Philosophers' Imprint

VOLUME 12, NO. 19
DECEMBER 2012

# Abstraction and the Origin of General Ideas

Stephen Laurence
*University of Sheffield*

Eric Margolis
*University of British Columbia*



## 1. Introduction[1]

Given their opposition to innate ideas, philosophers in the empiricist tradition have sought to explain how the rich and multifarious representational capacities that human beings possess derive from experience. A key explanatory strategy in this tradition, tracing back at least as far as John Locke's *An Essay Concerning Human Understanding*, is to maintain that the acquisition of many of these capacities can be accounted for by a process of *abstraction*. In fact, Locke himself claims in the *Essay* that abstraction is the source of *all* general ideas (1690/1975, II, xii, §1). Although Berkeley and Hume were highly critical of Locke, abstraction as a source of generality has been a lasting theme in empiricist thought. Nearly a century after the publication of Locke's *Essay*, for example, Thomas Reid, in his *Essays on the Intellectual Powers of Man*, claims that "we cannot generalize without some degree of abstraction…" (Reid 1785/2002, p. 365). And more than a century later, Bertrand Russell remarks in *The Problems of Philosophy*: "When we see a white patch, we are acquainted, in the first instance, with the particular patch; but by seeing many white patches, we easily learn to abstract the whiteness which they all have in common, and in learning to do this we are learning to be acquainted with whiteness" (Russell 1912, p. 101).

Despite the importance of abstraction as a central empiricist strategy for explaining the origin of general ideas, it has never been clear exactly how the process of abstraction is supposed to work. There are a number of reasons for this. One is that many philosophers who have written about abstraction have been more concerned with the role of abstraction in supporting a metaphysical agenda than with the psychological details of the process of abstraction. Interestingly, philosophers have appealed to abstraction in the service of opposing metaphysical positions. Some (*e.g.*, Locke and Reid) have called on it as a means for explaining generality in a way that is consistent

1. This article was fully collaborative; the order of the authors' names is arbitrary. We would both like to thank the referees for *Philosophers' Imprint*. Eric Margolis would also like to thank Canada's Social Sciences and Humanities Research Council for supporting this research.

with broadly nominalistic scruples, while others (*e.g.*, Russell) have understood it to be an essential ingredient for making sense of realism about universals.

Another reason why the psychological details of the process of abstraction have been so unclear is that philosophers have relied on introspection as the principal source of information about the process. Conflicting opinions regarding abstraction consequently turn on divergent claims about what introspection uncovers. While Locke takes it to be evident that introspection shows that general ideas like MAN or HORSE[2] are acquired through abstraction, others, including Berkeley and Hume, claim that they don't see this at all when they look into their own minds. But even if everyone were to agree about the deliverance of introspection, that would still leave us largely in the dark about the process. From a contemporary vantage point, it is well established that much of the mind isn't accessible to introspection and that introspective reports of psychological processes aren't always trustworthy. There is little reason to think that the processes involved in abstraction should be an exception.

We suspect, however, that the most important reason why the psychological details of the process of abstraction have remained obscure is that its adherents have not appreciated the need to provide a substantive explanation of how it works. As Chomsky has emphasized, this is often the case when it comes to the mind. "One difficulty in the psychological sciences lies in the familiarity of the phenomena with which they deal. A certain intellectual effort is required to see how such phenomena can pose serious problems or call for intricate explanatory theories. One is inclined to take them for granted as necessary or somehow 'natural'" (Chomsky 2006, p. 21). Consider how Locke peppers his discussion with phrases that are meant to highlight the obviousness of his subject matter. For example: "That this is the *way*, *whereby Men first formed general* Ideas, *and general Names to them*, I think, is so evident, that there needs no other proof of it, but the considering of a Man's self, or others, and the ordinary proceedings of their Minds in Knowledge…" (1690/1975, III, iii, §9).

Perhaps it is not surprising, then, that we know so little about abstraction. But given the recurring interest in abstraction, and given the importance of general ideas in thought, philosophers clearly need an explicit framework for understanding abstraction that isn't beholden to introspection and that is open to the findings of perceptual and developmental psychology and related fields. Our aim in this paper is to provide a general framework that fills this gap and to explore some of its philosophical implications. One of our motivations is to identify the extent to which a process that is broadly like the one invoked by Locke and other philosophers can explain the acquisition of general representations.[3] We should note at the outset, though, that while this paper takes its inspiration from early philosophical discussions of abstraction, our focus is theoretical rather than historical. We are primarily interested in the explanatory benefits that can be obtained by something akin to the traditional notion of abstraction, not with the historical controversies regarding how Locke and other philosophers in the modern era are best interpreted. We'll see that there are good reasons to abandon some of the features that figured prominently in traditional accounts of abstraction — including the link between abstraction and anti-nativist views of cognitive development. Nonetheless, we believe that philosophers like Locke were right to emphasize the significance of abstraction as a means of acquiring general mental representations. Even if they were wrong

2. We take Locke's Ideas to be mental representations and will use expressions in small caps to refer to mental representations.

3. In what follows, we will occasionally make reference to the acquisition of *concepts*, where a concept is understood as a type of mental representation. However, nothing essential turns on this way of thinking about concepts. On views that take concepts to be a type of abstract object, abstraction may still be important to the acquisition of general concepts by way of mediating access to these abstracta. On such a view, our talk of acquiring concepts via abstraction should be understood in terms of acquiring general representations that have concepts as their semantic values. In any case, our focus in this paper is on the question of how general mental representations are acquired; our use of the term *concept* can be read as stipulatively referring to general mental representations.

about significant details about how the process of abstraction works, abstraction does play an important role in explaining the origins of general representations.

### 2. Some General Representations Are Innate

In Book II of the *Essay*, Locke describes the process of abstraction, claiming that abstraction is the source of all of the mind's general representations. According to Locke, abstraction is the power of mind that involves "separating [Ideas] from all other *Ideas* that accompany them in their real existence; this is called *Abstraction*. And thus all its General *Ideas* are made" (1690/1975, II, xii, §1). Locke gives several examples that are meant to illustrate the workings of abstraction. Regarding the origins of the general representation WHITE, we are told:

> … the same Colour being observed to day in Chalk or Snow, which the Mind yesterday received from Milk, it considers that Appearance alone, makes it representative of all of that kind; and having given it the name *Whiteness*, it by that sound signifies the same quality wheresoever to be imagin'd or met with; and thus Universals, whether *Ideas* or Terms, are made. (II, xi, §9)

The claim is that a general representation for a simple quality is formed by (in some sense) leaving out specific details about where and when it originated, as well as other ideas that may have initially accompanied it. Later, in Book III, Locke discusses a different kind of example — the formation of a complex idea. He suggests that children may acquire MAN by first attending to particular individuals, such as their nurse or mother, and later observing that other things resemble those individuals. This leads children to:

> … frame an *Idea*, which they find those many Particulars do partake in; and to that they give, with others, the name *Man*, for Example. And *thus they come to have a general Name*, and a general *Idea*. Wherein they make nothing new, but only leave out of the complex *Idea* they had of *Peter* and *James*, *Mary* and *Jane*, that which is peculiar to each, and retain only what is common to them all. (III, iii, §7)



Locke scholars have debated how to interpret Locke's remarks about the nature of abstraction and even whether he has a single account. This is understandable, since there is some unclarity about whether Lockean general ideas are formed by retaining the full representations associated with the particulars that an agent perceives. To some readers, it sounds like the full representations *are* retained and that abstraction involves attending to certain features as opposed to others. However, to others readers, there is the suggestion that an abstract idea may involve the construction of a new representation, one that takes some features from the representations of experienced particulars while omitting others.[4] Regardless of what the right story is about Locke, it is clear that he views abstraction as a process that is grounded in perception and that operations on the representations resulting from contact with particulars are the source of the ability to represent far more than the items that were originally perceived — not just this white paper but all white objects, not just this man but all human beings, and so on.

But how exactly can abstraction be the source of all general ideas? To see the force of this question, we need to step back and consider more carefully what input gets the process going. If abstraction is to explain the origins of all general representations, what kinds of representations can it draw upon, and how do they depict the particulars that an agent perceives? We will argue that there are four models of the representational input that are available to Locke but that none of these models can provide a satisfactory account of the origins of all general representations. The result, we will argue,

4. The difference between these two approaches is nicely summed up by the contrast between J. L. Mackie's description of abstraction as *selective attention* and Jonathan Dancy's slogan that *abstraction is subtraction* (Mackie 1976, Dancy 1987).

is that abstraction cannot plausibly be the source of all general representations and that it is highly unlikely that *any* learning process could be the source of all general representations. If an organism has any general representations at all, then, in all likelihood, some of these must be innate.

