Default Defeaters in Explanation-Based Reasoning

Gilbert Harman, Department of Philosophy, Princeton University
Richard Cullingford, Information and Computer Science, Georgia Institute of Technology
Marie Bienkowski, Bell Communications Research, Morristown, New Jersey
Ken Salem, Department of Computer Science, Princeton University
Ian Pratt, Department of Philosophy, Princeton University

The purpose of this paper is to illustrate an approach to the theory of reasoning that takes all reasoning to be "explanation-based". In particular, we consider how to treat "default reasoning" as a special case of explanation-based reasoning and we indicate what implications this treatment of default reasoning has for handling cases where the legitimacy of default reasoning is defeated by special considerations.

We are particularly interested in the following question about default reasoning. Given a default principle of the form, "Normally A's are B's," one can normally infer that a given A is a B. But sometimes further information about an A can block this inference. The question is: How should the rules of inference accommodate these exceptional cases?

One method that is used in certain production systems is to have several rules, one for the default rule and one for each of the exceptional cases: "From x is an A, infer x is a B." "From x is an A and x is a C, infer x is not a B." Etc. Furthermore, a restriction is placed on the rules of inference saying that, if the left hand side of a rule R is satisfied, one can use R only if there is no satisfied rule whose left hand side includes all the conditions of R's left hand side plus some further conditions. Given A only, one can then use the first rule to infer B. But given A and C one cannot use the first rule, since the second rule's left hand side is now satisfied.

This method supposes that one has already discovered whether the case is exceptional before deciding whether to infer from "x is an A" to "x is a B" using the default rule. This does not account for the case in which the current evidence would allow the inference that x is an exception but this has not yet been inferred. McDermott and Doyle (1980) handle this case by using rules of the following form: "Given that x is an A and that x cannot be inferred to be a C, infer that it is a B." But even this approach does not handle a case in which the evidence indicates that there is a significant chance that x is a C, without being so strong as to allow the inference that x is a C.

The research reported here was supported in part by a research grant from the James S. McDonnell Foundation, by a research grant (487906) from IBM, by the Defense Advanced Research Projects Agency of the Department of Defense and by the Office of Naval Research under Contracts Nos. N00014-85-C-0456 and N00014-85-K-0465, and by the National Science Foundation under Cooperative Agreement No. DCR-8420948 and under NSF grant number IST8503968. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the McDonnell Foundation, IBM, the Defense Advanced Research Projects Agency or the U.S. Government.

In some cases of this sort, the possibility of x's being a C is relevant because the reason why A's are normally B's is that A's are normally C's, and C's are always B's. We suggest that in order to infer from x's being an A to x's being a B one must be able simultaneously to infer that x is a C. If the evidence indicates that there is a significant chance that x is not C, then it will not be possible to conclude that x is C, and that will prevent the inference that x is B. As we will now indicate, this way of handling certain default defeaters fits in with a general framework of explanation-based reasoning.

REASONING

A preliminary account of explanation-based reasoning occurs in Harman (1986). Reasoning is identified with a nonmonotonic process of "change in view". Such a change occurs only in the presence of an interest or goal of the agent, for example, an interest in the answer to a particular question. The process of reasoning tries to respond to this interest or goal by making a minimal change in the agent's beliefs that improves the explanatory coherence of the whole set of beliefs by addition to and subtraction from that set. The process is subject to a number of constraints discussed in Harman (1986) that will not be discussed here.

We envision a computer program, AR (for Artificial Reasoner), that modifies representations of beliefs in accord with the principles of explanation-based reasoning (Cullingford, et al., 1985).. When AR draws a new conclusion, this conclusion will normally take the form of a complex explanatory structure, in which beliefs are linked together by relations of intelligibility. Sometimes, explanation-based reasoning will involve new beliefs that are inferred as the best explanation of the truth of certain old beliefs. For example, when doctor AR infers that a patient has a particular disease, AR's conclusion is that the patient's having this disease explains why the patient has such and such symptoms. Sometimes explanation-based reasoning will introduce new beliefs whose truth is inferred to be explained by certain old beliefs. For example, when predictor AR predicts that an agent will do a particular action, AR's conclusion is that such and such motives will lead the agent to do that action and so will explain the agent's action. Other more complex cases are discussed in Harman (1986).

