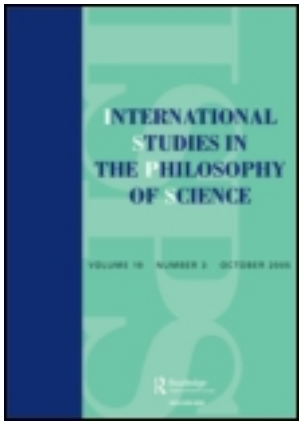

International Studies in the Philosophy of Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cisp20

Refining the Inferential Model of Scientific Understanding

Mark Newman

Published online: 12 Nov 2013.

To cite this article: Mark Newman (2013) Refining the Inferential Model of Scientific Understanding, International Studies in the Philosophy of Science, 27:2, 173-197

To link to this article: http://dx.doi.org/10.1080/02698595.2013.813253

International Studies in the Philosophy of Science, 2013
Vol. 27, No. 2, 173–197, http://dx.doi.org/10.1080/02698595.2013.813253

Routledge
Taylor & Francis Group

Refining the Inferential Model of Scientific Understanding

Mark Newman

In this article, I use a mental models computational account of representation to illustrate some details of my previously presented inferential model of scientific understanding. The hope is to shed some light on possible mechanisms behind the notion of scientific understanding. I argue that if mental models are a plausible approach to modelling cognition, then understanding can best be seen as the coupling of specific rules. I present our beliefs as 'ordinary' conditional rules, and the coupling process as one where the consequent of one ordinary rule (OR) matches and activates the antecedent of the rule to which it is coupled in virtue of the activation of an intermediate 'inference' rule. I argue that on this approach knowledge *of an explanation is the activation of ORs in a cognitive hierarchy, while* understanding *is achieved when those activated ORs are also coupled via correct inference rules. I do not directly address issues regarding the plausibility of mental models themselves. This article should therefore be seen as an exercise in refining the inferential model within an already presupposed computational setting, not one of arguing for the psychological adequacy of computational approaches.*

1. Introduction

What does it mean to *understand* a scientific explanation, rather than merely *know* it? This is a question currently receiving a good deal of attention from epistemologists and philosophers of science (de Regt, Leonelli, and Eigner 2009; de Regt 2013). There *does* seem to be a difference: for instance, Cushing (1991) argues that although one can derive Kepler's first law[1] from Newton's laws of motion, one really cannot *understand* why a planet takes a plane curve orbit unless one appeals to the causal explanation provided by Einstein's general theory of relativity. Cushing thinks causal explanations are required for understanding. For other authors, the key to understanding lies in a different property, such as the ability to manipulate variables

Mark Newman is at the Department of Philosophy, Rhodes College. Correspondence to: Department of Philosophy, Rhodes College, 2000 North Parkway, Memphis, TN 38112-1690, USA. E-mail: newmanm@rhodes.edu

in an explanation (Grimm 2006, 2010), or to build models based on explanatory knowledge (de Regt 2009). Although their accounts are helpful, none of these authors has provided a theory that clearly reveals the underlying *psychological* reasons we see a difference between explanations which provide understanding and those where we merely come to know that something is the case without understanding it.

In a previous article (Newman 2012), I tried to explain this psychological difference between knowing an explanation and understanding it with my inferential model of scientific understanding. The focus of attention was on understanding the explanation of a phenomenon, rather than understanding a model, theory, or some other object. I argued that if we take seriously the work done by cognitive psychologists it becomes apparent that understanding a scientific explanation is more akin to comprehending a story than either coming to know some sequence of facts or being able to solve problems using a theory. My goal in this article is to investigate a possible way in which one might model the cognitive processes involved in scientific comprehension, and draw some conclusions based on that investigation. Still focusing only on the explanation of some phenomenon, I propose a cognitive rule-based model, which, in contrast to current accounts of understanding, if correct, would be able to distinguish the specific cognitive mechanisms involved in the process of coming to understand something. I do not argue that a computational approach in general is the correct way to model cognition, but that if it turns out to be close, then understanding is a cognitive state that can best be understood as the coupling of specific kinds of rules. Hence, I *instrumentally* present our beliefs as conditional rules, and the coupling process as one where the consequent of one rule matches and activates the antecedent of the rule to which it is coupled. The activation of each rule requires the system break a 'cognitive threshold' set for that rule. The matter of how a threshold is established is mysterious, and I make some suggestions on how we might better understand this important computational mechanism. The important point is that on this computational approach, scientific understanding can be represented as the coupling of rules that represent our beliefs. That itself is enlightening. Empirical studies may help reveal whether the model is sustainable.

Before we get started, a word is in order about the overarching motivation for a rule-based mental model approach to representing the mind. One benefit of a rule-based account is that it maintains a commitment to representing the mind's functions in propositional terms, and hence accords nicely with our standard philosophical frameworks for describing knowledge states. This is not to say that our representations *have* to be given in sentential form, but by using a rule-based approach we are able to deliberate intelligibly about theories of knowledge and the nature of understanding. This may include, for example, considering whether our accounts of knowledge and understanding ought to be internalist or externalist. Alternative accounts that appeal to lower level mechanisms, such as Churchland's (1989) *prototype activation model*, cannot achieve this because they work in a domain below that where concepts like justification, belief, and truth can be clearly delineated.

A second reason for adopting mental models is the empirical inadequacies of neural network approaches. Most important (and this is reflected in Churchland's approach), neural networks have successfully been used to model low-level psychological processes like perception, memory, and categorization, but they leave entirely untouched important phenomena such as inference-making, and problem solving.[2] It seems intuitive that understanding is, if not constituted by then at least intimately connected to, these latter two processes.

A third reason supporting my selection of mental models is that even if neural network accounts come to provide a preferable overarching framework for representing mind, it seems increasingly likely the two approaches can be synthesized with theoretical neuroscience approaches. Such a synthesis is controversial but recent work indicates this may be possible. Scientists like Chris Eliasmith, Randy O'Reilly, and Terry Sejnowski certainly seem to be headed in this unifying direction (Thagard 2010).

Given these reasons, I think it perfectly legitimate to follow the rule-based mental models approach I adopt in this article. In section 2, I use an example from kinetic theory to illustrate precisely how the inferential model of scientific understanding works. In section 3, I explain the rule-based account of mental models. In section 4, the rule-based account is then used to illustrate how a cognitive system builds a situation model. In sections 5 and 6, I use that implementation to precisely tease apart our notions of knowledge and understanding. Potential objections are considered in section 7.

2. Scientific Knowledge and Understanding

Here I want to reiterate briefly an argument I have made elsewhere for thinking that scientific knowledge is different from scientific understanding.[3] To do this I will use a familiar example of a scientific explanation: the kinetic theory's explanation of the temperature of an ideal gas.[4]

(1) We assume a gas is made up of tiny, spherical, rapidly moving, perfectly elastic, molecules that have no extension and do not collide with one another.
(2) In a closed square container of length L and cross-sectional area A the pressure exerted on one surface S is due to the impacts of the gas molecules.
(3) The motion of any given molecule has x, y, and z components, but we assume the y and z components remain unaltered, whereas upon impact the x component reverses from $+v_x$ to $-v_x$.
(4) Each impact *therefore* produces a change of momentum $\Delta p_x = 2mv_x$.
(5) Each molecule traverses the distance from one side of the container and back again in a time $\Delta t = 2L/v_x$. This is the time between its collisions on the surface S.
(6) The reciprocal of this is the number of collisions per second: $1/\Delta t = v_x/2L$.
(7) *So* the rate of change of momentum is $\Delta p_x/\Delta t = 2mv_x/2L/v_x = mv_x^2/L$.
(8) *From Newton's second law* this is the average force exerted by a molecule on surface S.