We should note at the outset that this argument is intended as an inference to the best explanation, not a proof. We do not claim that it is logically impossible for all general representations to be acquired without there being some innate general representations. Rather, our point is that non-nativist models incur prohibitive explanatory costs. Also, to simplify the discussion, we will suppose that the general representation that we are trying to understand is WHITE and that the experience from which it is abstracted is the visual perception of a snowball (or a number of snowballs). We can now rephrase the issue as identifying how the snowball is initially represented so that WHITE can be abstracted from the experience. There are four potential models to consider.

*Model 1: Individual-representations and feature-representations.* The first model takes as input a combination of individual-representations (*i.e.*, representations which function like names or demonstratives and represent individuals *qua* individuals) and representations for each of the salient features of the experienced particular. Thus the snowball might initially be represented with such representations as THAT, COLD, SPHERICAL, and SOLID.

This model faces a number of problems, but the most serious is that it simply presupposes that the process of abstraction takes as input *general representations*.[5] This clearly won't do if the goal is for abstraction to explain the acquisition of *all* general representations, as the appeal to prior general representations will lead to a regress. Moreover, *color* will undoubtedly be among the salient general features of the snowball that would comprise the input to the acquisition process, and it would presumably be the perception of its color that would support the acquisition of WHITE. But then the process of acquiring WHITE would depend upon prior representations that include, among others, the representation WHITE. The model is plainly circular. It ends up saying that WHITE is the product of a process that takes WHITE as its input.

E. J. Lowe has made a related point in a criticism of Lockean abstraction (Lowe 1995, pp. 161–2), but there is an important difference between Lowe's criticism and our own. Lowe claims that abstraction can't get off the ground if the agent doesn't have a way to single out particulars in perception prior to abstraction taking place, and he claims that this requires being able to represent each particular under a sortal that provides a principle of individuation for things of the same type. Then the problem is that abstraction can't account for where these sortal representations come from, since they are a necessary precursor for abstraction to take place. Lowe gives the example of seeing an animal. He says that you may not have to know what type of animal it is, but you have to at least represent it under the sortal ANIMAL in order to single it out from other objects.[6]

We agree with Lowe that general representations are required to get the process of abstraction going, but not for the reason that he cites. The problem isn't limited to sortal representations and isn't primarily generated by the need to represent particulars. Rather, the problem arises for any of the salient features of a perceived object that, by hypothesis, are part of the input to the process of abstraction. Whether the representations of these features provide principles of individuation is irrelevant. Now we ourselves haven't yet argued that general representations must figure in the input to the process. For the moment, it is simply an immediate consequence of the first model that we are considering that they do. Our own argument for the need for general representations will emerge through consideration of the

5. The representations of shape, temperature, etc. in the input might be nonconceptual representations, as opposed to conceptual ones. But they would be general representations all the same.

6. Though it does not affect our point, Lowe wouldn't put things exactly as we do in the text, since he is agnostic about mental representations and prefers to couch the issue in terms of representational abilities.

various options regarding the input and through highlighting the necessity of explaining how learners can selectively attend to features of stimuli. But even then there is no reason to suppose with Lowe that sortals are required to isolate objects for further attention. There is good empirical evidence for a mechanism of visual attention that is able to track objects by focusing on their spatial-temporal features, not their kind-individuating features, and that this mechanism is present early in cognitive development (Scholl 2001). So while we'll see that Lowe is right to question whether abstraction can account for all general representations, his focus on sortals is too restrictive. The fundamental problem is just that the individual-representations-and-feature-representations model assumes that there are features of a particular that initially need to be represented as such; whatever these features are, the representations of these features cannot themselves be acquired via abstraction on this model.

*Model 2: Individual-representations only.* In order to address the problem with the previous model, one might suppose instead that particulars are initially represented only by individual-representations without any general representations coming into it until abstraction has taken place.

We don't know of any traditional empiricists who have proposed a model of this kind, however, and for good reason. Individual-representations alone don't provide enough information to get the process of abstraction going. If particulars are represented simply as objects, without representing any of their features, then the input just isn't rich enough. After all, with the canonical individual-representations — demonstratives — the whole idea is that they represent their referents directly, conveying no information about what the represented objects are like. But if all the mind has to go on in representing two white objects is THIS and THAT, it would have no basis for cognitively grouping the two together, and certainly no basis for bringing them under a specific general representation such as WHITE. By limiting the initial representations to representations of the individual objects as such, the agent is effectively representationally cut off from all the features of the objects.

Suppose, however, that we overlook the question of *why* different individual-representations are grouped together and simply allow that they are. Then a number of individual-representations could be combined, yielding a representation like THIS AND THIS AND THIS (each 'THIS' referring to one of three different white snowballs). Still, the resulting representation wouldn't do, since (1) it lacks the representational breadth of WHITE (WHITE is projectible, whereas the conjoined individual representations only pick out the particulars that have been encountered) and (2) it fails to single out the relevant feature that these objects have in common (whiteness, as opposed to, for example, sphericality, coldness, snowballness, etc.). It's one thing to represent a number of perceived objects that happen to be white and quite another to represent *whiteness* (or to represent white things in general). No finite conjunction of individual-representations of white things would constitute a general representation of whiteness.

*Model 3: Trope-representations.* We are asking what the input to the process of abstraction might look like on the Lockean assumption that abstraction is the source of all general representations. A third possibility, which is seen in the work of Thomas Reid, is that it is particularized properties or abstract individuals, also known as *tropes,* that the input representations represent as such.[7] A trope is property-like in that it constitutes a feature of a particular, but unlike

7. Reid remarks that "the whiteness of the sheet of paper upon which I write cannot be the whiteness of another sheet, though both are called white", and he goes on to add that "the whiteness of this sheet is one thing, whiteness another" (Reid 1785/2002, p. 367). For Reid, there is no such thing as the universal *whiteness*. There are only the individual color tropes that are inherent in each piece of paper, each snowball, etc. Still, the appearance of generality and the prevalence of general terms in natural language are both to be explained by reference to "general conceptions". Though Reid's general conceptions are very different from Locke's general ideas, and Reid himself was a trenchant critic of Lockean ideas, our criticisms of the trope view do not presuppose that general representations are akin to Lockean ideas and apply equally to Reid's general conceptions.

a universal, it can be present only in one particular. This is not merely because no other particulars happen to have that feature but because, by its metaphysical nature, a given trope can be possessed only by a single individual—tropes aren't multiply instantiable. Returning to the snowball example, the proposal is that the input to the process of abstraction includes a representation of the snowball's whiteness, where this is taken to be a trope that is inherent to the snowball; no other particular can participate in this very whiteness. In other words, the model restricts the input to representations of individuals (tropes, as abstract particulars) but offers the hope that the agent is no longer cut off from representing the features of the particulars she perceives (tropes, as particularized properties). Features can be represented without any general representations being illicitly smuggled into the foundations of the acquisition process.

Unfortunately, appealing to tropes doesn't help. In representing the whiteness of two white objects, an agent would have to deploy two distinct representations, WHITE$_1$ and WHITE$_2$, to represent each whiteness trope as such. Because these representations are essentially of individuals (namely, the two tropes), this gives rise to much the same sort of difficulties that arose for the previous model. There is a question about why these individuals are to be grouped together and how representing them together yields a fully general representation as opposed to one that merely picks out the individuals that have been encountered thus far.

One might think that some headway can be made on the question of why tropes are grouped together by saying that the agent also represents the similarity between the tropes. In the end, this suggestion doesn't help, but it turns out to be somewhat complicated to see why. This is because there are different ways in which the similarity might itself be represented.

