Assessing Others' Knowledge and Ignorance

Sarit Kraus[*]
Institute for Advanced Computer Studies and
Department of Computer Science
University of Maryland
College Park, MD 20742

Donald Perlis[†]
Department of Computer Science and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

1 Introduction

Autocircumscription [4] was introduced in part to modularize circumscription into a consistency part and a default part.[1] One problem that was solved using autocircumsciption is the 'Ronald Reagan' problem due to McCarthy [private communication]. Here the problem is to come to the conclusion

[*] Supported in part by ARO research contract no. DAAL03-88-K0087.

[†] Supported in part by ARO research contract no. DAAL03-88-K0087, NSF Coordinated Experimental Research grant no. DCR-8320136, and ONR/DARPA research contract no. N00014-82-K-0193. This author was on sabbatical leave at the University of Rochester while working on this paper.

[1] See Lin [1] for a related effort.

that, we do not know whether Ronald Reagan is currently standing or sitting. It amounts precisely to a judgement of our ignorance of certain facts (about Reagan), and so lends itself nicely to autocircumscription, which was designed to test ignorance in the agent's own reasoning.

However, a partner problem, also due to McCarthy, is the Reagan-Gorbachev problem. Here the problem is to come to the conclusion that Mikhail Gorbachev does not know whether Ronald Reagan is currently standing or sitting. Now we are in another area, in which we are trying to assess not our own but rather another's knowledge or ignorance. Clearly this is something we do all the time, and is very useful in commonsense situations.

Originally we thought that autocircumscription would have nothing to offer here, since it is employs a mechanism tied in an essential way to the subjective agent's own knowledge. However, we have found that it does serve as the basis for assessing ignorance on the part of others after all. This is reported in the remainder of the paper, after a short section acquainting the reader with preliminaries. To get a brief look ahead the kind of formal behavior we seek in a system is that it should produce the following kind of conclusions (in the Reagan/Gorbachev case):

I. As a default, we conclude G does not know P, in the absence of special information such as in the cases II-IV below. For instance, if the only information we have about α is un'unrelated' to Reagan, we should not be able to conclude that α orbachev knowess any specific details about Reagan.

II. If we know (or can prove) α does know whether P then we do not conclude he does not know (i.e., we preserve consistency). For example of α and R are shaking hands, then we may simply conclude α knows outright.

III. If something suggests to us that α might know whether P[2], then we

[2] This is somewhat in the spirit of [?]-H's likelihood operator L, when used at a 'low

do not conclude he does not know (yet we also do not get that he knows, at least on these grounds). For instance, if α and R are in the same room, this is suggests a way α might know whether R is sitting, hence we do not want to assume α does not know (or that he does know) without further knowledge. That is, the default about others' ignorance should not apply.

IV. Suppose we can prove α does not know whether P. In such a case it might be either that something suggests to us that α might know, or that nothing suggests to us that he knows. In both cases we will conclude that he doesn't know. For instance if we know or can prove α is asleep, then we conclude directly (i.e not by default) that he doesn't know about current facts.

The rest of the paper is organized as follows: The next section reviews relevant portions of autocircumscription. Then we present a formal system of axioms that demonstrate the four behaviors outlined above. Finally we conclude with some shortcomings of the present results and suggest future directions.

2 Review of Autocircumscription

We recall some details from [4] with some changes. Suppose a set of axioms $A = A[P] = A[P_0, ..., P_n]$ is given and the formula of which ignorance is being tested is represented by $W[P, y]$. Let $K(\alpha, y)$ be a predicate symbol[3] that can be read "α knows y". The autocircumscription schema is:

$$AUTO: \quad A[P'] \wedge \neg W[P', y] \rightarrow \neg K(\iota, W[P, y])$$

level' L^n for large n.

[3] the second argument of K is actually the name of a wff as we will review below

or, in more suggestive terms,

$$A[P'] \wedge V[P', y] \rightarrow Possible(\iota, V[P, y])$$

where now we are testing possibility of V in the form of $\neg W$ and interpreting possibility as "not known to me (i) to be false." (i.e, $Possible(i, w) \equiv \neg K(i, \neg w)$). The idea is (much as in circumscription see [2], [3]) that if there is an interpretation $P'_0, ..., P'_n$ of the predicates $P_0, ..., P_n$ such that $A[P'_0, ..., P'_n]$ holds than this is a possible interpretation for the P's, so that we must have been ignorant of any fact $W[P, y]$ such that $W[P', y]$ happens to fail.

