Understanding "Rules": When is Behavior Rule-Guided?

Ulrike Hahn (U.Hahn@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry, CV4 7AL, UK.

Nick Chater (N.Chater@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, U.K.

Abstract

The extent to which human cognition can be understood as rule-based is a classic issue in Cognitive Science and one which continues to provoke heated debate in a wide variety of areas, ranging from Implicit Learning through Inflectional Morphology to the acquisition of reading skills. Despite its centrality, the central notion of "rule" is far from well-defined. This paper examines a central feature of rule-based models, the concept of rule-following, and clarifies its role, its content, and some of the typical fallacies associated with its use.

Introduction

To what extent human cognition is based on rules is a cognitive question of longstanding interest. In the early days of AI, the rule-based nature of human thought was axiomatic; rules no longer have this general, dominant role, but rule-based accounts of particular tasks still abound. Artificial Grammar Learning (Reber, 1989; Brooks & Vokey, 1991; Redington & Chater, 1996) and Inflectional Morphology (Rumelhart & McClelland, 1986; Plunkett & Marchman, 1992; Pinker, 1991; Marcus et al., 1995; Nakisa & Hahn, 1996) are but two areas which are dominated by ongoing debate between proponents of rule-based accounts and supporters of alternative models such as exemplar-based or connectionist accounts. Despite this continued interest in rules, the very notion of rule is one of the most confused within Cognitive Science. This is manifest, to name just one example, in the lack of consensus about whether or not connectionist networks have or embody rules: statements to the effect that they clearly do not, and, hence, offer alternatives to rule-based accounts (Rumelhart & McClelland, 1986; Smith, Langston, & Nisbett, 1992) can be contrasted with the claim that "contrary to rumour, it is not the case that connectionist systems have no rules" (Bates & Elman, 1993, pg. 634).

Conceptual clarification is essential if debate about rules is to have substance. In service of such clarification, this paper focusses on a central aspect of the notion of rule -the dichotomy between behavior which is *guided* by rules as opposed to behavior merely *described* by rules. This distinction is fundamental to what it means for a behavior to be rule-based, yet confusion both about its content and its role prevails.

"Rules" in cognitive contexts: "strong" and "weak" readings

What exactly do researchers mean when they appeal to rules in explaining behavior? We can distinguish two different kinds of usage of "rule", which we will respectively refer to as "weak" and "strong". Examples of weak usage of the term rule are statements such as, for example, a general, behavioral claim that a language learner has succeeded in "mastering the rules of English" or the assumption that infants are born with "rules for looking" which guide their exploration of the visual environment. Statements like these use the term "rule" to refer to an external regularity (of English) or to an internal constraint without making a claim about mental architecture, i.e., without wishing to endorse a particular view about how the external regularity or the innate constraint are internally represented by the agent. Such a weak usage of the term rule in a cognitive context is NOT the focus of the debate about mental rules, nor is it the focus of this paper.

Rather, this debate is concerned with the "strong" use of rule. On the strong reading, speaking of an agent as possessing a rule is a statement about cognitive architecture. It is the claim that an agent has mental representations of a particular representational format, a format which is distinct from other types of mental representation. This stronger, more specific claim lies at the heart of the debates in Artificial Grammar Learning or Inflectional Morphology, where rule-based models are contrasted with exemplar or connectionist accounts.

Rule following

Most importantly, the strong use, on which we will focus below, claims an agent-internal role for the rule. The claim is one about the nature of the representations underlying a particular behavior. Stating that an agent possesses a particular rule is not merely saying that this agent's behavior displays a particular regularity, but rather that this "rule" has a causal role in producing this behavior: the behavior has the regularity it does, *because* the agent posseses the rule in question.

This is commonly phrased in terms of the distinction between rule-guided, or "rule-following", behavior and behavior which is merely conveniently described by rule (see e.g., Marcus et al., 1995). For an example of rule-following, one can think of legal systems and their effect (where documents encoding the law cause particular behaviors such as paying certain amounts of tax), whereas a standard example of rule-such an account, more must be given than the possibility of a rule-based account; it must be shown that the posited rules are causally efficacious.

Rule following and the causality of representation

Causal efficacy is typically cited as the hallmark of rule-guided behaviour. For instance, Searle, in a critique of Chomsky's (1980) Rules and Representations, holds that, in contrast to rules as used in the natural sciences which merely describe and explain, the use of rules in explanation of human behavior requires that the content of the rule must function causally in the production of the very behavior the rule seeks to explain (Searle, 1980).