(9) Assuming all N number of molecules in the container behave the same way, the total average force on S is $F_{\text{av}} = Nmv^2/L$.
(10) Pressure is force per unit area.
(11) *Thus*, $P = Nmv^2/LA$.
(12) The total speed squared of a molecule equals the sum of the squares of its components $v^2 = v_x^2 + v_y^2 + v_z^2$.
(13) Assume all molecules move in all directions with equal likelihood $v_x^2 = v_y^2 = v_z^2$.
(14) *Thus* $v_x^2 = \frac{1}{3}v^2$.
(15) Assuming that we can use v_{av}^2 as the average velocity of any molecule, *this entails* $P = Nmv_{\text{av}}^2/3LA$.
(16) Volume is $V = LA$.
(17) *Thus* we get $P = Nmv_{\text{av}}^2/3\,V$.
(18) *Since* Nm is total mass of the gas, the density $\rho = Nm/V$ and $P = \frac{1}{3}\rho v_{\text{av}}^2$.
(19) *This entails* $PV = \frac{1}{3}Nmv_{\text{av}}^2$.
(20) The translational kinetic energy (KE) of an object is $\frac{1}{2}mv^2$.
(21) *Thus* we get $PV = \frac{2}{3}N\text{KE}_{\text{av}}$ so pressure depends on the number of molecules per unit volume as well as their average translational KE.
(22) Importantly the ideal gas law tells us that $PV = NRT$, where R is a constant of proportionality and T is temperature.
(23) *Thus*, $\frac{2}{3}N\text{KE}_{\text{av}} = NRT$.
(24) *This means* temperature of an ideal gas is proportional to the average translational energy of its molecules.
(25) *So*, kinetic theory's explanation of the macroscopic property of temperature of a gas is explained in terms of the motion of its micro-constituents.

According to cognitive psychologists[5] when we are given an explanation like this we build a 'mental model'. For this example we actually build something more specific: a *situation model*—a representation of the situation described in the explanation. The process of encoding this information into a representation (the model) requires significant cognitive resources. For example, when we encode that molecules bounce around in a container, or that their impacts on a surface generates pressure, it requires we draw on and use a substantial amount of background information. For example, we need to draw on the meaning of terms like 'molecule', 'bounce', 'impact', 'surface', etc. These are difficult *referential inferences* regarding the meaning of terms in the explanation—we interpret the meaning of each statement stepwise in building our model. As we go through this process of selecting appropriate concepts to insert into the model, we also sequence them in a chain that mimics the explanation given. This process involves further, similar mechanisms that aid in building the model. Once we have a fairly detailed account that reflects the explanation we can be said to *know the explanation*.

However, it seems obvious that to *understand* the explanation, not merely know it, requires something more from us, cognitively speaking. It requires something like making connections between components of the situation model that goes further than sequencing them in a chain.[6] It is not merely a matter of making sense of each

expression in the story—it requires we put them all together in a coherent manner by connecting each segment in the correct way.

We should therefore like to know what it is to make these connections. What is it to construct a mental model where we link our explanations together in a way that makes sense to us—a way which reflects actually understanding the explanation, not just knowing it? If we look at the example, the difference between knowing the explanation (in something like a memorization sense) and understanding it lies in the *explanatory connections* (italicized above) between its components: it is *because* molecules generate an impact when colliding with surface *S* that they create pressure. It is *because* this pressure over the volume is proportional to temperature that we think it explains the temperature of the gas etc. But what are these explanatory connections?

I think they must be *inferential* connections because we use them to make inferences as we go through the explanation. This does not seem controversial. For example, look at steps 7–8: 'the rate of change of momentum is $\Delta p_x/\Delta t = 2mv_x/2L/v_x = mv_x^2/L$. *From Newton's second law* this is the average force exerted by a molecule on surface *S*'. Here it is necessary for us to *know* what the equation in line 7 means. We must also *know* that Newton's second law says $F = ma$, and use it to *infer* that mv_x^2/L is equivalent to ma. This inference provides us with knowledge of the average force of a molecule on *S*. A further inference from 8, 9, and 10 to 11 is required before we can be said to know why the rate of change of momentum for all molecules will get us to a characterization of pressure for a gas.

So the reasons a gas is generating pressure from its impacts is because it obeys Newton's second law. Even assuming we *know* this, it does not guarantee we will understand the entire explanation of temperature. We require an important conceptual connection between each of the steps in the explanation. We need to make all the appropriate inferences in order to reach step 25. Importantly, we have to combine the step 11 inference about pressure with our knowledge of translational KE (step 20) to get us to step 21. This latter inference helps provide adequate reasons for the explanation of temperature, and if it is right, we not only know the explanation, we also understand it. In this case the reasons for concluding temperature is just average translational molecular motion are causal—the pressure is due to impact forces. Thus, our understanding of temperature just is a result of knowing the *causal properties* of molecules and making inferences based on that knowledge.[7]

If this simple inferential account is correct, and I think something like it is, then we can now glimpse the difference between knowledge and understanding. Understanding this case relies on our *inferring* that molecules in motion generate temperature *because* they generate impact forces. If we merely knew that molecules in motion are responsible for temperature, and we did not know that this is *because* they generate pressure and that this is *because* of Newton's second law, then we would not really understand the explanation. We would not have made adequate *conceptual connections* between the properties of molecules, pressure, force, volume, mass, velocity, etc. We must make these inferences if we are to go beyond merely knowing *that* kinetic translational motion of molecules explains the temperature of a gas and achieve understanding of *why* this is so.

In this case understanding requires we possess the appropriate *causal* knowledge that enables us to build a situation model and make the necessary inferences: some properties of entities, events, etc. (impacts in this case) are such that they lead to or entail other properties of entities, events, etc. (pressure). So, the idea is that we can understand something when we make an explanatory connection, and that requires we know the properties responsible for making that connection the one that it is. We understand why molecules in motion explain gas temperature because we infer that it is the mechanical properties of massive objects impacting surfaces that they confer mechanical forces on objects around them. It is these mechanical 'bangings' that are responsible for the force on the surface *S*, and it is these that are also responsible for the temperature of the gas.

Let me generalize this idea regarding the nature of causal understanding: *for an agent A to causally understand an explanation of some phenomenon P, A must generate inferential knowledge of the reasons that are causally responsible for the cause C being the cause of P that it is. The agent therefore has to make inferences to the properties of C that are responsible for it causing P.*

Not all scientific understanding is causal of course, some of it perhaps taking the form of logical or probabilistic inference. We can therefore make the idea more general: *for an agent A to understand an explanation of some phenomenon P, A must generate inferential knowledge of the reasons that are causally, logically, or probabilistically responsible for C being responsible for P.*

I would further add that these inferences must be the *correct* inferences, else *A* simply *mis*understands *P*. They must also actually be made by *A*, or else *A* acquires only knowledge not understanding. This whole process is manifested in *A* constructing a mental model of the situation *P*, where there are inferential connections between the relevant components of the model, rather than inferential gaps, which would reflect mere knowledge of *P*. If this idea is along the right lines, it should enable us to begin delineating knowledge from understanding on any conceptual framework approach, be it mental models or neural networks.

Before closing this section I would like to address a concern that may arise with the upcoming transition to section 3. There I will unpack details behind a particular approach to mental representation, then go on to distinguish understanding from knowledge using that model.[8] Yet, at this point it might seem I am failing to think broadly enough about the concept of understanding and its role in relation to knowledge and representation. The idea is that while I am about to dive deep into representational issues, I have not yet given a convincing picture of what understanding overall amounts to—that I am losing the forest for the trees. One incarnation of this objection is to point out I have so far said nothing of the 'sense' of understanding and how it relates to knowledge.

In response, I will make three brief points. First, the above is a highly condensed summary of the arguments made in Newman (2012), where I think the reader can find a more complete analysis of the difference between knowledge and understanding, as well as of their contrast with other relevant issues, such as problem-solving abilities. Second, I am focused entirely on the issue of *scientific* understanding, and although

there is no doubt more to be said about the broader picture even there, we should not confuse that question with the much larger, and potentially overwhelming problem of understanding 'in general'. Third, the 'sense' of understanding we might experience, even within the sciences, is highly suspect as a phenomenal marker or indicator of genuine understanding. Trout (2002, 2005), de Regt (2004), and Grimm (2009) have written at length on that issue, and given the amount of controversy surrounding our feeling of understanding, I cannot see further discussion here serving our purpose well. We are after all concerned with the separate issue of outlining what might be a mechanism behind understanding itself.

3. Rule-based Mental Models

What we have so far is the barest sketch of a theory of understanding based on empirical studies. In this section, I describe the framework of mental models in terms of rule-based computation. In the next section, I use this framework with our example to analyse in far greater detail the potential difference between scientific knowledge and understanding.

On the account I am following,[9] a mental model **M** is a kind of mental representation that is used to model the properties, relations, and processes we perceive around us. We can use rules to show how we build a mental model of a container of an ideal gas **MG**. Rules are taken as the basic building blocks of all mental representations, and when they are activated at different levels of generality or specificity they form a hierarchy. A mental model is a specific activation of a complex interrelated hierarchy of condition-action rules **M**: $\{C_1, C_2, \ldots C_r\}$, each rule taking the form of an 'if–then' conditional. The rules can each have multiple conditions $C_1, C_2, \ldots C_r$ and an action A: $\mathbf{C} = \{C_1, C_2, \ldots C_r/A\}$. When a model is constructed it is a state of the hierarchy of rules, which is manifested by our activating a particular sub-network. The idea behind this approach is that the hierarchy undergoes updating of rule structure and rule strength with time-step execution cycles, otherwise known as learning.