The simplest way would be to use a general concept of similarity, one that quantifies over the respects in which similar things are similar to one another. However, if we are looking for a process that would allow us to explain the acquisition of all general representations, a process that appeals to an existing general representation (*i.e.*, SIMILAR) in explaining the acquisition of a new general representation is prohibited. Moreover, a completely general concept of similarity would be of little use anyway. Suppose that the agent deems that the referents of WHITE$_1$ and WHITE$_2$ fall under the fully general SIMILAR (IN SOME RESPECT OR OTHER). Since any two objects are similar in infinitely many ways (Goodman 1972), this does not bring us any closer to a general representation of WHITE, and it leaves the learner unable to represent the specific respect in which these two tropes are similar. Indeed, it seems that nothing short of a general representation in terms of SIMILAR WITH RESPECT TO WHITENESS will do the trick, since any two color tropes will be *color-similar* in indefinitely many respects as well (corresponding to indefinitely many ways of partitioning the color space that include both tropes). But if we need to appeal to a general representation along the lines of SIMILAR WITH RESPECT TO WHITENESS, we might as well admit that the learner must already have the general representation WHITE. We are driven back to the problem we saw earlier. The input to abstraction would presuppose the representations whose acquisition abstraction is supposed to explain.

There is, however, another option for explaining how different tropes might be deemed similar and consequently why the individual-representations for these tropes should be grouped together cognitively. This is that the perceived similarity between these tropes is itself explained in terms of a represented trope (namely, the trope of the similarity between white$_1$ and white$_2$), so that all of the representations in play are representations of tropes as such. In this case, the agent would represent the referents of WHITE1 and WHITE2 as being related via a relational *similarity-trope* that is unique to them and that no other individuals can participate in. Let's suppose that the similarity in such cases is picked out by SIMILAR$_1$.[8] Consider now

8. This is an oversimplification, since any two color tropes will stand in indefinitely many similarity relations (just as they would have indefinitely many features in common). But we will grant this simplification for the purposes of argument.

what happens when the learner represents a third white object, say, a white sheet of paper. The learner will represent the paper as being white$_3$ (with WHITE$_3$). She might then come to compare the paper's whiteness to the other two objects and notice the similarity between the referents of WHITE$_1$ and WHITE$_3$ and between WHITE$_2$ and WHITE$_3$. To represent these similarities, she could employ representations of the relational tropes involved — SIMILAR$_2$ and SIMILAR$_3$. Now the learner has three similarity representations, but how can she recognize that these similarity relations have anything in common? If we follow the prescription that perceived similarity requires representing a trope, then we'd have to say that, for each of these pairs, there is a higher-level similarity-trope representation of the similarity between these relations (SIMILAR$_4$, SIMILAR$_5$, SIMILAR$_6$), and presumably there would have to be yet another level representing the tropes that explain the similarity among these relations, and so on (see figure 1). A model that appeals exclusively to represented tropes ends up with a regress in which each further level of represented tropes fails to get us any closer to a fully general representation of whiteness.

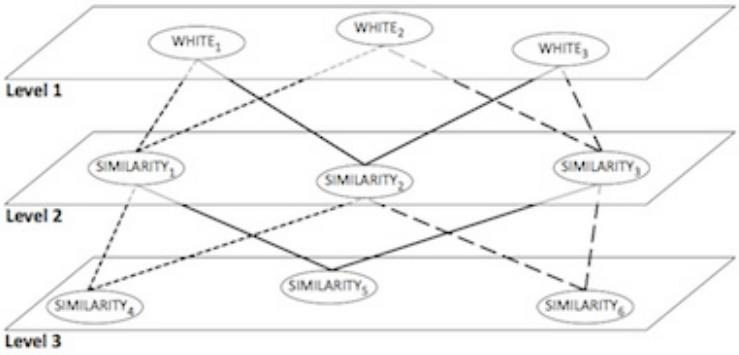


*Figure 1.* Representations of tropes of whiteness can be compared using similarity-trope representations, but further similarity-trope representations are required to compare these, generating a representational regress.

Bertrand Russell uses a related argument in the context of the purely metaphysical dispute about the status of universals. He argues that a nominalistic metaphysics that relies on resemblance between particulars isn't viable, since it would require "that these resemblances resemble each other, and thus at last we shall be forced to admit resemblance as a universal" (Russell 1912, p. 96). Russell is right that there is a need for a higher-order relation of resemblance, but in principle a nominalist could shun the universal *resemblance* by appealing to an infinite hierarchy of tropes, where the resemblance at any given level is captured by a trope that is unique to the resemblance tropes at the previous level (Campbell 1990). We aren't saying this is an especially appealing metaphysics (see Daly 1994 for criticisms), but so far as we can see, trope theorists are free to postulate an infinite hierarchy of similarity or resemblance tropes in this way if they like. In contrast, the argument that bears on the psychology of abstraction is much stronger. The reason is that an infinite hierarchy of *representations* of tropes has no psychological credibility whatsoever. Finite creatures like ourselves can't actually entertain an infinite number of representations. Yet that is exactly what we would have to do to appreciate whiteness in general if the input to abstraction is restricted to representations of tropes as such. Once again, it looks as if we need a richer source of input if we are going to explain how general representations are acquired.

*Model 4: Generality without discrete representations.* We have been looking at the various options regarding the input to the abstraction process, keeping in mind the goal of making abstraction the source of all general representations. We have ruled out a range of approaches that take some combination of representations of individuals as such and representations of features as such (models 1 and 2), and also approaches that take as input representations of particularized properties (tropes) as such (model 3). These come close to exhausting the options that ought to be considered. However, one further possibility is that more complex metaphysical entities than

individuals and features are represented in the input — something akin to events or states of affairs. In this case, the initial representations forming the input to the abstraction process might be unstructured representations that manage to pick out these more complex entities without any components representing the objects, properties, or tropes that are present in the event. For example, a snowball might be represented as being cold, spherical, *and white* but without separate representations corresponding to each of these features. The snowball's being cold, spherical, and white would be represented by a single unstructured representation (THIS-IS-COLD-SPHERICAL-WHITE), not by a structured representation composed of distinct representations capable of independently representing the object and these several features (THIS, COLD, SPHERICAL, and WHITE). In this way, WHITE wouldn't have to be a precursor to abstraction, nor would there have to be prior access to any other general representations corresponding to a particular's features.

Once again, however, psychological considerations need to be taken into account. And from a psychological perspective, such a model is not at all promising. One important feature of our systems of representation is *productivity*. The mind can represent an indefinite number of distinct combinations of features, for which the best explanation is that discrete representations are combined and recombined in accordance with a compositional semantics. However, the model under consideration (generality without discrete representations) is built on the assumption that the representational system doesn't have the compositional structure that this explanation requires. Instead, for each new combination of features attributed to an object, there would have to be a corresponding new and unique primitive representation. Unfortunately, this would require us to possess an astronomical number of primitive representations to serve as input to the process of abstraction. Since for any $n$ features there are $2^n$ possible combinations of these features, this means that with only a single object and 100 basic features and their combinations to represent — an absurdly conservative assumption — there would have to be $2^{100}$ distinct representations that could serve as input to the abstraction process. That's about 500 trillion times more representations than there have been seconds in the history of the universe (on the estimate that the universe is 20 billion years old or roughly $6.3 \times 10^{17}$ seconds). In our view, the truly staggering number of primitive representations at play is enough to undermine a model that relies wholly on unstructured representations.

But the problem with this model isn't just the sheer number of primitive representations that it would require. The real problem is with how it could account for our ability to acquire WHITE from such representations as THIS-IS-COLD-SPHERICAL-WHITE without a representational basis for homing in on just the whiteness in the experience. To mentally focus on whiteness itself would seem to require the prior ability to represent whiteness as such, but this amounts to helping ourselves to the general representation WHITE. Once again, the account in question seems to be circular: it cannot explain how the system could derive a representation of WHITE from the input without presupposing that the system already has the ability to represent whiteness.

It is easy to suppose that abstraction explains the origins of all general representations if you don't think through the psychological details. But what the failure of these four models shows is that there is a substantial burden for theorists who want to maintain this position. The principal options for getting the process of abstraction going are all problematic. They either presuppose a certain amount of general representation or are unable to support the acquisition of the target general representation. Of course, there is always the possibility that there might be some further model of how abstraction gets started that we have not considered, one that can (somehow) account for the origins of all general representations. For example, it might be said that abstraction isn't a representational process and hence that the input needn't include any representations at all, much less general ones. All that is required are causal interactions with

property instances. We grant that models of this sort aren't ruled out by anything we have said. But they are decidedly unattractive.[9] They effectively postulate mysterious neurological processes that inexplicably yield content-appropriate general representations simply on the basis of causal contact with the world. Indeed, without a well-developed account of how the process works, it is hard to see how a non-representational model of this kind is substantially different from a model that takes certain general representations to be innate and triggered by appropriate causal interactions. In any case, as we noted, we intend our argument to be an inference to the best explanation. The burden is on theorists who think that abstraction can account for the acquisition of all general representations to produce a model of abstraction that can plausibly meet this desideratum. Absent such an account, we conclude on grounds of explanatory plausibility that abstraction cannot explain the origins of all general representations and that some general representations are innate.