The unit of inference is an explanatory structure. When AR is considering whether or not to accept a given conclusion C, AR must consider whether there is some explanatory structure AR can accept of which C is a part. So, AR will have some principles for determining what possibly acceptable explanatory structures there are. AR will also have principles for deciding among competing explanatory structures (e.g. for choosing the best explanation). (These principles need not be wholly separate, because the possibly acceptable explanatory structures might be produced in an order that indicates how good they are. For example, there might be a preference for explanatory structures that involve fewer rather than more new beliefs. There might also be a preference for explanatory structures that account for more rather than less.)

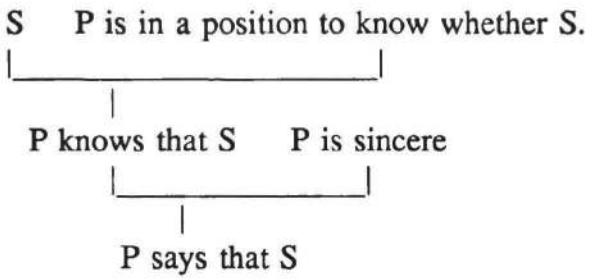

FIGURE 1: A (SLIGHTLY) COMPLEX EXPLANATION

STRUCTURES

In this view, beliefs are organized into complex explanatory structures. The basic links in such structures are immediate explanations. Immediately intelligible links represent connections that AR grasps without having to note intermediate links. An immediate explanation or *e-node* has two components, an *explanans* or list of (pointers to) immediately explaining beliefs and an *explanandum* or (pointer to) something immediately explained. Every belief is associated with *explainer* links to e-nodes of which the belief is the explanandum and *explained* links to e-nodes of which the belief is one of the explanans.

A complex explanation is a structure of immediate explanations, perhaps a tree, with the *ultimate explanandum* (thing explained) at the root, where it and other propositions in the tree have as immediate descendents e-nodes of which they are the explananda, where these e-nodes have as their descendents the propositions that are the explanans of the e-nodes. (Figure 1)

Some e-structures are more complex than this, since a new hypothesis might allow the explanation of more than one thing. For example, doctor AR should prefer a diagnosis that accounts for several of a patient's symptoms over a diagnosis that accounts for only one symptom. That involves an explanatory structure with more than one root. So we have to allow for cases in which there is more than one e-node below a proposition. It is not clear what to call this structure, but it consists basically in links among propositions and e-nodes.

So, we assume that AR has ɪ ·ocedures that produce possible explanatory structures of this sort containing the proposition it is "considering" and minimizing the number of new beliefs it adds and old beliefs it gets rid of.

INFERRING STRUCTURES

Suppose AR knows that A is F and AR also knows that, normally, x is F only if x is G. But AR does not know why this is so. In particular AR does not know whether something's being F is responsible for its being G, whether something's being G is responsible for its being F, or whether some other thing is responsible for this correlation. Still, AR can infer that A is G. But how is that to be represented as an explanatory inference?

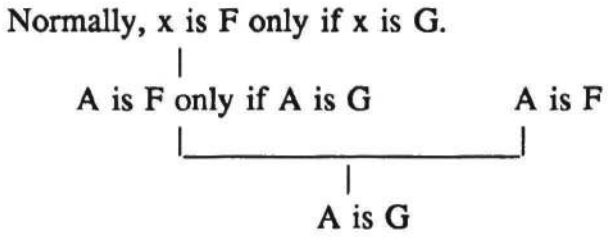

FIGURE 2: EXPLANATION USING A DEFAULT PRINCIPLE

We suggest the following answer: AR infers that the existence of the general correlation between something's being F and something's being G will account for the correlation in this particular case. That is, AR will add an e-node whose single explainer is (a link to) "Normally, x is F only if x is G" and whose explanandum is (a link to) "A is F only if A is G". AR will also add an e-node whose explainers are links to "A is F only if A is G" and "A is F" and whose explanandum is "A is G." These two e-nodes and the propositions they are linked to make up the explanatory structure that AR infers on this occasion. (Figure 2) (This is not to say that AR has to retain this structure in memory as time goes on. Rather: this is what AR accepts for the moment in coming to accept "A is G".)