The rules comprising the network out of which our models are constructed have different properties. Some rules are diachronic, while others are synchronic. The synchronic rules are useful for identifying (categorizing) what we are modelling, so they can be used to atemporally characterize our concept of a gas. For example, 'if X is composed of perfectly elastic molecules bouncing around without impediment, etc., then X is an ideal gas'. This rule takes the form as above, $\mathbf{C} = \{C_1, C_2, \ldots C_r/A\}$.[10] Synchronic rules also activate *associated rules* forming activations of conceptually related rules. For example, 'if X is a gas, activate the "molecule" concept' and 'if X is a gas, activate the "air" concept'.

Diachronic rules on the other hand are not concerned with categorization or association, but with prediction and action commands. They tell us what to expect in future states of the model and what to do in response to a stimulus. These temporal rules therefore tell us predictive things like, 'if the molecules of a gas lose average translational velocity, then the gas will decrease in temperature', and they provide action

commands such as 'if you see a drop in temperature, turn up the heat source'. Diachronic rules have the same formal structure as their synchronic counterparts.

A system can fire multiple sets of competing rules at a single time, and thus possesses considerable parallelism—an advantage the connectionist account had over traditional linear or serial processing accounts from the modelling paradigm. As a consequence, learning new rules can be characterized in this framework as the outcome of multiple competing rules battling it out for dominance, the new rule dominating in this competition and winning the right to represent the environment. For any single rule, the strength it brings to a competition can vary since rules are strengthened or weakened depending on their success at achieving the system's desired goals, and the strength of a connection reflects its probability of firing.

For example, take the rule 'if X has molecules, then X is a gas'. If this wins out in a situation where one is indeed working with a gas, then it will likely receive reinforcement from the environment. On the other hand if one is instead working with a solid object, then failed future inductions on the object's behaviour will cause a weakening of that rule. This rule is also rather naïve, so the system will cluster it with other rules, such as 'if X has point molecules, then X is an ideal gas' and 'if X has translational molecular motion, then X is an ideal gas' to provide a cluster of rules that together can be used to identify the object and make associations and prescriptions based on it. In a basic system the strength of a 'bid' to represent the environment made by a rule in competition with other rules can be determined by the strength it already possesses combined with the support it gets from other rules in the form of associative activations. This can be made more precise in the following way. The bid B made by rule C when its conditions are satisfied can be given by: $B(\mathbf{C}, t) = aR(\mathbf{C}, t)\ S(\mathbf{C}, t)V(\mathbf{C}, t)$ where a is a constant less than 1, $R(\mathbf{C}, t)$ is specificity of $\mathbf{C}$, $S(\mathbf{C}, t)$ is the strength of $\mathbf{C}$, and $V(\mathbf{C}, t)$ is the support of C, all taken at a time t. Specificity of a rule is determined by how many conditions it has that are matched to input data. The strength of a rule is its probability of firing. Support $\mathbf{V}$ for a rule $\mathbf{C}$ at some time t can be represented as $\mathbf{V}(\mathbf{C}, t)$. This is the sum of all the strengths of the 'bids' made by other rules (activations) in the set of all previous active rules $\{C^*\}$, and can be represented as follows: $\mathbf{V}(\mathbf{C}, t) = \sum_{\mathbf{C}\{\mathbf{C}^*\}} B(\mathbf{C}, t - 1)$. The strength of a rule is revised in the following way: a 'bid' by a rule to represent the environment will temporarily weaken that rule's strength by the amount of the bid $B(\mathbf{C}, t)$: $S(\mathbf{C}, t + 1) = S(\mathbf{C}, t) - B(\mathbf{C}, t)$. If the rule is accepted to represent the environment it is increased in strength by the size of the bid it makes divided by the number of rules that supported it.

What I have described is a framework for modelling the mind's construction of a mental model. It uses conditional rules as its primitive building blocks. A mental model is the activation of sets of these rules. There are different ways that rules can be structured in a mental model, which correspond to two types of rule substructure: *categories* and *default hierarchies.* Categories are just sets of rules that encode our probabilistic assumptions about what properties usually go with what other properties (gases usually disperse throughout a volume, for example). Categories are therefore what we otherwise call 'concepts'. Activation of clusters of these rules can activate

other concepts as well and gives us an idea of what to expect given the satisfaction of specific antecedents (if something is a gas, it probably will disperse at some point).

A mental model is the activation of part of a default hierarchy. Default hierarchies are different from categories in that they are sets of rules organized into hierarchical structures in virtue of our default expectations for an outcome, given subordinate and superordinate relations between concepts. For example, we expect that a gas will disperse throughout its volume and relax to average pressure throughout that volume. On the other hand, sets of rules can include exception rules, which can account for cases where the model has components that are not typical. Low temperature gases make a good example here because they behave in quite peculiar ways, for example, ideal Bose gases need to be modelled using exceptional partition functions in statistical mechanics. Default hierarchies are sets of rules that represent these kinds of different scenarios. In this way they can accommodate variability in the environment. A mental model, such as that of gas molecules in a container, can therefore be characterized as a set of rules which comprise states of a default hierarchy S which have won the right to represent the environment based on the strength of their 'bids'. Models are therefore the activation of specific rules in a default hierarchy.

There are two important functions comprising such a model, and these are determined by diachronic and synchronic rules (see Figure 1). The first function describes how states evolve over time and is given by a transition function, T. This function takes the initial activated state from $S(t)$ to $S(t + 1)$: $T\,[S(t), S(t + 1)]$. It is determined by activated diachronic rules.

The second function is a mapping function P and is determined by activated synchronic rules. The environment is far too complex to map every element into a mental model so a mapping function is required that takes elements from the world and inserts them into the model. In virtue of its selective nature this function is a partial isomorphism. But P is not a singular mapping; it may map many levels in the hierarchy generating different levels of specification or generality for a model. For instance, one mapping P_1 may map elements from the environment into just three in the model (molecules, container, surface), whereas another mapping P_2 may be less picky in its selection of categories and select two elements to put into the model (molecules, container). When a model uses multiple levels of mappings like this, each will be accompanied by a corresponding transition function T_i, taking each P_i from its initial state S_i into its final state $S_{i'}$. This transition function therefore dictates what the model is going to do next. Presumably in the case of an ideal gas in a container the function will map all current $T_{i'}$s from t to $(t + 1)$, which will be for the component molecules to traverse the volume of the container. The entire collection of transition functions $T_i - T_{i'}$ is known as a *quasi-homomorphism* or *q-morphism*. In sum, one can say that a q-morphism is given by the P and T functions, and these are composed of activated synchronic and diachronic rules.

A final concept in mental models, one that is going to be very important for us, is that of *coupling*. Two rules are coupled when one activates the other. For example, R_1 has a consequent that is the same as the antecedent of R_2. Each time R_1 is activated, it activates R_2. In order for a sequence of rules to represent a system as a mental model