First, from what we have said above, it is clear that Searle's position must be disagreed with in one respect: it is not "the content of the rule" which must function causally, but rather the statement (representation) of the rule. Its content, as we have seen is just the regularity in question; but, as we have also seen, this regularity can be exploited in different ways. It is only when it is used by a particular type of representational format that we speak of rule-based accounts; it is this particular type of representation that must function causally, not the regularity.

Second, we must ask what it actually means to "function causally" and how this can be ascertained. Loosely following Chomsky (1986) we assume that we are entitled to hold that an agent is following a rule R if our "best theory" of what the agent is doing, i.e., the best we can construct with the available evidence—invokes a mental representation of R. But this requires further clarification both of what it is a best theory of and what evidence must be taken into account.

We have already seen that rule-guided behavior is about a particular type of explanation. The sort of explanation which such a "best theory" seeks to provide is an account of behavior in terms of mental representations and procedures; For "causal efficacy" we require no more than that the rules are invoked in an explanatory account which involves procedures drawing on representations of these rules and that this explanatory account constitutes our "best theory" available.

Such explanatory accounts in terms of representations and procedures are exactly what researchers engaged in classic rule debates such as Artificial Grammar Learning or Inflectional Morphology are seeking. Most importantly, there is no restriction on what evidence is permissable or relevant.

In our past tense example, evidence for what constitutes the best theory is by no means confined to the ability of the models to produce the right past tense forms. Rather, both models exhibit a whole range of characteristics which give rise to further predictions. For example, they require different learning strategies (rule induction vs. instance storage) and as a result may produce quite different learning profiles: the time course of learning can differ, as might be what is easy and hard. One might be more tolerant of `noise' in the data and so on. *Any* such attributes can be called upon in assessing which theory best fits the data, as well as the desire for parsimony and coherence with other bodies of theory which we bring to the task.

The importance of levels

Additionally, the importance of levels of description must be emphasized. Levels of description are inherent in the context of biology—our theories can invoke brain regions, neurons, or neurotransmitters—as well as in computation where we can reiterate the question of how an algorithm is implemented proceeding downwards from "C-code" to assembly language, to logic gates, to silicon and so on. Hence, the issue of levels is unavoidable for cognitive theorizing. It is pertinent in this context, because accounts in terms of representations and procedures, putative "best theories", might be available at multiple levels as both biological and computational examples suggest.

Specifically, production-rule systems are Turing equivalent, that is, any effectively solvable algorithmic problem can be solved by a production system (Post, 1943). This means, any computation can be made "rule-based" and, as a result, any cognitive theory could be perceived as rule-based if there are no constraints on level. In particular, the nearest neighbour account of the past tense could be implemented using production rules, giving rise to the spurious claim that performance was "rule-based" after all.

Similarly, the constraints on what consitutes "connectionism" seem weak enough to allow implementation of virtually anything and connectionist implementations of "higher-level" cognitive accounts are regularly presented, e.g., Kruschke's (1992) ALCOVE, which implements an exemplar model popular in the categorization literature) or Touretzky and Hinton's (1988) implementation of a production-rule system.

This means commitment to a particular level of description is required. In particular, sweeping contrasts between connectionist and rule-based accounts of cognition, lacking commitment to a particular explanatory level, lack focus and, hence, substance.

The scope of the distinction

Inocuous as our rendition of what it means for behaviur to be rule-guided might look, it has a number of highly desirable properties. First, it applies equally to agent external, "public" rules and to agent internal, "private" rules", i.e., to rules I am told as well as rules I posit to myself;[1] furthermore, it can apply equally to rules, which are formulated in natural language and to "tacit rules" to which we typically have no conscious access. This is because it is defined, generally, in terms of causal efficacy of a representation with requisite format. Again, it does not obviate the need for decision on which formats qualify, but this is a question which itself arises equally for the natural language case and for putative cognitively impenetrable representations.

Second, our rendition of "rule-guided" allows one to see that what is generally treated as one of the many problematic issues about rules is ultimately a general issue of cognitive theorizing. "Rules" are not special: the rule-guided vs. rule-describeable distinction is all about the inference from salient regularities to cognitive models which exploit these regularities. For behaviour to be rule-based, more must be shown than the regularity itself, this "more" being the "causal effi-

[1] This contrasts, for instance, with Quine (1972).

come available if one were to adopt this view? To our knowledge, none have been put forward. This is in strong distinction from a classical rule-following system, like an expert system, where the rules which the system uses in inference provide a completely different level of explanation from the causal story about the workings of the underlying machinery.

