Observations on Cognitive Judgments

David McAllester
dam@ai.mit.edu

Abstract

It is obvious to anyone familiar with the rules of the game of chess that a king on an empty board can reach every square. It is true, but not obvious, that a knight can reach every square. Why is the first fact obvious but the second fact not? This paper presents an analytic theory of a class of obviousness judgments of this type. Whether or not the specifics of this analysis are correct, it seems that the study of obviousness judgments can be used to construct integrated theories of linguistics, knowledge representation, and inference.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the work described in this paper was provided in part by Misubishi Electric Research Laboratories, Inc. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124.

This paper appeared in AAAI-91. A postscript electronic source for this paper can be found in ftp.ai.mit.edu:/pub/dam/aaai91a.ps. A bibtex reference can be found in ftp.ai.mit.edu:/pub/dam/dam.bib.

1 Introduction

Consider the following two statements about the game of chess.

A king on an empty board can reach every square.

* A knight on an empty board can reach every square.

The first statement is clearly true. The second statement, while true, is not obvious. There is an analogy between the notion of an "obvious" statement and the notion of a grammatical sentence. By analogy with linguistic practice, an asterisk has been written in front of the second statement to indicate that it is not obvious.

The classification of a given statement as being either obvious or non-obvious will be called a *cognitive judgement*. In this paper we investigate the possibility of constructing analytic theories of cognitive judgments analogous to analytic theories of grammar — theories that predict which statements are obvious and which statements are not. One particular theory of a class of cognitive judgments, so called "inductive" judgments, is given in this paper.

Before considering a particular analytic theory of cognitive judgments, it useful to consider some further examples.

> Consider a graph with colored nodes such that every arc connects two nodes of the same color. Any two nodes connected by a path of arcs are the same color.
>
> * Any graph with five nodes and five edges contains a cycle.
>
> A five inch by six inch rectangle can be divided into squares where each square is one inch on a side.
>
> * A five inch by six inch rectangle can be divided into squares where each square is larger than one inch on a side.

Intuitively, a statement is obvious if it is *immediate* — one judges it to be true without experiencing intervening thoughts. The unstarred examples given above are obvious in this sense. If a statement is not immediately true, but can be seen to be true by considering some number of cases, examples, or other statements, then the statement is not obvious. The starred examples are not obvious.

Obviousness can not be equated with truth — many nonobvious statements are true, such as the starred statements above, and some obvious statements are not true. As an example of an obvious statement that is not true consider the statement "in a finite interval of time, a bouncing ball can only bounce a finite number of times." One can at least argue that this statement is false. This statement requires the additional assumption that there is a lower limit on the time taken by an individual bounce. In fact, to a close approximation, the time taken by successive bounces of a bouncing ball decreases geometrically with each bounce. This approximation predicts an infinite number of bounces in a finite amount of time. The mathematical model of geometrically decreasing bounce time is self-consistent and provides a counterexample to the statement. Fortunately, it seems possible to construct a predictive theory of cognitive judgments independent of the semantic notion of truth.

Linguistic theories of grammaticality are *generative* — an infinite set of grammatical sentences is generated by a finite grammar. Theories of cognitive judgments can also be generative. A *sequent* is an expression of the form $\Sigma \vdash \Phi$ where Σ is a set of formulas (premises) and Φ is a formula that may or may not be derivable from Σ. A cognitive judgement can be formalized as a sequent plus a specification of whether that sequent is obvious or nonobvious. We say that a set of inference rules *generates* the sequent $\Sigma \vdash \Phi$ if Φ can be derived from Σ using those rules. In order for a rule set to be a good predictor of cognitive judgments it should generate all, and only, the obvious sequents.

The theory of cognitive judgments presented here is linguistic — it is based on a particular knowledge representation language, a knowledge base, and inference rules. Of course, one can imagine non-linguistic theories of obviousness — for example, a theory based on "visual" processing. It remains to be seen whether one can find image-processing theories of cognitive judgments with the same predictive power as linguistic theories.

The remainder of this paper can be divided into two parts. The first discusses local rule sets and their role in theories of cognitive judgments. The general concept of a local rule set was introduced in [McAllester, 1990]. The second part of the paper discusses a class of "inductive" cognitive judgments. An inductive cognitive judgement consists of an obvious sequent where a formal (syntactic) derivation of the sequent appears to require reasoning by mathematical induction. Although no local rule set has been found that incorporates a rule for mathematical induction, aspects of the theory of local rule sets can be used to construct a formal theory of inductive cognitive judgments.