Locke was not alone in failing to appreciate the sorts of difficulties we have been pointing to and the need to attend to the psychological question of how abstraction works. Locke's account of abstraction was famously rejected by Berkeley and by Hume as well (largely based on Berkeley's vigorous criticism of the account). From a contemporary perspective, however, Berkeley's criticisms don't cut very deep, since an advocate of abstraction can drop the assumptions that Berkeley's criticisms turn on. And despite their spirited rejection of Lockean abstraction, the alternatives to abstraction embraced by Berkeley and Hume face much the same sorts of problems regarding the input to the process that we have been arguing Locke's account faces.

9. Much the same might be said for an Aristotelian model where sensible forms are taken to be literally transmitted from an object, through a perceiver's sense organs, into the mind. Adams (1975) succinctly describes such a view as follows: "Perception was interpreted as a transaction in which a form (the sensible form) is transmitted from the perceived object to the perceiver. … There is something (the sensible form) which literally comes into the mind from the object. This theory of perception is the basis for the Aristotelian empiricist answer to the question, how we get our ideas" (p. 73).



First consider Berkeley's criticisms, which primarily focus on Locke's construal of ideas as mental images and the view that these images can represent only what they resemble (Berkeley 1710/1975). Among other things, Berkeley points out that images are determinate in ways that bar them from achieving the generality that Locke requires. For example, you can't have an image of a generic man that represents men in general. To be recognizable as an image of a man, it would have to include specific details (*e.g.*, size, shape, color) that might be true of some men but not of others. While this may be a trenchant criticism of Locke, given the Lockean view of ideas, proponents of abstraction needn't be committed to the view that ideas are mental images or to the resemblance theory of content, not even for the representations that subserve perceptual processes. So, for contemporary theorists, these criticisms don't really identify the fundamental problems with abstraction.

Now consider Berkeley's own theory of the origins of general representations. According to Berkeley, a general representation arises as an image becomes used to represent a range of particulars that are similar to the one that the image initially picks out. In this way, a representation that is initially particular can become general. Berkeley gives the analogy of a drawing of a line in a geometrical proof. Although the line may be one inch long, it comes to represent all lines, not just one-inch lines, because the proof doesn't turn on its particular length:

> And, as that particular line becomes general by being made a sign, so the name *line*, which taken absolutely is particular, by being a sign, is made general. And as the former owes its generality, not to its being the sign of an abstract or general line, but of all particular right lines that may possibly exist, so the latter must be thought to derive its generality from the same cause, namely, the various particular lines which it indifferently denotes. (Berkeley 1710/1975, introduction, §12)

Hume described Berkeley's treatment of general representation as "one of the greatest and most valuable discoveries that have been made of late years in the republic of letters…" (1739/1978, I, i, §7).[10] But despite this high praise, it's hard to see why Berkeley's account is an improvement over Locke's. Basically, we are told that an image achieves generality because it is *used* as a general representation. An agent starts out with an image of a particular but then enlists it to reason about other things by ignoring irrelevant aspects of the image and focusing on just the relevant ones. The problem with this account becomes apparent when we ask how the mind manages to achieve this feat.

Suppose the image is of a specific snowball that a child has just seen and that she ignores the depicted shape and texture, among other things, in the service of thinking about white things in general. To do this, she needs to selectively attend to the color in the image. Yet Berkeley tells us nothing about how he proposes to account for the ability to selectively attend to certain aspects of an image while ignoring others. In order to psychologically focus one's attention on whiteness, one must, in effect, represent whiteness. But in order to do this, the options are essentially those we considered above for the Lockean account. Representing only particulars, whether concrete particulars or tropes, doesn't allow one to attend to whiteness as such. Employing a general representation of whiteness would, of course, allow one to attend to whiteness, but that would require prior possession of the general representation WHITE and hence reintroduce the problem of circularity.[11] And general representations aren't really an option for Berkeley anyway, since the whole point of his treatment of generality is that it is supposed to do away with fully abstract general ideas.

The situation for Berkeley isn't all that different from the situation for Locke, and it's the same for any theory of abstraction, or substitute process, once the need to specify the input is taken seriously. At least some general representations have to be available to get such a process going. *Some general representations have to be innate.* The moral we draw from these reflections is that the hope of providing a comprehensive theory of the origins of general representations should be abandoned. Still, a process worthy of the name *abstraction* might explain the origins of many general representations and thus be an important part of how human representational systems develop. In particular, the process of abstraction might profitably be seen as starting with relatively specific general representations as input (*e.g.*, a representation for a given shade of color or a narrowly circumscribed shape) and delivering another type of general representation as output (*e.g.*, broader color or shape representations, such as RED or TRIANGULAR). The input representations would capture the particularity of the represented qualities in experience — what is often called the fine-grainedness of perceptual experience. But the output representations would be comparatively *more general* in that they "abstract" from the particularities of the individually experienced colors, shapes, and so on. This is the idea that we plan to develop in the sections to come. It is a major departure from Locke and from other traditional accounts, but as we've seen, these accounts face insuperable difficulties in explaining how a theory of cognitive development can get by with anything less.

10. Hume's own treatment of general ideas has a strong affinity with Berkeley's, though the differences between them are worth noting. Hume doesn't follow Berkeley in claiming that we simply attend to relevant aspects of an idea and ignore others. He says, instead, that as we notice the resemblance between different objects, we give them the same name, and then later uses of the word call up related ideas. For example, the word 'triangle' may initially bring to mind an isosceles triangle, but, because of the association with other triangles, it may also bring to mind ideas of equilaterals. Reasoning about triangles in general would then amount to reasoning with an idea of a particular (say, just one isosceles triangle) and for this to be accompanied by much the same reasoning with other related ideas (other isosceles triangles, equilaterals, etc). In other words, you start with an image of an individual and consider the situation with respect to other images of other individuals that bear a resemblance to the first image.

11. Hume's account of generality is no better. Hume presupposes that people can recognize that different objects resemble one another. It's on the basis of the resemblance that the corresponding particular ideas become associated by a common word, such as 'triangle'. But Hume doesn't consider the question of how the resemblance is registered psychologically. He too would confront the same set of problematic options.

### 3. A Neo-Quinean Framework

In this section, we present a general framework for understanding abstraction. As will become clear, we think that there is a large family of related acquisition models that share important similarities and are equally deserving of the label *abstraction*. Since what is interesting from a philosophical point of view are the contours of the framework rather than the details of any particular model, our aim will be to sketch the broad outlines of the general framework. We take as our starting point W.V.O. Quine's treatment of learning in his paper "Natural Kinds" (Quine 1969). While Quine's account faces significant difficulties, it can be adapted and expanded in various ways to provide a promising basis for understanding abstraction.[12] The resulting neo-Quinean framework makes it possible to explain how abstraction can account for the origins of many general representations without falling prey to the difficulties that we presented for traditional accounts of abstraction in section 2.

Quine's discussion is couched in terms of an account of word learning. His account has three main components. First, Quine assumes that the learner can innately discriminate a range of fine-grained properties in the learning domain, for example, different shades of color in learning color words like 'white' and 'green'. These fine-grained discriminatory capacities provide the input to the process of abstraction. By building generality (albeit fine-grained generality) in from the outset in the form of general capacities for discriminating shades of color, Quine does limit the scope of abstraction. He doesn't take abstraction to explain the origin of *all* general discriminatory capacities. Nonetheless, for Quine, abstraction can explain how a general word like 'white' could be learned on the basis of the fine-grained discriminatory capacities associated with particular shades of color.

12. We should note that Quine doesn't describe himself as offering a theory of abstraction. Nonetheless, we will discuss Quine in these terms, since we take the learning process that Quine describes to be a good starting point for understanding abstraction.

The second component of Quine's account is a similarity metric. Quine assumes that the fine-grained discriminatory capacities are innately ordered in terms of similarity (an innate "spacing of qualities"), which he interprets behavioristically. "A standard of similarity is in some sense innate. This point is not against empiricism; it is a commonplace of behavioral psychology" (1969, p. 123). Quine's innate similarity metric incorporates a further element of innate generality, but it also facilitates learning, allowing the account to avoid the difficulties that earlier empiricist accounts of abstraction had in capturing the similarity in the input without general representations.