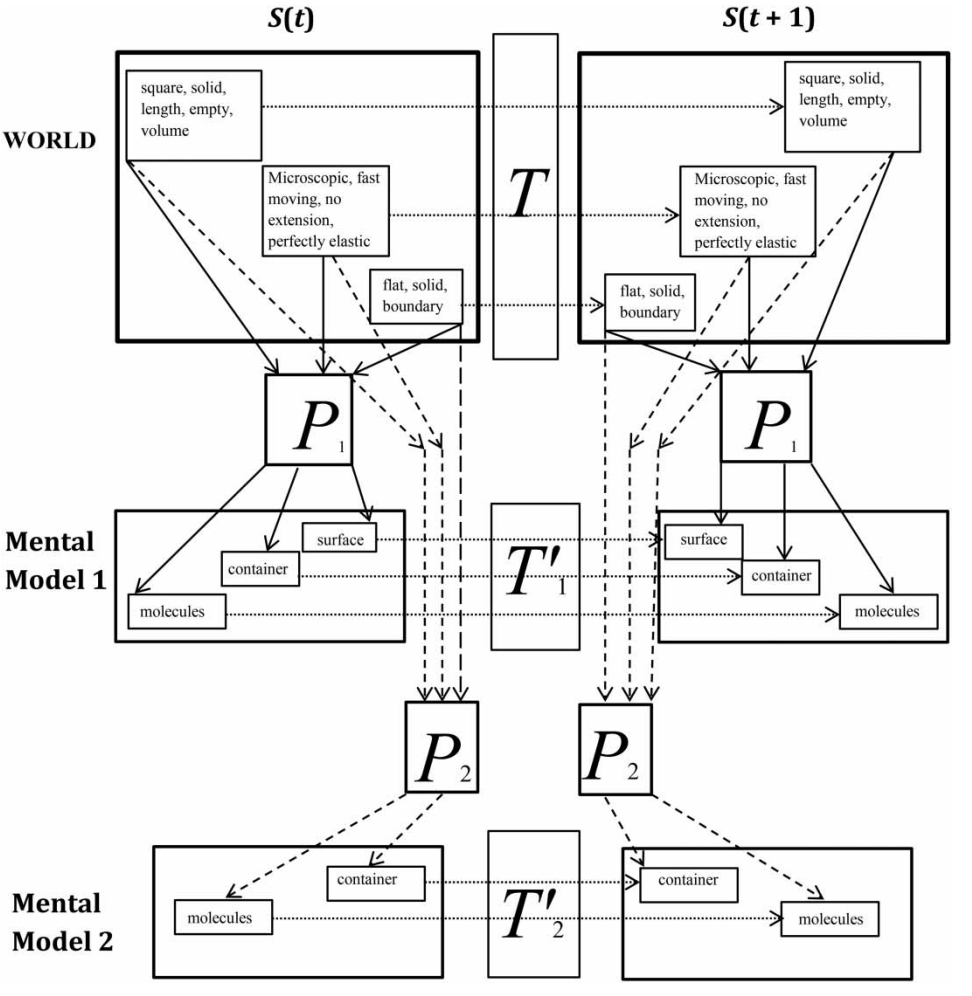

Figure 1 A hierarchy of rules as a *q*-morphism for part of our system. *P* functions are partial isomorphisms from the world to the selected mental model (solid lines for default *P* function, dashed for alternative *P* functions). *T* functions (dotted lines) are time-step transition functions taking members of the model from t to $t + 1$.

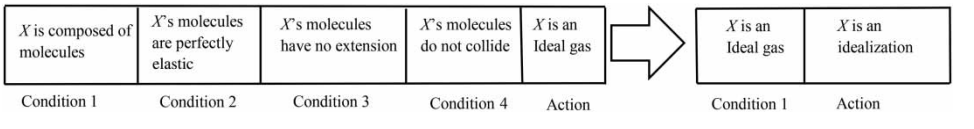

Figure 2 Two rules coupled by the match between action and condition.

they have to be coupled. Coupling therefore plays an essential role in the construction and maintenance of a mental model (see Figure 2).

Coupling is generated by two specific inductive mechanisms. I will argue that it is precisely these mechanisms that are responsible for our coming to understand a phenomenon rather than merely knowing it. These mechanisms are inductive

rule generalization and rule specialization (abduction and analogy are special cases of these).

Rule generalization comes in two varieties, condition-simplifying and instance-based. Condition-simplifying inductive generalizations are simply the cognitive system recognizing an unnecessary number of conditions in a rule and modifying the rule by cutting them. For example, take the rule 'if *X* has molecules, these molecules are perfectly elastic, they have no extension, and they do not collide with one another, then *X* is an ideal gas'. It may turn out, given the system's experience that this rule is activated just as well without the final condition. If so, then the rule can be simplified without harm.

Instance-based generalizations are the more familiar cases where a rule is developed or strengthened on the basis of similar conditions co-varying in the environment with similar actions. The system frequently sees massive objects, and they are impact-producing objects, so it establishes a rule reflecting the co-variation. This is basically a case of enumerative induction, but is essential for establishing rules that can fire to represent the environment.

Rule specialization is the second mechanism responsible for generating coupling between rules. This is a system's means of modifying a rule that is generally used in a situation in light of counterexamples. This might, for instance, occur in the situation mentioned above when we find that not all gases behave according to typical partition functions. Instead of throwing out the standard rules of applying these functions to gases, we just modify the relevant conditions to include 'and the gas is not a Bose gas'. This mechanism saves the system from discarding useful but overgeneralized rules.

4. Representation of the Example

Given this excursion into computational representation, we are now in a position to use this framework to describe a mental model of the explanation of temperature given by the kinetic theory. This will enable us to clearly distinguish knowledge from understanding on a mental models account. Since a mental model is supposed to be a cognitive system's representation of some part of the environment we start by giving the *P* function, which is a partial isomorphism from the environment into our situation model. We break the explanation into physical components (container, molecules) and the properties/relations of those physical components (velocity, pressure, etc.).

Start with the *P* function for physical objects. What components are in need of representation in the given explanation? The model in our explanation is simple: a container and the molecules inside it. The rest is empty space. That gives us only two categories of objects. The container has various properties, which we add to our representation: it is square, with length *L* and area *A*. It has surfaces that are perfectly solid. Molecules also have a number of properties: microscopic, spherical, fast moving, perfectly elastic, no extension. These two categories of objects comprise the physical states of our situation model that represent the environment. These categories

are just sets of synchronic rules for categorization of physical objects. When activated these rules generate activation of other associated rules. These additional activated categories will of course be those related to the properties of the entities in the model. For example, the container description might activate other concepts such as 'cube' or 'box', and the molecule description might activate rules for 'ball', 'hard', 'point', etc.

So far the activated network of rules forms the *P* function for the *physical object* part of the default hierarchy. There is also a non-object-based set of rules that needs activating in the model via the *P* function. There are two kinds of remaining components: those that are explicitly defined in the explanation, and those that are not. For those that are explicitly defined, we face the task of making correct categorizations for their properties, just as was the case for the *P* function with the container and molecules. Here are some of the explicitly defined concepts:

(i) Pressure: $P = Nmv^2/LA$
(ii) Total speed squared: $v^2 = v_x^2 + v_y^2 + v_z^2$
(iii) Length: L
(iv) Area: A
(v) Volume: $V = LA$
(vi) Total mass of the gas: Nm
(vii) Ideal gas law: $PV = NRT$
(viii) Translational KE: $\frac{1}{2}mv^2$

These are the properties and relations of the container and molecules not given by their object classification but essential to following the explanation (each concept being activated is again just a complex condition-action rule). There are also non-explicit properties and relations in the explanation, which if not activated in the cognitive system will undermine its ability to understand. Here are some:

(i) Momentum: $p = mv$
(ii) Rate of change of momentum: $\Delta p/\Delta t$
(iii) Change in time: $\Delta t = 2L/v_x$
(iv) Reciprocal: $1/x$
(v) Newton's second law: $F = ma$
(vi) Density: $\rho = m/V$

The *P* function provides a representation of the explanation's component parts, including the physical objects, their properties and relations, and other background-relevant definitions and principles. The other part of the situation model is given by the *T* function, which uses diachronic rules to provide expectations for what the gas and container are going to do in future time progressions of the model. Presumably nothing much is going to happen to the container—molecules are supposed to rebound off its interior surfaces with perfect elasticity. The molecules on the other hand are changing moment by moment. They are moving very rapidly back and forth in the x direction with no y or z motion at all, and make no collisions with one another. They collide only with y–z oriented surfaces of the container, and

when this happens they bounce back in the opposite x direction with the same speed. So, in comparison with the very detailed P function, our T function is really very straightforward: molecules bouncing back and forth like so many ping-pong balls in a box.

If the mental models approach turns out to be accurate, then what we have built here is a full representation of the *components* of the explanation. We do not yet have a representation of the explanation, merely its parts. Additionally, not all of those parts will appear to us consciously. Our conscious minds are not going to have a perfect picture of some particular number of particles going through this activity, but rather a general, somewhat vague, image of the scene. It is important to reiterate that if the rule-based approach were close to being correct, this conscious representation would not reflect the entire model. Much of our representations on this account may be non-conscious; leaving room for our knowing a great deal that is not being consciously represented.