2 Local Inference Rules

In linguistic theories of syntax it usually easy to determine whether or not a given string of words can be generated by a given grammar. For example, given any particular context-free grammar one can determine whether or not a given string is generated by that grammar in n^3 time where n is the length of the string. Most well known sets of inference rules are different from grammars in the sense that it is difficult to determine if a given sequent is generated by the inference rules. Inference, unlike parsing, tends to be computationally intractable.

The apparent computational intractability of inference is a problem in the theory of cognitive judgments for two reasons. First, one can argue that it is psychologically implausible that the human set of obvious statements is computationally intractable. Second, and perhaps more significantly, a computationally intractable theory is difficult to test against observed data. Given a theory of cognitive judgments, and a sequent that is judged to be non-obvious, one must show that the theory does not generate the sequent. For a complex rule set this can be difficult.

Fortunately, there is a class of sets of inference rules, the local rule sets, that are analogous to context free grammars — for a given local rule set one can determine, in polynomial time in the size of a statement, whether or not that statement is generated by the rule set. Let R be a set of inference rules. The following definitions and lemma are from [McAllester, 1990].

> **Definition:** We write $\Sigma \mathrel{\vdash\!\!\!\dashv}_R \Phi$ if there exists a proof of Φ from the premise set Σ such that every *proper subexpression* of a formula used in the proof appears as a proper subexpression of Φ, a proper subexpression of some formula in Σ, or as a closed (variable free) expression in the rule set R.

> **Lemma:** For any fixed rule set R, there exists a procedure for determining whether or not $\Sigma \mathrel{\vdash\!\!\!\dashv}_R \Phi$ which runs in time polynomial in the written length of Σ and Φ.

We write $\Sigma \vdash_R \Phi$ if there is exists any proof of Φ from Σ using the inference rules in R. The inference relation $\mathrel{\vdash\!\!\!\dashv}_R$ is a restricted version of $\vdash_R$. For any rule set R, the relation $\mathrel{\vdash\!\!\!\dashv}_R$ is polynomial time decidable. If the relation $\vdash_R$ is intractable, as is the case for any sound and complete set of rules for first order logic, then the polynomial time relation $\mathrel{\vdash\!\!\!\dashv}_R$ will be weaker than the relation $\vdash_R$.

> **Definition:** The rule set R is called *local* if the relation $\mathrel{\vdash\!\!\!\dashv}_R$ is the same as the relation $\vdash_R$.

An immediate consequence of the above definitions and lemma is that local rule sets are tractable, i.e., they generate polynomial time decidable inference relations. A variety of nontrivial local rule sets is presented in [McAllester, 1990]. An application of local rule sets to Schubert's steamroller is described in [Givan *et al.*, 1991].

3 Inductive Cognitive Judgments

There is a class of cognitive judgments, that I will call *inductive judgments*, which appear to be most simply analyzed by hypothesizing inference rules for mathematical induction. Consider the following examples, some of which are given above.

> By walking north, a person can never get south of where they started.
>
> If a maze containing a rat is placed in a sealed box then, no matter where the rat runs in the maze, it will not get outside of the box.
>
> Consider a graph with colored nodes such that every arc connects two nodes of the same color. Any two nodes connected by a path of arcs are the same color.
>
> A scrambled Rubic's cube is solvable, i.e., there exists a sequence of moves that will unscramble the cube.
>
> Given a bag of marbles, if marbles are removed one at a time, eventually the bag will be empty.
>
> A king, on an empty chess board, can reach every square.

To construct a set of inference rules that generates each of the above obvious statements, one must ask how these statements might be syntactically derived. The first three judgments can be seen as special cases of the following general principle.

> For any action A and property P, if P is true in the initial state, and, in any state where P is true, P remains true after performing action A, then P will be true in any state resulting from any number of applications of A to the initial state.

This general principle can be used to account for the walking north example if we assume that "not south of the initial position" is a property and "walk north" is an action. In the rat and maze example, "in the box" is a property preserved by "moving in the maze". In the colored graph example, "being the initial color" is a property that is preserved by "crossing an arc in the graph". The general principle, as stated above, is virtually isomorphic to the statement of the induction principle for natural numbers. Although the last three judgments do not appear to be direct applications of the above general principle, they all correspond to statements whose formal derivation appears to involve mathematical induction. The Rubic's cube statement can be proved by induction on the number of moves used to scramble the cube. The bag-of-marbles statement can be proved by induction on the number of marbles in the bag. The king-on-a-chess-board statement can be proved by induction on the distance between the king and a target square.

4 Polynomial Time Inductive Inference

This section gives a rule set that includes an inference rule for mathematical induction and that can be used to provide at least a partial analysis of each of the obvious statements given in the previous section. Although the rule set is not local, the theoretical framework of local inference relations can be used to construct a polynomial time inference procedure based on this nonlocal rule set.