This leads to the general question of *why* the rule-guided/rule-describale distinction really matters. From the point of cognitive theory, there appears to be a consistent set of generalizations concerning the behavior which classical rule-based system exhibit: e.g., it is possible discretely to add in extra pieces of knowledge to a rule-based system, which will then interact with previously stored rules; the system can learn by being "told" such knowledge, rather than learning from experience; and it is easy to achieve generalization across extremely disparate items. None of these properties apply to standard backpropagation networks, which have a different set of abilities, learning primarily from experience, where information is accrued incrementally, rather than in discrete packets, and most easily generalizing across similar items. Conversely, rule-based systems have problems in learning from experience, and have difficulty learning "quasi-regular" mappings which involve regular and exceptional cases, particularly if such mappings are governed by subtle effects of similarity. Connectionist networks excel in these domains. Overall, then, it is not clear that any of the important theoretical generalizations associated with rule-based systems carry over to standard backpropagation networks; hence, saying that these networks "follow" rules inappropriately suggests that the two kinds kinds of system share properties on which they actually differ.

There is one further interesting assumption in the Elman et al. quotation, namely, the remark that for merely rule-describable behavior, behavior only "accidentally" accords with the rule. Of course, there need be no accident about the fact that behavior corresponds to the rule; planets do not accidently have the orbits posited by the laws of physics. It is just that they do not use a statement of these laws to compute their orbits.

Assuming that rule-description is always only "accidentally" connected to observable behavior marginalizes the explanatory import that rule-description too can have. Issues of explanatory relevance, seem, to us, to underly the misunderstandings surrounding Chomskian linguistics, our second and final example.

We have repeatedly stressed that rule-following is about a particular kind of explanation and, above, we introduced Searle's comments that the use of rules in psychological explanation is distinct from that in the natural sciences. These issues deserve further elaboration. It is true that, as the case of the motion of planets shows, concise statements of regularities are central to the natural sciences and there are unquestionably perceived as "explanatory". Crucial to the explanatory power is the reduction of a complex behavior to a limited number of variables. Contrary to what Searle seems to suggest, this type of explanation has a role in psychology and Cognitive Science as well. Shepard's Universal Law of Generalization (1987) claims a universal function underlying generalization in humans and a range of non-human species on a variety of tasks. Similarly, "rational analysis" (Anderson, 1990) provides a form of explanation not immediately concerned with mechanism. Most frequently, however, descriptive statements of regularities, "weak" uses of rule, provide a form of explanation which is only partial and, hence, incomplete.

We can illustrate this latter category with a linguistic example, that of the German gender system. Linguistic study and connectionist modelling (see Koepcke, 1993) have isolated phonology as the key factor determining the assignment of gender to German nouns. This has made it clear that German gender, which was previously thought to be arbitrary, is, in fact, highly systematic. A highly complex system and corresponding linguistic behavior are reduced to a single critical variable. Discovering and stating this regularity does "explain" German gender.

From a cognitive perspective, however, we have said that this is always only a first step. Stating the regularity is not a full cognitive account, i.e., an account which explains behavior in terms of mental representations and procedures, simply, because—as seen above—this regularity might be exploited by the cognitive system in a myriad of ways.

The cognitive architecture underlying knowledge of gender might be simple exemplar storage, schemas which abstract families of similar words into more abstract internal representations (Koepcke, 1993) or sets of rules. All of these are conceptually distinct and give rise to different secondary predictions. It is precisely because of these different further predictions that these issues matter to the study of behavior. Finally, this step to internal representations and procedures matters, because it provides a lithmus test for the regularity in question. The "wrong" regularity, e.g., a spurious correlation, will ultimately yield only unsatisfying cognitive accounts, hence, theories in terms of representation and process feed back in to the evaluation of particular descriptive accounts.

All of these issues play a role in the continued debate about Chomsky. Chomskian linguistics, which Chomsky explicitly holds to be concerned with the psychology of the individual (Chomsky, 1980, 1986), aims to answer questions about the nature of linguistic knowledge through the specification of a grammar, i.e., a descriptive account (Chosmky, 1986). Such a grammar is viewed as a putative "best theory" from which we are allowed to infer the entities postulated are "real". This step from description of regularity (grammar) to mental representations is just the step from regularity to rule-based account, which, as we have seen, is an inference which requires further evidence to be justified. While Chomsky does not set out any bounds on "allowed evidence" for what constitutes our best theory, he also shows no positive sign of interest in the type of additional data one needs to resolve architectural in other areas of cognition.

In fact, Chomsky's "Knowledge of Language" (1986) shows considearble disdain for the rule-guided/rule-describeable distinction, a stance which seems to stem from the assumption that all the hard work, at least when it comes to syntax, is discovering and describing the salient regularities. However, if successful, such an account would have