In the formal terms introduced above we can say that the model for the gas **MG** is composed of a large set of rules $\mathbf{C_{MG}}$: $\{C_{\mathrm{MG}1}, C_{\mathrm{MG}2}, \ldots, C_{\mathrm{MG}n}\}$ some of which are coupled to others in that either condition or action components match and activate the condition or action components of another rule in $\mathbf{C_{MG}}$. On the rule-based account this model is the successful set of rules whose bid for representation was the strongest of all competing rule sequences, as given by $B(\mathbf{C}, t) = aR(\mathbf{C}, t)\ S(\mathbf{C}, t)\ V(\mathbf{C}, t)$. This model is therefore the overall best representation available for the system on balance of relevance, strength of matching, and support from background rules. It is also an instance of a *q-morphism* where the P function is just the set of synchronic rules activated and the T function is the set of diachronic rules. The progression of the gas molecules through the container can be characterized as the sequence of time-steps: $S(t), S(t+1), \ldots S(t+n)$: $T\,[S(t), S(t+1), \ldots S(t+n)]$.

5. Knowledge

In the previous section, I illustrated how a scientific explanation can be represented as the activation of a complex interrelated hierarchy of condition-action rules. What can this tell us about the difference between knowledge and understanding?

The nice thing about having our explanation characterized as a set of logically related rules is that we can see the explanation as a sequence of logically related sub-sequences of rules. We can view the entire explanation then as a model **MG** that is nothing but a complex sequence of rules, each comprising mapping (P) and transition (T) functions for their respective subset of the sequence. The whole sequence starts to look like one big complicated argument, each premise of which has won the right to represent the environment in virtue of winning a bidding war with other rules. Now if this is a useful way to think of mental representation, then it can be used productively to shed light on the problem at hand. Let us first consider what it might mean to *know* rather than *understand* an explanation using this model.

One suggestion is that knowledge of the proposed explanation as a model *just is* the activation of the appropriate mental model (here that would be rules 1–25). This

entails that to know an explanation of something like the temperature of a gas merely requires the activation of an appropriate set of rules. Here the idea is that explanations are known by an individual if that person can conjure up in their mind explicitly or implicitly, the model that is given to them by the explanation. On the rule-based approach this initially seems plausible, since we are assuming mental representations are just sets of activated rules in a hierarchy, and if these rules accurately reflect the explanation, then surely the system can be said to 'know' that explanation.

Even though I believe this is along the right lines, it is a little too quick. One point to note is that our traditional notion of knowledge requires something like *justified true belief* and so far I have said nothing about these concepts and how they are to be characterized in the system. The activation of a complex set of rules is not obviously a satisfactory view of knowledge, even if they do somehow reflect reality.

I cannot hope to do more than scratch the surface here, but will attempt some computation-compatible responses. We can address independently each of the three concepts: justification, truth, and belief. Starting with truth, a naïve correspondence account is adopted, so all that is required is an accurate representation of the propositions provided in the explanation. The mental models account does not require the representations be linguistic in nature, but whether given as propositions or images, or sounds, or something else entirely, the rules of the model are expected to correspond to the given input. The explanation addresses what it is for a *gas* to have temperature. If the rules activated were instead to represent what it is for a *liquid* to have temperature, the knowledge criteria would obviously fail.

Justification is much trickier if for no other reason than epistemologists are still significantly divided on what this concept means. I assume here the possibility of implicit knowledge for our system. Since supporting reasons for a belief are therefore potentially inaccessible I suggest we should adopt an externalist epistemology. I leave entirely open the details of such an account except to mention it should have the usual defeater clauses.

Perhaps one will object that the adoption of an externalist epistemology makes life too easy for my account.[11] The internalist may suggest that to (merely) 'know that' something is the case itself requires we be able to make the kinds of conceptual connections that I require of understanding. Therefore, understanding and knowledge are not so easily divorced as I suppose. My rule-based approach with its sympathy for implicit connections is easier to accept on externalist grounds, which do not require transparent access to inferential steps. For internalists then, the kinds of relationships required for justification are much harder to discriminate, and it is not clear they can be separated from understanding.

I want to make just a couple of points in response to this challenge. First, it is not clear to me that the internalist/externalist distinction holds a lot of water anymore. The leading advocate of externalist epistemology, Goldman (2012), himself acknowledges process reliabilism requires an evidentialist component. This would alleviate the most pressing problem for externalists—how to accommodate scenarios where subjects have justification without transparent access to evidence for their beliefs. On the other hand, internalists trying to keep their approach free of implicit processes

do not fare much better—especially when trying to make sense of a plausible externalist semantics within an internalist account of mental states.[12] I therefore do not think it terribly helpful to condemn a theory of scientific understanding on a presumption regarding which final epistemology will prove least subject to devastating internal tension.

The second point is far less polemical: my account may seem to make life easy for an externalist, but there is no principled trouble with accommodating all I have said so far within an internalist epistemology. I suggest Terry Horgan does just this with his *iceberg epistemology*.[13] For Horgan, understanding relies heavily on what he calls 'implicit conscious content'—information implicitly present in our synchronic experience. He portrays this as possible through a metaphor called the 'chromatic illumination of morphological content'. The basic idea is that an internalist can quite happily appreciate implicit unconscious mental content that may causally contribute to the fixing of beliefs because there may be justifying experiential content in the evidence of our senses. This can be achieved through conscious appreciation of our evidence that is 'illuminated' by non-explicitly represented information. A paradigm example of this is when we 'get a joke'—a great deal of the inference work done in 'leaping' to the conclusion of a joke, given its premises, is of the implicit variety. Still, we do it, and we do it reliably, because of all the information dormant in those premises. This information is however not explicit to us. Thus, we can have justification for a belief even if it is not transparently accessible.

Regardless of how cogent one finds Horgan's story, my rule-based account should cohere with it in virtue of the nature of winning rules: all that implicitly inferential work being done by rules competing to represent. As such, I take it even an internalist can potentially work within this mental models framework.

Moving back then to the main thread of justification, the concern would seem to be this: how are we to accept an activated model, or even a single activated rule, as being justified? In response, this concern can be assuaged I think by recalling the means by which a rule is activated, and since the entire model **MG** can be treated as a complex set of rules, by transitivity this will apply to the entire representation. Recall that a rule wins the bidding war, and gets activated by the system if it has the highest bid, where its bid is given by $B(\mathbf{C}, t) = aR(\mathbf{C}, t)\ S(\mathbf{C}, t)\ V(\mathbf{C}, t)$. This boils down to the claim that the winning bid has the best overall combination of specificity, strength and support. These components map quite nicely onto some of epistemology's most historically respected virtues for a justified belief. But we are not committed to the necessity of precisely these three. Whatever one takes to be the desiderata for belief-selection could similarly be modelled.

That leaves belief. How is a model a belief? This is perhaps the most difficult of the issues relating mental models to knowledge. As such, I will avoid making any particular commitments. It is generally accepted in philosophy that a belief is a propositional attitude and to characterize this notion adequately we require a theory about such attitudes. Representationalists take beliefs to be states of the mind that are representations with the content of a proposition as their object. Stories as to what this involves diverge. Dispositionalists care less about the internal structure of the mind per se,

and argue that beliefs are behavioural dispositions with the content of the thing believed. They also come in many varieties. Then there are interpretationists for whom beliefs are similarly behaviourally determined, but for whom a stance or interpretation reveals which individuals do and which do not possess beliefs—Dennett (1971) is one clear example. Whatever one's account of beliefs, the approach to modelling knowledge pursued in this article should be compatible with any naturalistically informed theory of mind. It may sound from what has so far been said that the picture I have of mind is functionalist and hence particularly susceptible to a causal-dispositional story, but nothing necessitates such a reading. Representationalism is compatible with functionalism, and causal accounts can just as well be non-functionalist. What I have been describing seems to be both representationalist and causal. In line with my previous suggestion though, what is sketched here should not constrain us to one particular story of mind (eliminativism versus functionalism, for example). Similarly, it should not constrain our account of mental content, and in particular of the propositional attitudes.

6. Understanding

If the above line of reasoning is plausible, then **MG** can be considered knowledge in virtue of being a set of beliefs that correspond to reality and are justified. The important question now is, How is understanding **MG** different from knowing **MG**? I suggested that the difference lies in the explanatory connections between the components of an explanation. For this example these connections are causal. We must have knowledge of the causal properties for each component that causes it to manifest the causal behaviour it does in the given explanation: we must know the *cause of the causes*.

But that rather general argument can now be made precise in our framework. The assertion is now capable of being characterized in terms of rule-based mental models. Rather than merely knowing that rules 1–24 entail rule 25, we are asking what it means to understand this sequence. We need to know what it is that makes each cluster of rules trigger the next set, rather than leave these representations as isolated knowledge units. What we want is an account of how these activated rule clusters connect to one another to form the sequence given in the explanation. For the situation model to include the entire explanation the cognitive system has to represent explicitly the claims made in the explanations 1–25, *and also* the inferential connections between these rules. Such inferential connections are revealed by (but may not be limited to) the italicized inference indicator words in steps 1–25.