The third and final component of Quine's account is a selection process. Quine assumes that learners engage in hypothesis testing, where overt behaviors (*e.g.*, calling a color sample 'white') are selected through positive and negative feedback in accordance with the principles of conditioning. The selection process operates in tandem with the innate quality space to isolate a region within that space corresponding to a conventional term (*e.g.*, the white region within the innate similarity space). In this way, the innate similarity space can come to be partitioned in culture-specific ways.[13]

The structural features of Quine's basic account — innate fine-grained generality, an innate similarity space, and a selection process to isolate regions within that similarity space — provide the foundation we have been looking for to develop a workable theory of abstraction. However, the details of Quine's account are problematic in various ways. The most serious difficulties stem from Quine's behaviorism. Consider his explanation of the innate similarity space. Quine's account of what it is to have an innate similarity space is essentially that we are innately disposed to respond to certain stimuli in a similar manner. "A response to a red circle, if it is rewarded, will be elicited again by a pink ellipse more readily than by a blue triangle" (1969, p. 123). This explanation,

13. Quine also envisions more radical changes to the similarity space through further language learning, formal education, and the impact of science. One way to think about some of these more radical changes is that they alter the character of the similarity space by, for example, introducing new dimensions.

however, is little more than a restatement of the phenomenon to be explained. It is no better than saying that we tend to respond to certain stimuli similarly (explanandum) because we are innately disposed to respond to those stimuli similarly (explanans). True enough, but what we need to know is *why* people have the same response to the stimuli. This requires, at the very least, the outlines of a synchronic mechanism. For this reason, a better account would be one that explains the innate sense of similarity in terms of an innate computational process operating over an innate class of fine-grained representations, where features of the representations and the computational process result in representations being ordered so as to produce the similarity effects. Many computational-representational systems are possible here, and so the details are best left to empirical psychology. But we will assume that some such account of similarity is the right way to proceed, as an account that sticks purely to behavioral dispositions isn't substantive. This is the first step in developing the neo-Quinean framework for understanding abstraction. And once a computational-representational system is used to explain the similarity space, it's only natural to adopt representational versions of the other components of Quine's account — the fine-grained discriminatory capacities and the selection process. So our neo-Quinean framework will also include innate fine-grained representations and a selection process that is a computational process — one that operates over a quality space of representational states, not a field of behavioral dispositions.[14]

14. Without a representational account of the selection process, we would need an explanation of why reinforcement has its effects on overt behavior, and we would face difficulties arising from the fact that the principles of conditioning don't apply to many instances of learning, including word learning (Chomsky 1959). Citing only external factors (the impingement of stimuli, the imposition of rewards, etc.) is inadequate, since these clearly don't have the same effects on every physical system. There has to be something about the intrinsic character of the learning system that explains why conditioning shapes its responses. The best account that psychology has to offer is that, in many cases, the mechanism is deeply cognitive. It's because of the way that the contingencies of rewards and punishments are represented that the principles of conditioning have any purchase on changes in behavioral regularities (Gallistel 1990; Gallistel & Gibbon 2002).



Another aspect of Quine's account that should be addressed is the character of the selection process. Quine narrowly focuses on a single type of selection process (hypothesis testing driven by conditioning). Though a representational version of this suggestion can account for the acquisition of general perceptual representations, there are numerous possibilities for how a selection process might function, and the neo-Quinean framework should be taken to encompass the full range of such possibilities. Not all models will involve hypothesis testing, and among those that do, there will be differences in the assumptions they make. The processes involved in isolating a region in the innate quality space can range from relatively unconstrained processes (*e.g.*, summation of positive instances, or hypothesizing simple regularly shaped regions containing positive instances and excluding negative instances) to highly constrained processes (*e.g.*, where hypotheses are drawn from a highly circumscribed set or where the hypothesis space evolves in an innately specified manner). We will offer examples along these lines below. The important point for present purposes is that a wide variety of options are available for the selection process, each of which, in its own way, isolates a region of the innate quality space in response to the fine-grained representations that are taken as input.

There are also a number of other important sources of potential variation that Quine himself does not discuss but which ought to be included in the neo-Quinean framework. For example, the fine-grained representations that form the basis of abstraction needn't always be innate. In some cases, they might be learned. Likewise, the innate quality space might not be developmentally fixed. The size or dimensions of this space might be altered. Relational parameters within a quality space might also be altered, or new relations superimposed onto the space. There could also be multiple distinct quality spaces and quality spaces that stand in different relations of psychological accessibility to one another. Taken together, these and the previously mentioned sources of variation introduce considerable

flexibility within the neo-Quinean framework.[15] While we won't be able to systematically explore all these different possibilities, some will be discussed below.

In sum, the neo-Quinean framework that we are proposing takes the following form: Abstraction is a computational-representational learning process that operates over a quality space of fine-grained general representations that are ordered by a similarity metric. Abstraction involves a selection process that isolates regions of the quality space. The similarity metric needn't be simple. In fact, it might be quite complex and multifaceted. Likewise, the selection process can take many different forms. But one thing that all variations on this basic model have in common is that, by building in enough structure right from the outset (some general representations and a suitable similarity metric), the criticisms that were so damaging to traditional theories of abstraction are avoided.

If we return to the example of the general representation WHITE, there are numerous alternative models for how such a representation might be acquired in the neo-Quinean framework. One possibility, just to get the feel of the framework, would be a model much like the computational-representational analog of Quine's own account of color words. In this case, a learner comes equipped for the task with general representations for different shades of white (among other colors), as well as an innate similarity metric that organizes her color space. Then, upon encountering different instances of white things (snowballs, paper, milk, etc.), she would represent those particular shades and, through a process of positive and negative feedback, develop a representation that incorporates all of the shades that received a positive signal and none of the shades that that received a negative signal.

15. Also open to investigation is the class of representations that might be acquired by such a process. This is likely to include standard perceptual representations (*e.g.*, representations for colors, textures, and odors). But it might also include representations involved in bodily sensations (pleasure, pain, heat, etc.) and representations of cross-modal and amodal categories (*e.g.*, shape and spatial relations), among others.



This is just one example, but notice that such a model avoids the difficulties that we raised in the previous section for Locke and others, and does so specifically by abandoning the Lockean ambition of trying to explain the origins of all general representations via abstraction. Instead, the model works by supposing that some general representations are innate (*e.g.*, the fine-grained but still general representations of particular shades of white). Abstraction, according to the neo-Quinean framework, can't account for all general representations, but that is of no matter, since no framework can account for all general representations. What this new framework does do, however, is very much in the spirit of traditional theories of abstraction, in that it explains how general representations can be learned on the basis of fine-grained perceptual experience.

## 4. Implications of the Neo-Quinean Framework

We've sketched the general outlines of a workable framework for understanding abstraction, but many questions remain regarding how the framework should be developed and regarding the implications it has for philosophical theories of the mind. In this section, we make some programmatic suggestions. We offer these in the spirit of an initial exploration of a poorly understood area that is ripe for philosophical attention. But even at this early stage of inquiry, we think there are some important and perhaps surprising conclusions that can be drawn. We organize our remarks around three general issues: (1) the empiricism-nativism debate, (2) the output of the process of abstraction, and (3) human uniqueness.

### *(1) Abstraction and the Empiricism-Nativism Debate*

Abstraction has historically been seen as a distinctively empiricist acquisition process. However, we will argue that there is nothing about abstraction per se that limits it to an empiricist psychology; abstraction is equally compatible with nativist views of the mind. To see why, we need to step back and consider the characteristic features of nativism and empiricism.

Empiricists and nativists disagree about the way that psychological traits (psychological faculties, states, dispositions, etc.) are acquired.[16] Empiricists maintain that most psychological traits are acquired on the basis of a small number of general-purpose psychological systems, while nativists maintain that numerous specialized systems are needed as well. Although commentators sometimes lose sight of the point, both nativists and empiricists appeal to innate psychological traits in accounting for the acquisition of further psychological traits. For example, empiricists who are opposed to innate knowledge nonetheless suppose that basic psychological faculties for perception and memory are innate. Another common misunderstanding is the supposition that empiricists are alone in giving a large role to learning. But nativists aren't opposed to learning. They just disagree with empiricists about how learning takes place and about the systems involved. Empiricists only invoke general-purpose learning systems (*e.g.*, principles of association), while nativists also invoke specialized learning systems (*e.g.*, an innate language-acquisition device).