How does AR infer from "S says that P" to "P"? Perhaps via the generalization, "Usually, x says that m only if m." Then this is an instance of the sort of inference just discussed.

However, AR might also have a view as to why the generalization holds. AR might believe that the generalization holds because, usually, when x says that m, that is because x believes that m and x wants to say whether m, and, furthermore, x believes that m because m and x is in a position to know whether m. It is only because AR accepts such an explanation that AR can avoid inferring "P" from "S says that P" on those occasions on which AR believes S does not want to say whether P or on those occasions on which AR thinks S is not in a position to know whether P. (Eventually, we consider how AR's acceptance of this sort of explanation allows AR to avoid these bad inferences.)

To represent this we need to allow for explanatory structures that link propositions with variables in them. The relevant structure here contains an e-node whose explanandum is (a link to) "x says that m" and whose explainers are (links to) "x believes that m" and "x wants to say whether m". It also contains an e-node whose explanandum is (a link to) "x believes that m" and whose explainers are (links to) "m" and "x is in a position to know whether m." (Figure 3)

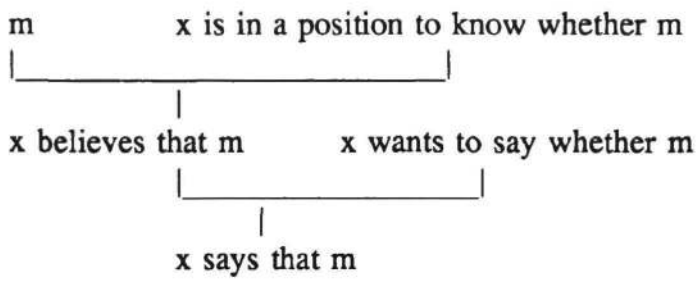

FIGURE 3: EXPLANATORY STRUCTURE WITH VARIABLES

Call this structure ST, then the relevant belief is that, usually, when x says that m, then ST. Notice that this involves quantifiers whose scope is the whole explanatory structure. (Also the phrase, "usually when x says that m," is a kind of quantifier here. Its scope is also the whole structure.)

LIMITED REASONING

The amount of processing required to tell whether a conclusion is currently inferable in N steps is an exponential function of N. Furthermore, anyone who has taught a logic course knows that students often have trouble with more than one or two steps of inference at a time. Proofs have to be broken down into manageable stages, each of which must be absorbed before going on to the next. In general, people are capable of only a few steps of inference at any given time. So we wanted AR to be subject to the same limitations.

Our first implementation of this idea was to limit AR to N steps of immediate implication or immediate inductive projection, where N is 1 or, anyway, small. But the reasoning involved in many elementary activities, e.g. story understanding, often involves rather complex chains of inference.

Reflection on various examples seemed to us to indicate that how many steps of immediate inference are possible depends on how familiar the area is. So, our second implementation of AR replaced the limitation to N steps of inference with a set of rules specifically spelling out what deductions or projections AR was capable of at any given time, where these rules might allow (in principle) for unlimited chains of implication and projection, if the chains were of a "familiar" sort. (What chains of reasoning were to count as familiar dictated the choice of rules. The notion of familiarity did not play an explicit role in the rules.)

Now, in order to make further progress here, we saw we had clearly to distinguish between a step of *inference* (i.e. a step of *belief revision)* and a step of *implication* or *inductive projection.* This distinction is easy to appreciate for the case in which belief revision includes the elimination of some prior belief, because eliminating a belief is clearly different from inferring an implication or projection from prior beliefs. But the distinction is also important for reasoning that does not eliminate any old beliefs, in other words, for reasoning that adds new beliefs that are implications

and/or projections of old beliefs. What gets added in such a case is (we claim) an implicational-explanatory structure. There will often be a complex chain of implication and projection in the structure. In a sense this represents a complex chain of reasoning. But in another sense (we want to say) this might be only one step of reasoning.

