion session at the July 2008 conference on Tool Use and Causality at Warwick University. Earlier versions of this material have also been presented to my seminar in Columbia University, to David Charles's discussion group in Oxford in March 2010, and to a conference at the Institute of Philosophy in London in June 2010. I thank David Charles, Dorothy Edgington, Michael Martin, Daniel Rothschild, Nick Shea, Michael Strevens and the editors of this volume for comments helpful in multiple ways. Though I have not had the time to develop the point further, I hope it is clear that the points and distinctions made in Section 3 can be elaborated to provide a critique of interventionist and 'difference-making' accounts if presented as constitutive theories of the nature of causation. For a treatment of the relations between causation and certain kinds of explanation that maps very neatly in many respects on to the distinction between intraworld causal explanation and interworld comparative explanation, see M. Strevens, 'The Causal and Unification Approaches to Explanation Unified - Causally', *Noûs* 38 (2004) 154-176. The notion of comparative explanation, together with the various kinds of comparison that can be made, can be used to characterize many distinctions and to explain many theses prominent in the theory of scientific explanation.