Is this what the debate on rules was about?

Ulrike Hahn

School of Psychology, Cardiff University, Cardiff CF10 3YG, United Kingdom.
hahnu@cardiff.ac.uk
http://www.cf.ac.uk/psych/home/hahnu/index.html

Abstract: The key weakness of the proposed distinction between rules and similarity is that it effectively converts what was previously seen as a consequence of rule or similarity-based processing, into a definition of rule and similarity themselves – evidence is elevated into a conceptual distinction. This conflicts with fundamental intuitions about processes and erodes the relevance of the debate across cognitive science.

The target article recommends a conceptual distinction between rule-based and similarity-based processing based on the number of object properties involved in the decision. Where a "small subset of the relevant object properties" are involved (sect. 2, last para.), we are dealing with a rule process, otherwise the process moves, along a continuum, towards one of similarity.

One concern is the workability of this proposal: is it "small" in terms of absolute numbers or proportions that is key, and how is the set of "relevant" properties to be determined? The proposed criterion for relevance as those properties "uniquely common to the instances of each category" (sect. 2, para. 3) will not do, given the fundamental insight of Wittgenstein, brought to cognitive psychology by Rosch and others, that natural language categories typically do not possess minimal sets of features shared by all instances. Therefore, typically, there will be no feature set which, as a whole, is unique to a category; but nor will features of instances likely be unique to a category when considered individually: both dogs and cats are "mammals" and both houses and cars "have doors."

A further ambiguity surrounds what it means for properties to be "involved" in the decision. This will be illustrated with an example. Two people are asked to classify letter sequences as grammatical or agrammatical. Person A uses edit-distance; that is, person A determines the number of operations required to change one string into another as a measure of their overall similarity. If the similarity between a training and a test string is high, A adopts the same classification. Person B uses a finite state grammar and classifies each item according to whether or not the sequence can be generated with the grammar. Both are given a grammatical training item for illustration, say sequence MSSX, and a subsequent test item, say MSSSX. Both classify this item as grammatical according to their respective procedures. Crucial to the example, we know exactly what procedures they used because they told us so, and we have no reason to doubt their reports, which are entirely consistent with all other aspects of their behaviour during classification (e.g., A went back and forth between the training and the test item on his sheet, whereas B looked only to the test item, sketched out his grammar, etc.).

The further ambiguity in the proposal is about whether the properties involved must be "many" or "few" with regard to (1) the stimulus currently under consideration (i.e., the individual test item), or (2) the entire set of properties evidenced across the category as a whole. In the present example, both interpretations coincide. The same number of properties are involved in the decisions of A and B, because the procedures of both require them to consider every feature of the novel item in determining their classification (1) and trivially the same structural relationships obtain between the old and new item for both (2).

In effect, the procedures of A and B which constitute paradigm cases of similarity- and rule-based classification, respectively, and which are meant to be two ends of the continuum collapse into each other. The target article considers this issue in section 4.4, where patterns of classification such as the present one are discussed. The claim there is that sometimes rules and similarity will converge, and if we then want to classify participants as rule *or* similarity oriented, as opposed to both, we must consider assemblies of judgements. Specifically, we should supplement the above data set by providing the test sequence MSSSSSSX. If this string is accepted as grammatical, then a rule process was involved.

This suggests that it must be the second sense of number of "properties involved" that is intended, namely, whether there were many or few with regard to the structural relationships between items across the category as a whole, because only this second sense now distinguishes the edit-distance and the finite state grammar case. With regard to just the properties of the test item itself (1), the processes continue to be the same. To classify the novel sequence MSSSSSSX, both person A and person B consider *all its letters*: A, in the context of the edit-distance comparison, must determine the number of operations required to generate the test sequence from any of the old items; B, with the finite state grammar, must check that the grammar allows generation of the entire sequence for it to be grammatical.

What has changed, however, are the structural relationships between items in the data set. Specifically, what has changed is the amount of feature overlap between old and new items. We see that this is the only interpretation of the proposed criterion by which these two paradigm cases of rule and similarity could *ever* be distinguished. However, basing the meaning of rule or similarity on structural relationships in the data is ultimately no more satisfying than basing it on the numbers of object properties of the test item considered, even if it disambiguates the edit-distance and finite state grammar application.