For example, a person may accept as a background "belief" a complex general explanatory structure. A single step of inference might involve the acceptance of a particular instance of that general structure, accepted as an instance of that structure. Then the conclusion accepted can involve a number of steps of implication and projection even though there is only one step of inference between the general explanatory belief and that conclusion.

So, we can combine the original idea that there is a limit to the number of steps of reasoning a person can do at any one time with the observation that some reasoning involves a complex chain of considerations, once we distinguish between a step of reasoning and a link in an implicational-explanatory chain.

DEFAULT REASONING

Suppose AR believes -

Normally, if x is an F, x is a G.
Normally, if x is an F and x is an H, x is a Q.
G and Q are contraries.
a1 is an F.

Then AR can accept the conclusion "a1 is a G." More specifically AR can accept the explanatory structure in Figure 4. This involves adding one new belief, namely "a1 is a G", which is linked to beliefs previously accepted.

The competing explanatory structure involves adding two new beliefs (Figure 5). The two new beliefs added here are "a1 is an H" and "a1 is a Q". The explanatory power of this structure is the same as that of the previous structure, so the previous structure is preferred to this one.

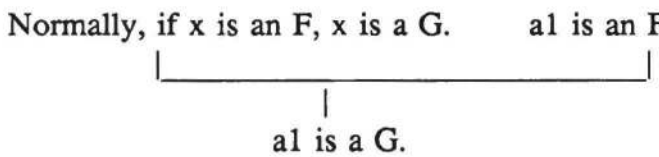

FIGURE 4: DEFAULT STRUCTURE

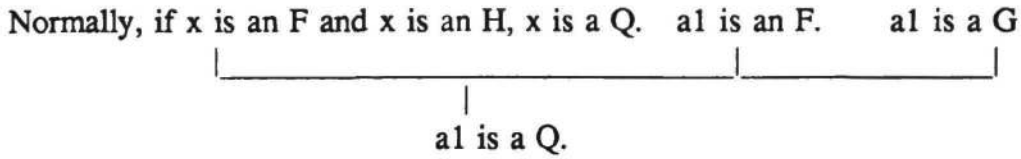

FIGURE 5: COMPETING STRUCTURE

If AR's evidence included both "a1 is an F" and "a1 is an H", then the second structure would be preferred because it links "a1 is an H" to the newly inferred belief, whereas the first structure would only link "a1 is F" to the newly inferred belief.

McDermott and Doyle (1980) would allow the first inference only given the further premises, "I have no reason to infer 'a1 is H'". But this premise is not needed on the explanation-based reasoning approach.

DEFEATERS

Suppose AR believes

(1) Normally, if x is F, then x is G.
(2) a is F.

Then AR ought to be able to infer a is G. But not if AR also believes:

(3) (1) holds because, normally, if x is F, then x is H, and any H is G.

together with other things that prevent AR from using (3) to infer that a is H, either because AR believes things that imply a is not H, or because AR believes things that imply that there is a significant chance that a is not H despite (2) and (3).

Notice that we cannot capture this within a "nonmonotonic logic" by asserting that, when something like (3) is believed, (1) must be replaced with

(4) Normally, if x is F and *it is not inferable that* x is not H, then x is G.

This is not enough of a modification in (1), because it would not prevent the inference in certain cases in which the inference should be prevented, namely, those in which it is not inferable that x is not H but it is also not inferable from (3) that x is H (e.g. because we have other evidence that indicates a significant likelihood that x is not H (for example the evidence might indicate that there is a 50-50 probability that x is not H).

ALGORITHM

Here is a quasi-algorithm that AR uses, given an interest in answering a question. It first forms a list of possible e-structures it might infer that contain an answer. It orders this list in terms of how good these explanations are as measured by the extent

of change required (the more change the worse the explanation) and effect on explanatory connections (the more the better). This is discussed in Harman (1986) without settling on a precise measure. We do not have space here to discuss possible measures. In any event, call the ordered list L. AR takes the first item from L. Call this item I. AR forms a list of competing e-structures that might be inferred. It considers whether I is better than any e-structure in this last list. If so, it infers I. If not, it considers the next item in L and goes back three steps. If L is empty, no answer to the question can be inferred.