The inference rules are stated in a particular knowledge representation language. Although a denotational semantics is not required for a formal theory of cognitive judgments, the inference rules are much easier to "understand" and remember if such a semantics is provided. The knowledge representation language given here has been designed to be the simplest possible language in which an induction rule can be incorporated into a local rule set. The language contains a Kleene star operation to express an indeterminate number of iterations of an operation. The induction rule is similar to the induction rule of propositional dynamic logic [Pratt, 1976], [Harel, 1984], [Kozen and Tiuryn, 1990]. The language described here is also closely related to the knowledge representation language described in [McAllester *et al.*, 1989].

The classical syntax for first order logic involves two grammatical categories — formulas and terms. The knowledge representation language described here also involves two syntactic categories — formulas and class expressions. Formulas denote truth values and class expressions denote sets.

- A *class expression* is one of the following.
  - A class symbol.
  - An expression of the form $(R\ C)$ where R is a binary relation symbol and C is a class expression.
  - An expression of the form $(R^*\ C)$ where R is a binary relation symbol and C is a class expression.
- A *formula* is an expression of the form $(\texttt{every}\ C\ W)$ where C and W are class expressions.

A *semantic model* of the language defined above consists of an assignment of a set to every class symbol and an assignment of a binary relation (a set of pairs) to every binary relation symbol. If $\mathcal{M}$ is a semantic model then we write $\mathcal{V}(e, \mathcal{M})$ for the semantic value of the expression e in the model $\mathcal{M}$. If C is a class expression then $\mathcal{V}(C, \mathcal{M})$ is a set and, if Φ is a formula, $\mathcal{V}(\Phi, \mathcal{M})$ is a truth value, either $\mathbf{T}$ or $\mathbf{F}$. The semantic value function $\mathcal{V}$ is defined as follows.

- If P is a class symbol then $\mathcal{V}(P,\ \mathcal{M})$ is the set that is the interpretation of P in $\mathcal{M}$.
- $\mathcal{V}(\texttt{(}R\ C\texttt{)},\mathcal{M})$ is the set of all d such that there exists an element d' in $\mathcal{V}(C,\mathcal{M})$ such that that the pair $<d,\ d'>$ is an element of the relation denoted by R.
- $\mathcal{V}(\texttt{(}R^*\ C\texttt{)},\mathcal{M})$ is the union of $\mathcal{V}(C,\ \mathcal{M})$, $\mathcal{V}(\texttt{(}R\ C\texttt{)},\mathcal{M})$, $\mathcal{V}(\texttt{(}R\ \texttt{(}R\ C\texttt{))},\mathcal{M})$, $\mathcal{V}(\texttt{(}R\ \texttt{(}R\ \texttt{(}R\ C\texttt{)))},\mathcal{M})\ldots$.
- $\mathcal{V}(\texttt{(every}\ C\ W\texttt{)},\ \mathcal{M})$ is $\mathbf{T}$ if $\mathcal{V}(C,\ \mathcal{M})$ is a subset of $\mathcal{V}(W,\ \mathcal{M})$.

As an example, suppose that `a-red-node` is a class symbol that denotes the set of all the red nodes in some particular graph. Suppose that `a-neighbor-of` denotes the binary relation that contains the pair $<d,\ d'>$ just in case d and d' are nodes of the graph and there is an arc between d and d'. In this case the class `(a-neighbor-of a-red-node)` denotes the set of all nodes that are one arc away from a red node. The class `(a-neighbor-of`* `a-red-node)` is the set of all nodes that can be reached by crossing zero or more arcs from a red node.

Figure 1 gives a sound set of inference rules for the above knowledge representation language.[1] For ease of exposition, let $\texttt{(}R^n\ C\texttt{)}$ abbreviate $\texttt{(}R\ \texttt{(}R\ \cdots\ \texttt{(}R\ C\texttt{)))}$ with n occurrences of R. The expression $\texttt{(}R^*\ C\texttt{)}$ denotes the union over all $n \geq 0$ of $\texttt{(}R^n\ C\texttt{)}$. Inference rules 5 and 6, together with rules 2 and 3, ensure that for any $n \geq 0$ we have $\texttt{(every (}R^n\ C\texttt{) (}R^*\ C\texttt{))}$. Inference rule 7 is an induction rule. Consider the statement that if every neighbor of a red node is red, then every node connected by some path of arcs to a red node is also red. This statement contains a premise equivalent to the formula