Structural relationships between old and new items have always been viewed as *evidence* for rules or similarity, but the present proposal, it seems, turns what was evidence of the conceptual contrast between rules and similarity into their definition. The degree of featural overlap among members of a category is no longer a consequence of a particular kind of mental process, that is, a signature effect which can be used to infer the use of this mental process; rather, it becomes constitutive of being that mental process. Minimal property overlap in classification is not a consequence of rules, it is now what "rule" *means.*

The problem with this development is that it now necessarily renders numerous other sources of evidence conceptually irrelevant. It is entirely irrelevant in the above example that we know exactly what procedure A and B are using because they could tell us so. A further, counterintuitive consequence is that the characterization of what A and B are doing changes as the test set is expanded because the structural relationships between old and new grammatical items change once we add the further sequence MSSSSSSX. Though B, in particular, did the very same thing with his first test string MSSSX as he then did with MSSSSSSX, the original "process" involved in classifying MSSX was *conceptually* a rule- and similarity-based one in equal measure, whereas it is now a rule-based one. We are forced to accept that it has subsequently changed *in nature,* even though it has remained the same process, clearly described to us by B, all along.

This is a necessary consequence of defining "processes" exclusively in terms of their outcomes, and one that makes clear that, despite the terminology of the article, the distinction drawn is (if it is to be compatible with the examples discussed in the article) not about process at all. However, the proposed reduction of two kinds of processes to structural patterns of behavioural data is not just problematic because it means that the same process can be "similarity" or "rule" depending on the data to which it is applied. Whether or not peoples' behaviour draws on many or few properties might or might not be an interesting question in its own right. It is a contrast that seems to be at the heart of at least some studies within the psychological literature on categorisation (e.g., Johansen & Palmeri 2002; Regehr & Brooks 1995). However, the situation in the wider cognitive science literature is quite different. People cared so much about the debate on mental rules in general precisely because it involves a fundamental debate about the computational characteristics of human information processing, that is, a debate about the kinds of *representations* and *processes* that underlie human cognition. Patterns of behavioural

https://doi.org/10.1017/S0140525X05340015 Published online by Cambridge University Press

data were of interest (only) because they were thought to inform this debate. This is true not only of the expert systems debate in Artificial Intelligence, but in particular of the debate in the context of language, from the longstanding controversy between rules and connectionism prompted by Rumelhart and McClelland's (1986) model of the past tense to more recent contrasts between grammars and data-oriented parsing (e.g., Bod 1998). Past proposals of the rule/similarity distinction such as Hahn and Chater (1998) have (successfully or not) sought to characterise the representations and processes that might count as rules or similarity and how they relate to data. The proposed distinction in the target article does more than turn this relationship on its head, in that patterns of data are all that is to remain.

Rules and similarity – a false dichotomy

James A. Hampton

Psychology Department, City University, Northampton Square, London EC1V OHB, United Kingdom. **hampton@city.ac.uk**
www.staff.city.ac.uk/hampton

Abstract: Unless restricted to explicitly held, sharable beliefs that control and justify a person's behavior, the notion of a rule has little value as an explanatory concept. Similarity-based processing is a general characteristic of the mind-world interface where internal processes (including explicitly represented rules) act on the external world. The distinction between rules and similarity is therefore misconceived.

In order to maintain a meaningful theoretical distinction between two explanatory notions such as rules and similarity, it is necessary to be clear about how the terms are to be used. As Pothos notes, there has been much discussion about whether "similarity" can be rendered as a useful theoretical notion (Goldstone 1994a; Goodman 1972). Similar issues arise in defining the notion of a rule.