Far more could be said about the empiricism-nativism dispute, but even with this brief outline, it ought to be clear that abstraction isn't intrinsically empiricist; nativist versions of abstraction are also possible. Whether a given occurrence of abstraction should count as empiricist or nativist depends on how the details are filled in. The crucial factors have to do with the character of the innate similarity space and the types of selection processes that are invoked. For instance, where the selection process is domain-general and subject to few if any innate constraints, the result will be an empiricist model. But where it is domain-specific and subject to significant innate constraints, the result will be a nativist model. To illustrate that abstraction is neutral regarding the disagreement between empiricists and nativists, we return to the case of color.

Currently, there is a lively debate regarding the extent to which the acquisition of general color representations — concepts like WHITE, BLUE, and GREEN — is innately constrained. Some researchers view the learning of color categories in strongly empiricist terms. For example, in a recent review of the literature on color categorization, Regier & Kay (2009) provide a description of a view that should sound familiar:

> Debi Roberson and colleagues … concluded that there are no universal foci, that categories therefore cannot be organized around them, and that "color categories are formed from boundary demarcation based predominantly on language" … subject to the constraint of 'grouping by similarity': namely, that categories must form contiguous regions of color space. The implication is that apart from that rather loose constraint, category boundaries are determined exclusively by local linguistic convention. (Regier & Kay 2009, p. 442)

Put in these terms, Roberson et al.'s position bears a striking resemblance to Quine's (minus the behaviorism). In support of their view, Roberson et al. point to cross-cultural evidence demonstrating significant variation in color representations. For example, in an important study, Davidoff, Davies, & Roberson (1999) report that the Berinmo of Papua New Guinea use five basic color terms that crosscut the basic color terms in English; one Berinmo term covers both yellow (*i.e.*, what's called 'yellow' in English) and numerous shades that English speakers think of as green. On Roberson et al.'s account, color representations are learned by identifying different culturally salient regions within a common initial similarity space. Since there are only weak internal constraints on the learning process, color representations will vary significantly cross-culturally.

16. Historically, concerns about the nature and origins of psychological traits were often conflated with epistemological questions about justification (Cowie 1999). From a contemporary perspective, however, it is clear that justification is one thing and psychology another. In principle, a belief that requires empirical justification could be innate (*e.g.*, the belief that spiders are dangerous), while a belief that is justified a priori might not be (*e.g.*, the belief that arithmetic is incomplete).

However, other evidence suggests that the acquisition of color representations is guided by strong innate constraints. In an important early study, Bornstein, Kessen, & Weiskopf (1976) showed 4-month-old infants examples of a primary hue until the infants began to lose interest and then showed them novel instances of the same hue as well as equally novel instances that crossed a hue boundary. For example, infants were familiarized with a shade of blue (480-nm light; nm = nanometer) and subsequently shown a novel shade of blue (450-nm light) and an equally novel shade of green (510-nm light). The result was that the infants looked significantly longer at the novel shade of the new hue (green) but not at the novel shade of the old hue (blue). Franklin & Davies (2004) have recently replicated these findings using a more rigorous metric for measuring distances between stimuli. They found boundaries not only between primary color categories (blue-green) but also between secondary color categories (blue-purple). Together with the evidence of adult variability from Roberson et al., this evidence suggests that the abstraction process may begin not with an equipotent innate similarity space with no category boundaries but with a similarity space that comes with its own innately bounded regions that are modified in light of later experience. Such a model is still fully compatible with the neo-Quinean framework for abstraction. It's just a model in which the selection process is a *nativist* one, involving adjustments around innately specified boundaries in the similarity space.

Other evidence that suggests that the learning process is guided by nativist constraints points in the direction of a different sort of nativist model. For example, Terry Regier and colleagues examined color naming in 110 languages from nonindustrialized societies around the world (Regier, Kay, & Cook 2005). They found that the best examples of color terms across this diverse sample tended to cluster around the best examples of the English terms 'black', 'white', 'red', 'yellow', 'green', and 'blue'. On an empiricist model, this is highly surprising — if there are no built-in ways to group colors, why should people in every culture wind up with highly similar best examples of colors? Moreover, Regier et al. also found that the best examples of colors across these languages were more closely clustered than the center points of the color fields associated with each language's color terms. This suggests that the best examples are not simply derived from the color fields associated with the terms but rather that the best examples are primary and the color fields form around them. A natural model that takes account of this fact would be to have innate focal colors around which color fields are built through a process of learning. Or another possibility is to have *innate focal color fields*, where the best examples of colors must lie within these fields but the precise locations are open to linguistic influence and consequently subject to cross-cultural variation.

Our purpose here is not to settle the issue of whether a nativist or an empiricist model provides the best model of color concept acquisition. Rather, the illustrations are intended to show that, despite the historical affiliation between abstraction and empiricist approaches to representational-conceptual development, there is nothing in the process of abstraction that exclusively ties it to an empiricist psychology. Empiricists and nativists alike can help themselves to the process of abstraction. Theorists can even mix and match the two approaches by adopting empiricist processes for some domains of abstraction and nativist processes for others. What will determine whether the process is an empiricist or nativist one isn't merely whether abstraction takes place but rather the character of the innate structure of the similarity spaces and the innate constraints that guide the process as it unfolds.

*(2) The Output of the Process of Abstraction*

Most of the mind's representations are complex representations. They have constituent structure in accordance with the principles of compositional semantics. The concept WHITE CIRCLE, for example, is composed of the simpler concepts WHITE and CIRCLE. Primitive representations, on the other hand, do not have compositional semantic structure. They are the semantic atoms from which complex representations are built. For Locke, many of the products of abstraction

seem to have been primitive representations of this sort (*e.g.*, WHITE). This raises the interesting question of whether representations that are learned via abstraction within the neo-Quinean framework could be primitive, since it is widely assumed that primitive representations cannot be learned. As Steven Pinker describes the consensus:

> On the nurture side, empiricists tend to make do with the abstemious inventory of sensori-motor features, invoking only the process of association to build more complex ones. On the nature side, nativists argue that a larger and more abstract set of concepts, such as "cause," "number," "living thing," "exchange," "kin," and "danger," come ready-made rather than being assembled onsite.
>
> Both sides, if pressed, have to agree that the simple building blocks of cognition — like the keys on a piano, the alphabet in a typewriter, or the crayons in a box — must themselves be innate. Type on a standard typewriter all you want; though you can bang out any number of English words and sentences, you'll never see a single character of Hebrew or Tamil or Japanese. (Pinker 2007, p. 93)

According to this *building blocks model* of representational-conceptual development, the primitive representations must be innate, and the rest of our representations and concepts are assembled from these primitives. However, if abstraction offers a way to learn new primitive representations, then it argues against the building blocks model. It would show that a compelling and extremely influential view about the origins of concepts is misguided.[17]

One of the benefits of having an explicit framework for understanding abstraction is that it renders such questions tractable. Consider again the case of color representations. Given the neo-Quinean framework, we can take the input to the process of abstraction to be a set of representations of various specific shades within a similarity space (several particular shades of white, corresponding to the colors of several experienced white objects). A selection process operating on this input results in the demarcation of a field within this similarity space (a region in the color space corresponding to whiteness is delimited). Let's suppose that this process also generates a new representation, WHITE, that is linked to each of the representations in the selected field such that the activation of any element in the field brings about the activation of this new higher-level representation. Now the semantics of this higher-level representation could be handled in a number of different ways. One way would be for the content of the representation to be determined by its causal dependence on the environmental conditions that it has the function of responding to (Dretske 1995); the internal representations for specific shades would simply mediate this mind-world link between external conditions (whiteness) and the representation WHITE. Elsewhere, we have called such mediating factors *sustaining mechanisms* (Margolis 1998; Laurence & Margolis 2002). A sustaining mechanism doesn't directly determine a representation's content but indirectly makes its contribution by establishing the mind-world relation that does constitute the representation's content. On such an account, the products of the process of abstraction — representations like WHITE, CIRCULAR, SMOOTH, etc. — would have their content determined not compositionally but rather by the mind-world relations established by the sustaining mechanisms.[18] Hence abstraction *would* result in new primitives. So it looks like it is possible to learn new primitive concepts via abstraction on the neo-Quinean framework.

17. If new primitives can be learned via abstraction, this would strengthen the case that we've made elsewhere against the building blocks model (Laurence & Margolis 2002); see also Carey (2009).