Importantly, over and above the explanation, those inferences must be represented by a cognitive system in order for it to have a complete understanding. All the italicized statements are inferential in nature, and do the explanatory work in the model. Each can be represented by the system as a complex condition-action rule that is activated in the situation model through the *P* function. They are therefore just more rules, like those we have already been treating, it is just that they are what I will call 'inference rules' (IRs) as opposed to the 'ordinary rules' (ORs) we have so far been addressing.

They do not tell us what is going to happen in time progressions of the model so they are not diachronic rules for the model. They represent abstract inferential relations between parts of the physical system and background theory. We can represent each of the IRs used in the example as conditional statements where each number refers to a line from our initial explanation:

(IR_i): $(1 + 2 + 3) \rightarrow (4)$
(IR_{ii}): $(4 + 5 + 6) \rightarrow (7)$
(IR_{iii}): $(7) \rightarrow (8)$
(IR_{iv}): $(8 + 9 + 10) \rightarrow (11)$
(IR_v): $(12 + 13) \rightarrow (14)$
(IR_{vi}): $(11 + 14) \rightarrow (15)$
(IR_{vii}): $(15 + 16) \rightarrow (17)$
(IR_{viii}): $(17 + 18) \rightarrow (19)$
(IR_{ix}): $(19 + 20) \rightarrow (21)$
(IR_x): $(21 + 22) \rightarrow (23)$
(IR_{xi}): $(23) \rightarrow (24)$
(IR_{xii}): $(24) \rightarrow (25)$

(IR_i)–(IR_{xii}) represent the essential inferential steps in the explanation given by kinetic theory. Without these IRs being fired there is no connection between rules representing the propositions in the explanation and all that is achieved is knowledge, rather than understanding.

This can all be explained through the important concept I mentioned earlier, *coupling*. Recall that two rules are coupled when the firing of one initiates the firing of another. I have not yet explained what causes one rule to fire another, and we need to understand this mechanism in order to understand coupling. To do this we need a means of characterizing our rules in a more precise way. We can start by adopting a classic approach from classifier systems theory: treat each rule not only as a set of conditions that entail an action, but treat each condition and action as defined as a member in a class of 'messages'. These messages are treated as binary strings of fixed length, k over an alphabet of three symbols {1, 0, #}. The # can be used as a 'do not care' element to fill the gaps in the k length chain that are not occupied by either 1 or 0. For example, a simple condition/action rule previously schematized as $\mathbf{C} = \{C/A\}$, where $k = 4$ might be {11##/0011}. Here the input conditions are four different values, two of which do not matter. This rule might represent 'if X has perfectly elastic molecules, and they have no extension/then X is an ideal gas'. This rule may be triggered by any previously activated rule where the output is {1111}, {1100}, {1110}, or {1101}.

This way of using binary messages to code the condition and action components of rules allows us to better appreciate how coupling works. The important operating principle is that of *matching*. Two rules can be coupled if they have activated messages that have matching antecedent and consequent. Remember that to get a rule **C** activated that rule must win a bidding competition in virtue of having the greatest value for B, where recall again that $B(\mathbf{C}, t) = aR(\mathbf{C}, t)\ S(\mathbf{C}, t)\ V(\mathbf{C}, t)$. We have

already seen that knowledge of an explanation can be represented as a sequence of activated rules, where the rules that comprise these sequences are activated in virtue of having won their bidding competitions. They won the right to represent part of the explanation as a rule. Now we can understand how these ORs can be coupled to one another: *to activate the connection and couple already activated matching 'ORs', which otherwise would just provide knowledge, we require the activation of an 'IR' that connects those rules.*

We know that for two ordinary rules (OR_1) and (OR_2) to be coupled requires that a match be found between the messages in the action output of the first rule and the condition part of the second rule. And since our explanations are merely collections of such rules, the coupling of inferences works in the same way. Two rules are coupled when their respective rules are similarly 'matched' but also activated via a further 'coupling' rule being activated. All it takes to activate this mechanism is for this IR to win the right to represent connections in our larger model.

Here is another way to think of it. In propositional logic we learn the IR called *hypothetical syllogism* (HS), which says, $\{[(p \supset q)\ \&\ (q \supset r)] \supset (p \supset r)\}$. We can apply this IR in sub-proofs whenever we come across structurally similar antecedents, such as $[(A \supset B)\ \&\ (B \supset C)]$. We can make the inference to $(A \supset C)$ on the basis of the IR (HS). Think of each of our inferences in the gas explanation as being like these component conditionals, and think of the IR that couples these component rules as being like our rule (HS). We require the activation of (HS) in order to couple $[(A \supset B)\ \&\ (B \supset C)]$ to $(A \supset C)$. Similarly, we require the activation of an IR to couple our activated gas rules together and provide a coupling between them.

This returns us to the issue of what is required to activate a rule, since to understand why and when an IR is activated presupposes we understand how an OR is activated (assuming they operate on the same principles). To see how this works, we need to recall that each sequence of rules is activated because of its specificity, strength, and support. So to figure-out why a rule sequence is successful in establishing itself as representing a situation, and hence generating its connection to the next sequence of rules, we need to ask where these values *R*, *S*, and *V* come from. How does a rule get these values? Furthermore, how does a rule originate in the first place? Answering these two questions will get us to the heart of what constitutes the coupling of rules. Answering these two questions will therefore get us to the heart of what differentiates knowledge from understanding on this computational approach.

In the literature these two problems are solved in a myriad of ways, but in general the issues revolve around what are known as the 'genetic algorithm' and the 'learning algorithm'. The former accounts for how a rule originates, the latter tells us how it is revised or removed from the system. If we are to understand understanding we need to know how these mechanisms generate and update not only our ORs, but also IRs that connect ORs.

The strengthening of a rule comes from its success in the bidding competition and feedback from the environment. So long as the rule which is activated does not suffer negative input in the form of disconfirmation, then it is rewarded as previously mentioned in the amount of its bid, shared amongst all its supporting rules. So, if *x* is the

number of rules supporting the bid for rule **C** then the strength of the rule $S(\mathbf{C}, t)$ is updated after time sequence $(t + 1)$ as follows: $S(\mathbf{C}, t + 1) = S(\mathbf{C}, t) + (1/x)\, B(\mathbf{C}, t)$. This is just to say that the strength of a rule is updated by the strength of its bid shared with all other supporting rules for that bid. This accounts for how a rule is strengthened over successful bids. For a failed bid, the weakening of the rule is simply the loss of the bid previously made by the rule: $S(\mathbf{C}, t + 1) = S(\mathbf{C}, t) - B(\mathbf{C}, t)$. Further, there is the possibility that a rule that wins out over all other rule bids is still not strong enough to be activated. This happens when the *activation threshold* fails to be met. An IR will only couple rules if it is activated, so it must overcome this threshold. This simple set-up explains the necessary components in a successful coupling: matched rules are coupled by a further IR that operates to activate the coupling only if that IR reaches activation threshold.

We still need to know where these rules come from, otherwise we have a coupling mechanism but no origination. This is given by a combination of our previously mentioned tendency to make generalizations and the genetic algorithm for the system. The idea is that we generate associations between examples with common properties and tend to generalize from those experiences. For instance, we see a series of unsupported massive objects and they all fall to the ground. We therefore generalize that all unsupported massive objects fall to the ground. Those generalizations, if successful, are reinforced which gives greater credibility to the condition properties being associated with the action properties in further rules. That is, the concept of a massive object gains credibility in the system. A cognitive system then builds on these successful properties by constructing new rules through using them in further cutting and splicing. For instance, if the system has repeated exposure to a property like *mass* and successfully associates it with forceful impacts, then the *mass* category will be reinforced by confirmed instances. A specific property like this can then be treated as a plausible category to the degree that the rules in which it is embedded are successful. That is, the building blocks for rules in the system are selected based on their success as components of rules extracted from experience. For instance, let m represent the property of *mass*. The strength of this building block for constructing new rules is given by the average value v of its success in all the rules using it:

$$v(m, t) = \sum_{j=1}^{n} \frac{S(C_j, t)}{n}.$$

So, if mass has been a successful building block in a lot of rules, then it will be a good candidate for constructing new rules where it seems appropriate. These rules are then eligible to enter bidding wars to represent the environment when their components are matched to incoming information. If they are successful in both representing and receiving confirmation from the environment, then these rules are strengthened by the amount of their bid, as we have already seen. This is how rules originate.