`(every (a-neighbor-of a-red-node) a-red-node).`

Inference rule 7 allows us to immediately conclude the formula

`(every (a-neighbor-of`* `a-red-node) a-red-node).`

Let I (for Induction) be the rule set given in figure 1. Recall that the inference relation $\vDash_I$ is a restricted version of the inference relation $\vdash_I$. The restricted inference relation $\vDash_I$ is polynomial time decidable. By definition, the rule set I is local if and only if these two relations are the same. Unfortunately, the rule set I is not local. In particular we have

$$\{\texttt{(every}\ A\ B\texttt{)},\ \texttt{(every (}R\ B\texttt{)}\ B\texttt{)}\}\ \vdash_I\ \texttt{(every (}R^*\ A\texttt{)}\ B\texttt{)}$$

[1] These rules are apparently not semantically complete. The formula $\texttt{(every (}R^*\ A\texttt{)}\ B\texttt{)}$ semantically follows from $\texttt{(every}\ A\ B\texttt{)}$, $\texttt{(every}\ A\ C\texttt{)}$, $\texttt{(every (}R\ B\texttt{)}\ C\texttt{)}$, and $\texttt{(every (}R\ C\texttt{)}\ B\texttt{)}$. However, there appears to be no proof using the above inference rules. A proof could be constructed, however, if we allow intersection class expressions with appropriate inference rules for intersection. In that case we could show that R preserves the intersection of B and C.

(1) (every C C)

(2) $\dfrac{\texttt{(every } C\ W\texttt{)} \quad \texttt{(every } W\ Z\texttt{)}}{\texttt{(every } C\ Z\texttt{)}}$

(3) $\dfrac{\texttt{(every } C\ W\texttt{)}}{\texttt{(every (}R\ C\texttt{) (}R\ W\texttt{))}}$

(4) $\dfrac{\texttt{(every } C\ W\texttt{)}}{\texttt{(every (}R^*\ C\texttt{) (}R^*\ W\texttt{))}}$

(5) (every C (R^* C))

(6) (every (R (R^* C)) (R^* C))

(7) $\dfrac{\texttt{(every (}R\ C\texttt{) } C\texttt{)}}{\texttt{(every (}R^*\ C\texttt{) } C\texttt{)}}$

Figure 1: Some inference rules

but

$$\{\texttt{(every } A\ B\texttt{)}, \texttt{(every (}R\ B\texttt{) } B\texttt{)}\} \not\Vdash_I \texttt{(every (}R^*\ A\texttt{) } B\texttt{)}.$$

The problem is that the proof underlying the first sequent involves the formula (every (R^* B) B) (this is derived from inference rule 7 and the desired result can then be derived from inference rules 4 and 2). Unfortunately, the formula (every (R^* B) B) is not local — the class expression (R^* B) does not appear in the desired sequent. Since only local formulas are allowed in proofs underling the inference relation $\Vdash_I$, this proof can not be used to generate the second sequent listed above.

A second set of inference rules is given in figure 2. Let I' be the rule set given in figure 2. For sequents that do not involve formulas of the form (preserves R C), the inference relation $\vdash_{I'}$ is equivalent to the inference relation $\vdash_I$. However, the restricted relation $\Vdash_{I'}$ is considerably more powerful than the restricted relation $\Vdash_I$. For example, we have

$$\{\texttt{(every } A\ B\texttt{)}, \texttt{(every (}R\ B\texttt{) } B\texttt{)}\} \Vdash_{I'} \texttt{(every (}R^*\ A\texttt{) } B\texttt{)}.$$

As with all locally restricted inference relations, the inference relation $\Vdash_{I'}$ is polynomial time decidable. We can take $\Vdash_{I'}$ as a generative theory of obvious inductive sequents, although any theory with reasonable coverage of the actual obvious sequents would require a richer knowledge representation language and more inference rules.

Unfortunately, the expanded rule set I' is still not local — there are sequents generated by $\vdash_{I'}$ that are not generated by $\Vdash_{I'}$. However, these examples are difficult to find and seem to have little, if any, significance in practice. It is not known whether the rule set I' can be further expanded to a truely local rule set.