18. Notice that, on this treatment, the representations of the various fine-grained shades aren't *constituents* of the general representation, unlike WHITE and CIRCLE in WHITE CIRCLE. Theorists who opt for sustaining mechanisms rather than constituency relations often do so because it weakens the relationship between the representations in the sustaining mechanism and the concept whose content is indirectly established, thus allowing for the possession of a given concept across a great deal of perceptual and cognitive variability (see Laurence & Margolis 1999, Dretske 1981, Fodor 1987).

This sort of model isn't mandatory, however, and other models that are consistent with the neo-Quinean framework would have the output of the process of abstraction be a complex representation, not a primitive. Once again, consider the case of color representations. As before, we can take the input to abstraction on the neo-Quinean framework to be a set of representations of various specific shades within a similarity space, and a selection process will result in the demarcation of a field within the similarity space. This time, though, we will suppose that this process also generates a new representation that is a disjunctive representation whose many disjuncts are just the representations that appear in the demarcated field. On this model, the semantics of the abstracted representation is plainly compositional. The content of WHITE is a function of the contents of its constituents and the compositional structure in which they inhere.

Both the compositional model and the sustaining mechanism model are compatible with the neo-Quinean framework. Abstraction can produce complex representations that *incorporate* the fine-grained representations that are the input to the process, or it can produce simple representations that are *activated by sustaining mechanisms* that incorporate the fine-grained representations. Nonetheless, several considerations suggest that the sustaining mechanism model may be preferable. One is the computational load for processes that occur at the level of the abstracted representation. If these processes have to operate on a highly structured representation and deal with each of its numerous constituents, this is likely to place a heavy processing burden on the system. On the other hand, if the processes can stick to an unstructured representation and ignore all of the structure that is inherent in its sustaining mechanism, the computational load would be considerably eased. There may also be advantages in the informational loss that is inherent to the employment of an unstructured representation. For example, if what matters in applying a learned rule is the more general category *white*, then a representation that focuses attention on just that category (and not on some particular shade) puts the emphasis just where it should be. If it doesn't matter



which precise shade is at issue, it's important not to fixate too strongly on any particular shade.

For purposes of this paper, we don't need to settle the question of whether abstracted representations are, in fact, primitive. We simply want to call attention to the fact that the neo-Quinean framework allows for the possibility that new primitives can be learned. Since standard theories of development so often suppose that new primitives must be innate, this is a possibility of considerable philosophical interest. On the model we have sketched, the neo-Quinean framework would allow us to acquire new primitive concepts, thereby increasing the combinatorial expressive power of our representational-conceptual system.

*(3) Is Abstraction Uniquely Human?*

As Locke sees things, the ability to form abstract ideas is a uniquely human capacity, one that is associated with our linguistic abilities. Locke takes the fact that animals don't use public signs to be a good indication that they aren't able to have any general ideas at all:

> … the power of *Abstracting* is not at all in them; and that the having of general *Ideas*, is that which puts a perfect distinction between Man and Brutes; and is an Excellency which the Faculties of Brutes do by no means attain to. For it is evident, we observe no foot-steps in them, of making use of general signs for universal *Ideas*; from which we have reason to imagine, that they have not the faculty of abstracting, or making general *Ideas*, since they have no use of Words, or any other general Signs. (1690/1975, II, xi, §10)

Locke is not alone in these views. Thomas Reid, for one, wholeheartedly agrees that animals "have not the powers of abstracting and generalizing; and that in this particular, Nature has made a specific difference between them and the human species" (Reid 1785/2002, p.

388). And a number of contemporary philosophers have picked up on at least the strand of Locke's view that ties the notion of a concept to language, claiming that animals lack bona fide concepts because they lack the necessary participation in a linguistic community (*e.g.*, Davidson 1975, Dummett 1994, McDowell 1994).[19] Though we can't go into the issues regarding concept possession, we do want to address the question of whether animals must lack general representations and, more significantly, whether the capacity for abstraction as understood in the neo-Quinean framework sets us apart as a species.

To begin, we should note that it is by applying general representations to individuals, and by relating one general representation to another, that agents are able to draw inferences, form expectations, and learn from experience. Most animals would not survive very long without them. A deer might manage to quench its thirst when drinking from a pool of water, but no matter how many pools it drinks from, it wouldn't be able to infer that the next pool is also able to quench its thirst. Similarly, a wildebeest that escapes a lion's attack or even multiple attacks wouldn't have the wherewithal to infer that the next lion ought to be avoided because it too is dangerous. So it is unsurprising then that psychologists have documented that general representations are widely distributed in the animal kingdom. As Richard Herrnstein notes in a review and analysis of work on animals, categorization and hence general representation "has turned up at every level of the animal kingdom where it has been competently sought" (Herrnstein 1990, p. 138).

In fact, one of the central projects in animal psychology has been to determine whether, and to what extent, different species are capable of discriminating sundry categories. Researchers routinely train animals on natural and artificial stimuli to see if they can respond not

19. Interestingly, Locke's claim that animals do not use any public signs turns out to be false. Though animals don't possess anything as rich as human natural language, there are species whose systems of communication include public signs that are under an animal's control, including nonhuman primates (Cheney & Seyfarth 1990), meerkats (Manser 2001), and even the humble chicken (Evans, Evans, & Marler 1993).



just to the items in the training regimen but also to novel instances of the category they exemplify. For instance, in a representative and now classic study, Herrnstein and his colleagues (1976) trained pigeons to distinguish pictures of trees. The subtlety involved in these discriminations is impressive, since the training set is very diverse and the contrasting stimuli are, in many respects, highly similar to exemplars of the target category (*e.g.*, while Herrnstein et al.'s pigeons had to give a positive response to a picture that showed just the top corner of a tree in the background of a scene, they had to give negative response to a picture that showed a celery stock front and center with its leaves intact). Though it is possible that the general representation TREE is innate in pigeons, other work leaves no doubt that pigeons are capable of learning new general representations. Pigeons have been trained to selectively discriminate such artificial categories as automobiles and chairs (Lazareva, Freiburger, & Wasserman 2004). They have even been trained to discriminate Monets from Picassos, and Stravinsky from Bach (Watanabe, Sakamoto, & Waikta 1995; Porter & Neuringer 1984). Our neo-Quinean framework provides a plausible account of how the underlying general representations are learned. According to this framework, the animals initially represent fine-grained (yet fully general) perceptual properties of the stimuli and, through training, come to represent broader categories in a previously established similarity space.

Quine himself, we should point out, does recognize that nonhuman animals are capable of generalizing. Unfortunately, he draws the wrong moral from this similarity between humans and animals, suggesting that our apparently sophisticated inductive abilities should be downgraded to "an animal" model.

> [O]ther animals are like man. Their expectations, if we choose so to conceptualize their avoidance movements and salivation and pressing of levers and the like, are clearly dependent on their appreciation of similarity. Or to put matters in their methodological order, these

> avoidance movements and salivation and pressing of levers and the like are typical of what we have to go on in mapping the animals' appreciation of similarity, their spacing of qualities. Induction itself is essentially only more of the same: animal expectation and habit formation. (Quine 1969, pp. 124–5)

Quine gets things exactly backwards here, attempting to reduce a sophisticated representational ability in humans to something more brute in the form of an unexplicated notion of animal expectation. Contrary to what Quine suggests, inductive inference in humans requires a substantive explanation, one that implicates representational states and processes. And, for the most part, animal expectation must be understood on the human model in terms of representational states and processes. Quine seems to be succumbing to the tendency, noted above, to be content with a superficial treatment of ordinary mental phenomena. But ordinary mental phenomena, whether in humans or in animals, mask a great deal of complexity that our explanations need to register and do justice to.

In any event, humans are by no means special in their ability to represent general categories, nor, in all likelihood, to arrive at them via abstraction. Of course, this doesn't mean that animals are capable of developing the very same general representations as humans. It ought to be clear enough that humans can develop a large assortment of representations that are unavailable to other animals. In some cases, these may be representations that do indeed require natural language, since they depend upon culturally acquired information that cannot be conveyed in any other way. In other cases, they may be representations that are grounded in domain-specific representational systems that are themselves unique to the human lineage. Regardless, it shouldn't be controversial that general representations aren't all on a par. It's one thing to have a general representation like WHITE and quite another to have a general representation like PROTON.