These ideas can now be used to illustrate our coupling of rules through the activation of IRs in the following way. Two rules may have matching components, such

as OR_1: {1111/0011} and OR_2: {0011/0001}. Let the (0011) segment be our *m* in this example. These two rules are eligible for coupling. However they may both be activated by the system and yet remain uncoupled if there is no further rule which activates OR_2 *in virtue* of its *m* segment matching the *m* segment in OR_1. This is the situation I have claimed holds when we merely know rather than understand a set of rules. We require an activated IR to tell the system that the match is also active. This IR, let us call it IR_α, is just another rule activated by OR_1 and OR_2 when they fire, so it is activated by *association*. The activation of OR_1 and OR_2 each activate associated concepts/rules, such as that of IR_α, and will only occur if IR_α is beyond activation threshold. That is, they occur only if IR_α is a sufficiently strong rule to represent the relation between OR_1 and OR_2. The strength of IR_α is given by the formula for any rule's strength, and will only suffice if it breaches activation threshold. If it does not, then there is no coupling, and all the system achieves is knowledge, not understanding.

So, to be a little more specific, coupling between rules occurs when IR_α is activated where IR_α matches components of ORs. The IR_α is an association rule between ORs that have already won the right to represent the model. Just as 'gas' activates the associated concepts 'molecule', 'air', etc., a coupled set of rules are coupled in virtue of the activation of synchronic association rules between their matching components. Similarly we can be said to understand rather than merely know why $\{[(A \supset B)\ \&\ (B \supset C)] \supset (A \supset C)\}$ only if we can match the relevant component parts of the argument to the rule HS. If the associations fail to fire, the matches remain uncoupled. We fail to understand.

Here then is a final characterization of the difference between knowledge and understanding on the rule-based mental models account being suggested:

(K): Knowledge of an explanation is the activation of ORs in a cognitive hierarchy that correctly represent the explanation's propositional content.

(U): Understanding an explanation is achieved when those activated ORs are coupled by the correct IRs.

So far the picture I have painted provides us with a means for specifying exactly what it is that differentiates knowledge from understanding. Let us see how this is supposed to work for our kinetic theory example. To do this we merely have to go through our IR_α rules as identified by (IR_i)–(IR_{xii}) at the end of section 3. It would be laborious to cover the entire set, but we can quickly run through a couple.

(IR_i) says $(1 + 2 + 3) \rightarrow (4)$. To make this inference is to couple rules (1)–(3) with rule (4). To do this we have to match concepts from the antecedents to concepts in the consequent. In this case the crucial concepts are those of mass, velocity, momentum, and impact. We are told that this is a system of molecules, which should provide by mere referential inference the activation in our situation model of the concept 'mass'. If we did not even know that a molecule has mass, it is unclear if we could make sense of statement (1) and would likely fail to even know this part of the explanation. We would have a seriously deficient situation model. So, it is a first step merely to understand the statements. We are told in (3) that each molecule reverses its velocity in the *x* direction upon impact with a wall of the container. We have to combine then

the concepts of mass and velocity to infer that momentum, which again by referential inference entails the activation of the concept *mv*, is doubled when a molecule reverses direction. The molecule loses all its momentum, coming first to a halt upon impact, and then regains the same amount of velocity in the other *x* direction. So, the inference is that the change in momentum is double the initial momentum of each molecule. This gets built into our situation model but only in virtue of the activation of IR_i.

(IR_{ii}) says $(4 + 5 + 6) \rightarrow (7)$. To make this inference merely requires a bit of algebra, though it is no trivial step in the explanation. We are told in (4) that change of momentum $\Delta p_x = 2mv_x$. We are also told in (5) that each molecule traverses the distance from one side of the container and back again in a time $\Delta t = 2L/v_x$. This is the time between its collisions on the surface *S*. (6) tells us that the reciprocal of this is the number of collisions per second: $1/\Delta t = v_x/2L$. We now have the raw material to perform some conceptual replacements in moving to (7): we multiply Δp_x by $1/\Delta t$, which requires the activation of an IR. This rule is simply of the general form (*x* multiplied by $1/y$ produces x/y). Replacing for our case we get $\Delta p_x/\Delta t$. We then have to infer that since $\Delta p_x = 2mv_x$ and $1/\Delta t = v_x/2L$ that this entails $2mv_x/2L/v_x$. This inference requires the activation of an IR of the general form (*x* multiplied by y/z produces $x/z/y$). Finally, we have to further infer that $2mv_x/2L/v_x$ is equivalent to mv_x^2/L. These separate inferences take some deliberative work, and matching in each case requires the activation of generalized algebraic rules. Still, assuming there is no principled reason to think structural matching is problematic, the inferences work in the same way as cases for concepts that match in a more concrete way, such as we found with 'mass'.

It is a similar process for each of the remaining terms IR_n, where to make the appropriate inference necessary to move through the explanation one must activate already present IRs. Further elaboration would, I think, only belabour the point.

7. Conclusion

The account of understanding as activation of coupled rules via IRs raises a host of questions. I have space here to address only the most pressing. First, one might worry that I have said nothing so far about what dictates the activation threshold for any particular rule, be it an ordinary (diachronic or synchronic) rule, or a coupling IR. What parameters describe this difference and how are they generated? It is an interesting question but not one a philosopher need feel the burden of answering. This account is a characterization of what goes on in our minds when working towards a specific cognitive achievement, not a study of what the relevant empirical values must be in order to actually do it. Whether the threshold for a rule is set at one value rather than another is a contextual and empirical question for cognitive science. This should not be thought a drawback for my account since the task never was to specify the actual values for particular rules, rather my aim has been to show that there are different kinds of rules in use when one understands in contrast to when one merely knows. I have given an objective account of the meaning of that distinction, not a means for testing for it.

A second related worry is that in adopting the bidding system for rules I am just assuming that the correct bid formula must be comprised of the specificity, strength, and support for that rule. Even if the idea of a bidding competition is empirically vindicated why think these the appropriate parameters? In response I admit that this is merely one form of modelling which adopts a specific kind of algorithm (known as the 'bucket brigade algorithm' (BBA)). This was the first widely adopted credit assignment scheme in the learning classifier systems community in machine learning, and although subject to a number of problems it is still considered the benchmark. One notable alternative scheme is the Q-learning-based systems, which have become the most popular current implementation. In fact, Q-learning overcame inherent limits of BBA, but the systems are basically equivalent, now being implemented in a hybrid—the Q-learning bucket brigade algorithm (QBB). New versions (such as X-level category system) do not focus on the strength of a rule $S(\mathbf{C}, t)$ so much as its accuracy. New accuracy-based systems can retain rules that do not accumulate much reward but are nevertheless still reliable classifiers. Ultimately though, the question of whether we select strength, support, and specificity as the appropriate parameters is going to be answered empirically, not *a priori*. Whether the process by which we come to represent a situation is given by these specific variables on a rule-based system is open to empirical challenge. My use of mental models as computational algorithms in its traditional formulation has been more a matter of expository efficacy than a claim to empirical veracity. Still, the important thing is that we have a useful model for how to think about and identify the characteristic relations between knowledge and understanding.

A third concern is that with the introduction of the IR terms, it seems understanding is achievable only by running the risk of an infinite regress of rules. Since IR terms are required to activate OR terms, what rules are responsible for activating IR terms? Meta-IR terms? And what further rules will be required to activate *those* rules, etc.? This concern is misplaced for I do not claim that one must have a coupling between an OR and the relevant IR, only that the IR be activated, and this is a result of it passing threshold and winning the bidding war with other potential IR terms. Recall that although it is a necessary condition on coupling that there be an activated IR, this is not the same as the basic algorithm for activation which is given by $B(\mathbf{C}, t) = aR(\mathbf{C}, t)\ S(\mathbf{C}, t)\ V(\mathbf{C}, t)$. Coupling was defined as holding between two rules where the antecedent of one is the consequent of the other. IR terms and OR terms do not have this relationship.

A fourth worry is that understanding clearly comes in degrees and is therefore a continuous notion, yet my account commits us to a discrete mechanism whereby rules are either activated or not. This really is not a problem at all. The objection mistakes knowledge for understanding. The more IR terms couple OR terms, the more we understand, but we can of course have incomplete understanding of an explanation by only activating a fraction of the correct coupling between available OR terms. A useful analogy here is that of a jigsaw puzzle. If each piece represents a proposition in some explanation which itself is the complete picture, then mere knowledge of an explanation with no understanding will be like the unassembled pieces lying in a

pile. Put some of the pieces together by 'matching' parts of the pieces and you begin to develop understanding of the picture. Put all the pieces together and you have exhaustive understanding of the explanation. In this way we can perfectly well have degrees of understanding of an explanation by only connecting some of the propositions inferentially.