(1) $\texttt{(every } C\ C\texttt{)}$

(2) $$\frac{\begin{array}{l}\texttt{(every } C\ W\texttt{)}\\ \texttt{(every } W\ Z\texttt{)}\end{array}}{\texttt{(every } C\ Z\texttt{)}}$$

(3) $$\frac{\texttt{(every } C\ W\texttt{)}}{\texttt{(every (}R\ C\texttt{) (}R\ W\texttt{))}}$$

(4) $$\frac{\texttt{(every } C\ W\texttt{)}}{\texttt{(every (}R^*\ C\texttt{) (}R^*\ W\texttt{))}}$$

(5) $\texttt{(every } C\ \texttt{(}R^*\ C\texttt{))}$

(6) $$\frac{\begin{array}{l}\texttt{(every } C\ W\texttt{)}\\ \texttt{(every (}R\ W\texttt{) } C\texttt{)}\end{array}}{\texttt{(preserves } R\ C\texttt{)}}$$

(7) $$\frac{\begin{array}{l}\texttt{(every } C\ W\texttt{)}\\ \texttt{(preserves } R\ W\texttt{)}\end{array}}{\texttt{(every (}R\ C\texttt{) } W\texttt{)}}$$

(8) $$\frac{\begin{array}{l}\texttt{(every } C\ W\texttt{)}\\ \texttt{(every } W\ C\texttt{)}\\ \texttt{(preserves } R\ C\texttt{)}\end{array}}{\texttt{(preserves } R\ W\texttt{)}}$$

(9) $\texttt{(preserves } R\ \texttt{(}R^*\ C\texttt{))}$

(10) $$\frac{\begin{array}{l}\texttt{(every } C\ W\texttt{)}\\ \texttt{(preserves } R\ W\texttt{)}\end{array}}{\texttt{(every (}R^*\ C\texttt{) } W\texttt{)}}$$

Figure 2: An equivalent, more nearly local, rule set

It seems likely that one can construct large local rule sets that include rules for mathematical induction. Such rule sets, or even "nearly local" rule sets such as that given in figure 2, may have important engineering applications in areas such as automatic program verification.

5 Conclusion

Cognitive judgments, i.e., judgments about whether a given statement is obviously true, can be viewed as a source of data about the structure of human cognition. Although there is a rich source of fairly unambiguous cognitive judgments, it appears impossible to gain direct introspective access to the underlying computational mechanisms. On the other hand, it does appear possible to construct generative analytic theories of these judgments.

Analytic theories of cognitive judgments can be viewed as being analogous to analytic theories of grammaticality. Local rule sets are analogous to context free grammars in that the there exists a procedure for determining, in polynomial time in the size of a sequent, whether or not the rule set generates that sequent. Local rule sets provide a formal framework for the construction of linguistic theories of cognitive judgments.

This paper is, at best, only a first step in the construction of compelling predictive theories of cognitive judgments. A richer language is clearly needed for expressing generative inference rules. A theory is needed of the translation of English sentences into formulas of the internal knowledge representation language. Ideally, the knowledge representation language used to express cognitive inference rules should be the same as the language used to express the logical form (semantic representation) of natural language statements. This would allow existing theories of logical form to be used in constructing theories of cognitive judgments. It remains to be seen whether the basic approach outlined here can lead to a convincing integrated theory of linguistic logical form and cognitive judgement data.

References

[Adjuciewicz, 1935] Kasimierz Adjuciewicz. Die syntaktische konnexitat. *Studia Philophica*, 1:1–27, 1935. Translated as "Sytactic Connection" in Strolls McCall (ed), *Polish Logic: 1920-1939* (Oxford University Press, 1967).

[Bach, 1988] Emmon Bach. Categorial grammars as theories of language. In Richard Oehrle and Edmond Bach, editors, *Categorial Grammars and Natural Language Structures*, pages 17–34. D. Reidel, 1988.

[Brachman and Schmolze, 1985] Ronald Brachman and James Schmolze. An overview of the kl-one knowledge representation system. *Computational Intelligence*, 9(2):171–216, 1985.

[Givan *et al.*, 1991] Robert Givan, David McAllester, and Sameer Shalaby. Natural language based inference procedures applied to schubert's steamroller. In *AAAI-91*, July 1991.

[Harel, 1984] David Harel. Dynamic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic II: Extensions of Classical Logic*, pages 497–604. Reidel, 1984.

[Kozen and Tiuryn, 1990] D. Kozen and J. Tiuryn. Logics of programs. In J. Van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, pages 789–840. MIT Press, 1990.

[McAllester *et al.*, 1989] D. McAllester, R. Givan, and T. Fatima. Taxonomic syntax for first order inference. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 289–300, 1989. To Appear in JACM.

[McAllester, 1990] D. McAllester. Automatic recognition of tractability in inference relations. Memo 1215, MIT Artificial Intelligence Laboratory, February 1990. To appear in JACM.

[Nebel, 1988] Bernhard Nebel. Computational complexity of terminological reasoning in back. *Artificial Intelligence*, 34(3):371–384, 1988.

[Pratt, 1976] V. Pratt. Semantical considerations on floyd-hoare logic. In *FOCS76*, pages 109–121, 1976.