A natural question, at this point, is to ask what *other* types of processes, besides neo-Quinean abstraction, might support the acquisition of general representations. This question is closely connected with a number of important philosophical issues, including the influence of language on thought, the innate structure of the mind, the origins of human creativity, and the nature of theory change in science. However these are to be settled, in our view, there is no one key acquisition system responsible for representational-conceptual development; human representational and conceptual systems stem from a highly varied collection of systems of acquisition. Likewise, the difference between human and animal minds does not depend on a single powerful source from which all uniquely human representations derive but instead depends on an eclectic potpourri of sources.

Just to give a flavor of this diversity, we will mention two proposals about how humans are able to acquire certain representations via cognitive resources that animals lack. The first is Susan Carey's proposal that many concepts can only be learned via a process she refers to as *bootstrapping* (Carey 2009). Bootstrapping occurs when an agent relies on an uninterpreted or partially interpreted symbol system whose symbols act as placeholders for the representations to be learned. Interpreting the system is achieved through what Carey calls *modeling processes*. These typically involve drawing an analogy between two systems of representation, the uninterpreted system and a system that already has some meaning for the learner. Carey's flagship example of bootstrapping is an account of how children learn the positive integers. On this account, children first have to learn the counting procedure as a meaningless routine and also have to directly pick up on the meanings of the first few count terms. Then, after a protracted period and much effort, children come to see an analogy between the cognitive models they use in connection with the first few count terms and what happens with the sequence in the count list. The idea of *next word* is mapped on to the idea of *adding a single individual to a set*. Carey suggests that animals aren't capable of learning in this way, since they lack the ability to work with

uninterpreted symbol systems and engage in the modeling processes that render them meaningful. If she is right, bootstrapping may be an important part of the explanation of why we human beings have a conceptual system whose expressive power far exceeds what is found elsewhere.

The other proposal we wish to mention is one that we have developed in previous work (Margolis 1998, Laurence & Margolis 2002). On this approach, some concepts depend upon an innate template that underlies the acquisition of a range of concepts in a given domain. One model that illustrates this approach has it that human beings have a template for animals or living kinds that contains slots for information regarding properties that are highly indicative of kind-membership — shape, color markings, characteristic motion, etc. When a learner confronts a new type of animal, the information required by the template is associated with a new representation whose processing is constrained by a disposition to treat kind-membership as a matter of having an underlying nature that is responsible for the kind's more accessible properties. We've argued that together these components can establish the mind-world causal relation that is constitutive of a concept's content according to an information-based semantics approach. A similar account can be developed for artifact concepts. In this case, the constraint on processing is perhaps a disposition to defer to the creator's intent regarding issues of kind-membership (Bloom 1996). So another way that a general concept might be acquired is for this type of cognitive machinery to be engaged when a learner sees a new item that falls under the purview of an innate template. And while animals may share some of the cognitive machinery that supports concept acquisition via innate templates, it is doubtful that they have the very same templates or all of the cognitive dispositions that turn our templates into the many natural kind and artifact concepts that occupy much of human thought.

A lot more could be said about the neo-Quinean framework, but we hope that these brief remarks indicate that its treatment of abstraction has important philosophical implications. It speaks to such questions as how to understand the empiricism-nativism dispute, what kind of structure concepts have, and whether humans have unique concept-forming abilities. But most importantly, the neo-Quinean framework offers an explicit treatment of an otherwise mysterious process, and because of this explicitness, it allows for the formulation of a range of realistic possibilities regarding concept acquisition.

## 5. Conclusion

One of the central projects in the philosophy of mind is to explain the origins of our representational capacities. The aim of this paper has been to clarify one important part of the explanation — the process of abstraction — by providing an explicit framework for understanding how it works. Just as Locke supposed, general representations can be learned via a process that begins with fine-grained experience that arises through contact with particulars. Abstraction can still explain the acquisition of representations with greater generality from more fine-grained ones, and it can still explain the acquisition of a broad range of different kinds of general representations. However, on the neo-Quinean framework that we have presented, the process of abstraction differs from the traditional empiricist picture in a number of important respects. Unlike the traditional notion of abstraction, it is perfectly consistent with a nativist psychology and applies to humans and animals alike. But these departures from the traditional empiricist picture are in no way deficits of the neo-Quinean framework. They are advantages, giving the framework greater flexibility and broader applicability. Perhaps the most significant departure from the traditional empiricist picture is that the neo-Quinean framework requires a certain amount of general representation to be present from the start, so it cannot explain the acquisition of all general representations. But this is not a deficit of the framework either, since no account can do that. We conclude that while abstraction cannot be the whole story about the origin of general representations, it is nonetheless one central and important part of the story.

**Bibliography**


Adams, R. M. (1975). Where do our ideas come from? — Descartes vs. Locke. In S. Stich (ed.), *Innate Ideas*. Berkeley, CA: University of California Press.

Berkeley, G. (1710/1975). *A Treatise Concerning the Principles of Human Knowledge*. In M. R. Ayers (ed.), G. Berkeley, *Philosophical Works*. Totowa, NJ: Rowman & Littlefield.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, 60.1, 1–29.

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 2.1, 115–129.

Campbell, K. (1990). *Abstract Particulars*. Oxford: Blackwell Publishers.

Carey, S. (2009). *The Origins of Concepts*. Oxford: Oxford University Press.

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Chomsky, N. (1959). A review of B. F. Skinner's *Verbal Behavior*. *Language*, 35.1, 26–58.

Chomsky, N. (2006). *Language and Mind*, 3rd edition. New York: Cambridge University Press.

Cowie, F. (1999). *What's Within?: Nativism Reconsidered*. New York: Oxford University Press.

Daly, C. (1994). Tropes. *Proceedings of the Aristotelian Society*, 94, 253–61.

Dancy, J. (1987). *Berkeley: An Introduction*. Oxford: Blackwell Publishers.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398.6724, 203–4.

Davidson, D. (1975). Thought and talk. In his *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dummett, M. (1994). *Origins of Analytical Philosophy*. Cambridge, MA: Harvard University Press.

Evans, C. S., Evans, L., & Marler, P. (1993). On the meaning of alarm calls: functional reference in an avian vocal system. *Animal Behaviour*, 46.1, 23–38.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Franklin, A., & Davies, I. R. L. (2004). New evidence for infant colour categories. *British Journal of Developmental Psychology*, 22.3, 349–377.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Gallistel, C. R., & Gibbon, J. (2002). *The Symbolic Foundations of Conditioned Behavior*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and Projects*. New York: Bobbs-Merrill.

Herrnstein, R. J., Loveland, D. H., & Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2.4, 285–302.

Herrnstein, R. J. (1990). Levels of stimulus control: A functional approach. *Cognition*, 37.1–2, 133–66.

Hume, D. (1739/1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (eds.), *Concepts: Core Readings*. Cambridge, MA: Bradford Books/MIT Press.

Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, 86.1, 25–55.

Lazareva, O. F., Freiburger, K. L, & Wasserman, E. A. (2004). Pigeons concurrently categorize photographs at both basic and superordinate levels. *Psychonomic Bulletin & Review*, 11.6, 1111–7.

Locke, J. (1690/1975). *An Essay Concerning Human Understanding*. Edited by P. H. Nidditch. Oxford: Oxford University Press.

Lowe, E. J. (1995). *Locke on Human Understanding*. London: Routledge.

Mackie, J.L. (1976). *Problems from Locke*. Oxford: Oxford University Press.

Manser, M.B. (2001). The acoustic structure of suricates' alarm calls varies with predator type and the level of response urgency. *Proceedings of the Royal Society B,* 268.1483, 2315–24.

Margolis, E. (1998). How to acquire a concept. *Mind & Language*, 13.3, 347–69.

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. New York: Allen Lane.

Porter, D., & Neuringer, A. (1984). Music discrimination by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 10.2, 138–48.

Quine, W.V.O. (1969). Natural kinds. In his *Ontological Relativity & Other Essays*. New York: Columbia University Press.

Regier, T., Kay, P., & Cook, R.S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences of the United States of America*, 102.23, 8386–91.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, 13.10, 439–46.

Reid, T. (1785/2002). *Essays on the Intellectual Powers of Man*. Edited by D. Brookes & K. Haakonssen. Edinburgh: Edinburgh University Press.

Russell, B. (1912). *The Problems of Philosophy.* Oxford: Oxford University Press.

Scholl, B.J. (2001). Objects and attention: The state of the art. *Cognition*, 80.1–2, 1–46.

Watanabe, S., Sakamoto, J., & Wakita, M. (1995). Pigeons' discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior*, 63.2, 165–74.