A last worry is more interesting, and opens up lines for further investigation. Surely there are many different ways to understand some phenomenon: causally, intentionally, functionally, mechanically, mathematically, etc. This makes it plausible to think that we ought to avoid a single objective notion of understanding. Yet my account seems to suggest understanding an explanation itself boils down to just one objective thing, the activation of IRs to couple ORs. Even if the distinction between these two sorts of rules can be made coherent there is no means here of differentiating which form of inference is correct.

This concern opens up an important line of inquiry regarding the relation between understanding and kinds of explanation. My account is supposed to give an *objective* characterization of scientific understanding, so it can accommodate the contextual nature of explanations by appealing to empirical facts. An agent is correctly understanding an explanation if the inferences she makes are of the correct variety—causal inferences for causal explanations, mathematical inferences for mathematical explanations, etc. Since the debate over the objective nature of explanation is still far from settled it would be hubris to suggest understanding must be of only one particular variety, and my account does not make that claim. For now we can merely say that these should be correct inferences—ones that correctly match the way an explanation claims the world is. Likewise, a 'good explanation' in the objective sense should be one that is both correct and enables a subject to make the relevant correct inferences between its component propositions (ORs)—one that enables the subject to understand the explanation. Cushing's opposition to non-causal explanations can I think be seen as an expression of frustrations along these lines. Physical explanations are satisfying when they allow us to make causal inferences, others are less so since they leave us with the activation of mere mathematical IRs.

Acknowledgements

Thanks to two anonymous referees of this journal, the editor James McAllister, and Pat Shade for very helpful comments on an earlier draft of this article.

Notes

[1] This law says the orbit of each planet is an ellipse with the sun at one focus.
[2] For a full argument supporting this claim, see Thagard (2012).
[3] For the full argument, see Newman (2012). Nothing in this section depends on any particular account of how to represent cognition.
[4] This is a little long-winded, I am afraid, but all these steps are necessary for what follows.
[5] See, for example, Way (1991), Otero, Léon, and Graesser (2002), Holyoak and Morrison (2005), and Tapiero (2007).

[6] That understanding hangs crucially on making connections is a very common assumption in the literature, made by among others Zagzebski (2001), Kvanvig (2003), Grimm (2006, 2010), and Elgin (2007).

[7] I am assuming that inferences between mathematical steps in this explanation are still causal since the mathematics reflects physical properties and relations, rather than abstract objects.

[8] Here I am addressing an issue raised by one reviewer and the editor of this journal. I very much thank them for pressing me on this point.

[9] See, for instance, Anderson (1983, 1993), Holland et al. (1986), Newell (1990), Thagard and Litt (2008), and Thagard (2012). The same computational architechture can be found used in learning classifier systems, a branch of machine learning. For that use, see Wyatt (2002), Bull and Kovacs (2005), and Urbanowicz and Moore (2009).

[10] Notice however that here the disjunctive antecedent would actually be manifested as a cluster of rules with identical consequents.

[11] Thanks again to one of my reviewers and to the editor for raising this concern.

[12] This is a sticking point for one of internalism's most well developed theories, evidentialism. See especially Connee and Feldman (2012).

[13] Horgan's views on this issue can be found throughout most of his work on epistemology and mind over the last couple of decades. A nice place to find them collected is his recent co-authored book (Henderson and Horgan 2011).

References

Anderson, J. R. 1983. *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. 1993. *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum.

Bull, L., and T. Kovacs. 2005. *Foundations of Learning Classifier Systems*. Heidelberg: Springer.

Churchland, P. M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

Conee, E., and R. Feldman. 2012. "Replies." In *Evidentialism and Its Discontents*, edited by T. Dougherty, 283–323. Oxford: Oxford University Press.

Cushing, J. T. 1991. "Quantum Theory and Explanatory Discourse: Endgame for Understanding?" *Philosophy of Science* 58: 337–358.

Dennett, D. 1971. "Intentional Systems." *Journal of Philosophy* 68: 87–106.

Elgin, C. 2007. "Understanding and the Facts." *Philosophical Studies* 132: 33–42.

Goldman, A. 2012. "Toward a Synthesis of Reliabilism and Evidentialism? Or: Evidentialism's Troubles, Reliabilism's Rescue Package." In *Evidentialism and Its Discontents*, edited by T. Dougherty, 254–280. Oxford: Oxford University Press.

Grimm, S. 2006. "Is Understanding a Species of Knowledge?" *British Journal for the Philosophy of Science* 57: 515–535.

Grimm, S. 2009. "Reliability and the Sense of Understanding." In *Scientific Understanding: Philosophical Perspectives*, edited by H. W. de Regt, S. Leonelli, and K. Eigner, 83–99. Pittsburgh, PA: University of Pittsburgh Press.

Grimm, S. 2010. "The Goal of Understanding." *Studies in History and Philosophy of Science* 41: 337–344.

Henderson, D., and T. Horgan. 2011. *The Epistemological Spectrum: At the Interface of Cognitive Science and Conceptual Analysis*. Oxford: Oxford University Press.

Holland, J., K. Holyoak, R. Nisbett, and P. Thagard. 1986. *Induction: Processes of Inference, Learning and Discovery*. Cambridge, MA: MIT Press.

Holyoak, K., and R. Morrison, eds. 2005. *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press.

Kvanvig, J. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
Newman, M. 2012. "An Inferential Model of Scientific Understanding." *International Studies in the Philosophy of Science* 26: 1–26.
Otero, J., J. A. Léon, and A. C. Graesser. 2002. *The Psychology of Science Text Comprehension*. Mahwah, NJ: Lawrence Erlbaum.
Regt, H. W. de. 2004. "Discussion Note: Making Sense of Understanding." *Philosophy of Science* 71: 98–109.
Regt, H. W. de. 2009. "Understanding and Scientific Explanation." In *Scientific Understanding: Philosophical Perspectives*, edited by H. W. de Regt, S. Leonelli, and K. Eigner, 21–42. Pittsburgh, PA: University of Pittsburgh Press.
Regt, H. W. de, ed. 2013. "Special Section: Understanding Without Explanation." *Studies in History and Philosophy of Science* 44: 505–538.
Regt, H. W. de, S. Leonelli, and K. Eigner. 2009. *Scientific Understanding: Philosophical Perspectives*, edited by H. W. de Regt, S. Leonelli and K. Eigner. Pittsburgh, PA: University of Pittsburgh Press.
Tapiero, I. 2007. *Situation Models and Levels of Coherence: Toward a Definition of Comprehension*. Mahwah, NJ: Lawrence Erlbaum.
Thagard, P. 2010. "How Brains Makes Mental Models." In *Model-based Reasoning in Science and Technology*, edited by L. Magnani, W. Carnielli, and C. Pizzi, 447–462. Berlin: Springer.
Thagard, P. 2012. "Cognitive Architectures." In *The Cambridge Handbook of Cognitive Science*, edited by K. Frankish and W. Ramsay, 50–71. Cambridge: Cambridge University Press.
Thagard, P., and A. Litt. 2008. "Models of Scientific Explanation." In *The Cambridge Handbook of Computational Psychology*, edited by R. Sun, 549–564. Cambridge: Cambridge University Press.
Trout, J. D. 2002. "Scientific Explanation and the Sense of Understanding." *Philosophy of Science* 69: 213–233.
Trout, J. D. 2005. "Paying the Price for a Theory of Explanation." *Philosophy of Science* 72: 198–208.
Urbanowicz, R. J., and J. H. Moore. 2009. "Learning Classifier Systems: A Complete Introduction, Review, and Roadmap." *Journal of Artificial Evolution and Applications.* http://www.hindawi.com/archive/2009/736398
Way, E. C. 1991. *Knowledge, Representation and Metaphor.* Dordrecht: Kluwer.
Wyatt, J. 2002. "Reinforcement Learning: A Brief Overview." In *Perspectives on Adaptivity and Learning*, edited by I. O. Stamatescu, 243–264. Berlin: Springer.
Zagzebski, L. 2001. "Recovering Understanding." In *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, edited by M. Steup, 235–252. New York: Oxford University Press.