The prototypical notion of a rule is an explicit code that governs conduct – a school rule or a traffic rule would be a good example. A legal code, for example, is a set of rules that governs the behavior of those working in the legal/justice system. In framing rules of this kind, lawmakers are meeting three aims. First, they select the relevant dimensions on which decisions and actions should be based, thus ensuring that legal decisions are not based on prejudiced or arbitrary grounds. Second, they provide a basis for the public justification of legal decisions; the application of the rules allows a judge to make explicit the grounds for a decision using deductive logic. Third, an explicit set of rules allows for the sharing of beliefs. Any competent member of the community can reasonably be expected to understand and apply the rules to their own behavior. The rules provide the conceptual framework within which appeals and argument can take place.

How can this central notion of a rule be applied to models of cognitive psychology? An uncontroversial use of the notion would be to consider rules as explicitly held beliefs that people use to direct their actions. To spell a word correctly, I remember the rule "i before e except after c." To avoid a hangover, I apply the rule of never drinking spirits after dinner. This sense of rule as explicitly codified principle can be seen in a number of cognitive models. RULEX (Nosofsky et al. 1989) is a good example: A learner classifies a set of stimuli by choosing an explicit rule, and then learns to spot the individual exceptions. This type of learning is familiar from the experience of learning a new language in the classroom, where the teacher provides the rule for forming a past tense and then the student learns the irregular exceptions. Until the student becomes more fluent, she may explicitly apply the rule when forming a sentence in the new language.

Where the notion of rule becomes problematic, and quite possibly empty, as an explanatory tool is when it is applied to describe regularities in behavior of which the agent has no explicit knowledge. In such cases (such as using the syntax of one's native language, or following the rules of social interaction in everyday contexts) the person can be said to be *following* a rule, but this is not evidence that the rule itself is represented in the part of the mind/brain directing the behavior. Behaving in a regular manner "as if" following a rule is a property of many different types of system, including physical systems with no mental representations at all. Water flows downhill as a rule, but does not represent this rule in itself. Rule-governed behavior is not sufficient evidence for a model in which the internal representation of those rules has a causal role in the production of the behavior.

I would propose then that the notion of "rule" in cognitive science should be restricted to those rules that can be explicitly stated by the person following the rule. (It then becomes an interesting question whether the rule is causally efficacious or merely used for post hoc justification.) Of course such a restriction will be very constraining on the range of situations in which we can explain behavior in terms of a rule. There are, however, clear examples. Situations in which rules control behavior would include the classic concept identification experiments conducted by Bruner et al. (1956) and experiments on inductive reasoning where rules have to be hypothesized to account for observed data (Wason 1960). More recently, Ashby et al. (2002) have a range of very telling dissociations between learning contexts that involve explicit reasoning and those that use implicit associative learning, and also have evidence that different brain systems are involved.

The danger of not restricting the notion of rule in this way is that, effectively, any systematic cognitive process could be thought to involve a rule. Short-term memory follows rules (most recent items are recalled first); attention and perception work according to rules – the notion of rule simply becomes the notion of an observed regularity. No causal mechanism involving representation of the rule can be implied.

Having restricted the meaning of "rule" narrowly enough for it to have some distinct explanatory value, we can then ask whether "similarity" is the best concept with which to describe other forms of behavior that are not directly controlled by explicit rules. Here again I find the notion problematic, and indeed the dichotomy between rules and similarity to be false. Consider how a rule is applied in a given situation. A rule generally has two parts: a condition that must be satisfied to trigger the rule, and an action that follows once the rule has been triggered. In deciding whether the triggering condition of a rule has been satisfied, it is inevitable that similarity will be involved. Some situations will trigger the rule in a clear prototypical fashion. Others will partially match the conditions, and will result in slow and uncertain application of the rule. A learner who has decided to follow the explicit rule of putting all red blocks in one pile and all orange blocks in another will need to use similarity judgments when faced with colors intermediate between red and orange. Generally speaking, with the exception of artificial microworlds such as chess or baseball, there will always be the potential for vagueness and uncertainty in how the rule applies to an individual case. All processes that involve the interface between internal processes and the external world will exhibit similarity-based effects, regardless of whether explicit rules are involved or not.

https://doi.org/10.1017/S0140525X05340015 Published online by Cambridge University Press