INTERSECTION EXAMPLE

We conclude with a different sort of example which we have been examining. Suppose AR starts with an interest in answering the question, "Do the two streets Harrison and Aiken intersect?" It collects a list of e-structures it might infer that contain an answer, using backward chaining. In this case the possible answers are yes and no, i.e. "Harrison and Aiken do intersect" and "Harrison and Aiken do not intersect".

AR discovers the following possible e-structure that it might infer: Harrison and Aiken do intersect, because Harrison and Aiken are near each other and perpendicular to each other, and, normally, when roads are near each other and perpendicular, they intersect. This e-structure is inferable only because AR already believes (1) that Harrison and Aiken are near each other and perpendicular to each other and (2) normally, when roads are near each other and perpendicular, they intersect.

AR discovers no other inferable e-structure.

AR next considers whether there are competing e-structures that might be inferred and discovers none. So AR infers that Harrison and Aiken intersect.

The rules of backwards chaining applicable in a case like this are quite similar to ordinary backward chaining inference rules EXCEPT that they lead to an e-structure containing all the "premises".

It might be good to say a bit more about "near" and "perpendicular". In saying that Harrison and Aiken are near and perpendicular what is meant is that there is a point X on Harrison and a point Y on Aiken such that X is near Y and the orientation of Harrison at X is perpendicular to the orientation of Aiken at Y. AR might know about such points, e.g. the intersection of Aiken and Princeton and the intersection of Harrison and Nassau.

Now suppose that AR starts by believing this:

(*) Normally, if
X is a point on S1,
Y is a point on S2,
X and Y are near each other, and
the orientation of S1 at X is perpendicular to the orientation of S2 at Y,
then: S1 intersects S2.

Then AR comes to believe that (*) holds because

(a) the lines going through X and Y with the orientations of S1 at X and S2 at Y intersect in a point Z that is near or at both X and Y and
(b) normally, given a point P on a road R, the road continues at least a short distance along in the same direction from that point, so that, if P' is near or at P and P' is on the line through P that has the same orientation that R has at P, then R continues from P to R -- so
(c) S1 continues to Z and S2 continues to Z, so
(d) S1 and S2 are both at Z, so
(e) S1 and S2 intersect at Z.

Now, if AR is considering whether Harrison and Aiken intersect, backchaining leads to a much more complicated e-structure, which might be expressed in words as follows: Harrison and Aiken do intersect because there is a point X on Harrison and a point Y on Aiken such that X is near Y and the orientation of Harrison at X is perpendicular to the orientation of Aiken at Y and that means there is a point Z that is near X and near Y and Harrison continues beyond X to Z and Aiken continues beyond Y to Z, where Z is the intersection of the lines going through X and Y that are oriented as Harrison and Aiken are at X and Y, and where all this is so because of the generalizations alluded to above.

This modification of backchaining requires that some change be made to the original generalization linking it to its explanation. The backchaining rule has to be sensitive to this link in such a way that it yields the e-structure just given.

This means that AR will be unable to infer that Harrison and Aiken intersect if it cannot infer the more complex e-structure, perhaps because AR believes that Aiken does not continue on in the indicated way but instead dead ends into a park.

BIBLIOGRAPHY

Richard Cullingford, Gilbert Harman, Marie Bienkowski, & Ken Salem (1985) "Without Logic or Justification: Realistic Belief Revision," *Proceedings of the National Academy of Sciences Workshop on Artificial Intelligence and Distributed Problem Solving*. Washington, D. C. National Academy of Sciences.

Gilbert Harman (1986) *Change In View: Principles of Reasoning*. Cambridge, Massachusetts. M.I.T. Press.

Drew McDermott & Jon Doyle (1980) "Non-monotonic logic, I," *Artificial Intelligence* 13: 41-72.