This schema may be compared with one of the more usual schemas of circumscription. For example:

$$A[P'] \wedge (\forall x)(W[P', x] \rightarrow W[P, x]) \wedge \neg W[P', y] \longrightarrow \neg W[P, y]$$

Notice that in autocircumscription, unlike the minimizing versions of circumscription, all wffs W can be ignorance-tested at once. Thus $AUTO$ is a schema in W as well as in P'. Whereas minimizing (non-modular) circumscription contains within it default information (W's are rare events), autocircumscription is much less bold in that it merely records that W is not known to be true. Thus autocircumscription must be supplemented with an explicit default axiom in order to capture non-monotonic reasoning.

In order to use AUTO, the syntax of the underlying first-order language L must include names for formulae, so that they can appear as terms in formulas. Therefore, for each formula w in the language, there is a designated term t_w whose free variables are those of w. Thus t_w is a function symbol with variables (or constant symbol if w has no free variable). Call a language 'reified' if it is so endowed with terms. Since context makes clear the usage, we will simply use w for t_w. Thus Kw is really $K(t_w)$. We also assume that L has the dyadic predicate letter SMK ($SMK(a,b)$ intuitively stands for "something suggests a might know b").

3 The Solution

The idea is to invoke a default axiom whose intuitive reading is that if we do not know anything that suggests a might know b, then assume a does not know b. Formally:

$$\neg K(i, SMK(x,y)) \to \neg K(x,y)$$

We will illustrate it's application to the Reagan/Gorbachev problem, showing that at least in certain examples it accomplishes the three conditions (I,II,III,VI) set out in the Introduction.

To give more motivation: Ignorance is the 'normal' state of affairs. An agent comes to know whether something, P, is true, by interacting in a special way with things that lead him to know whether P. These things together may constitute a sort of determination that P is known, but separately they are mere suggestions toward potential knowing. Thus such can be taken as 'suggesting α might know whether P'. That is, knowledge necessarily depends on there being the appropriately linked pieces leading to the knowledge, and thus there always are suggestions that α might know whether P is in fact α does know whether P. Of course, this is not to say that we or anyone else know about these suggestions. But if someone does, then he ought to refrain from an automatic use of the ignorance default, i.e., ought not to conclude G is ignorant whether P at least on the basis of the default.[4]

[4] In private communication, John McCarthy mentioned the idea of assessing 'the kind

Let L be a reified language such that $Reagan, Gorbachev \in L$. Suppose the agent has the general axioms
$A[InSameRoom, Sit, SMK, K, Unconscious, ShakingHands]$:

$$Sit(x) \wedge InSameRoom(r, x, y) \rightarrow SMK(y, Sit(x) \tag{1}$$

$$\neg Sit(x) \wedge InSameRoom(x, y) \rightarrow SMK(y, \neg Sit(x)) \tag{2}$$

$$Unconscious(y) \rightarrow (\forall x) \neg K(y, x) \tag{3}$$

$$Sit(x) \wedge SakingHands(x, y) \rightarrow K(y, Sit(x)) \tag{4}$$

$$\neg Sit(x) \wedge ShakingHands(x, y) \rightarrow K(y, \neg Sit(x)) \tag{5}$$

$$\neg K(i, SMK(x, y)) \rightarrow \neg K(x, y) \tag{6}$$

$$K(x, y) \rightarrow SMK(x, y) \tag{7}$$

In addition, the agent has some axioms about the ability of other agents to make deductions. For example:
$SMK(p,\text{``}x \rightarrow y\text{''}) \wedge SMK(p,\text{``}y \rightarrow z\text{''}) \rightarrow SMK(p,\text{``}x \rightarrow z\text{''})$
$SMK(p,\text{``}x\text{''}) \wedge SMK(p,\text{``}x \rightarrow y\text{''}) \rightarrow SMK(p,\text{``}y\text{''})$

We can also add to A axioms such as:
$Sit(x) \wedge TalkingOnThePhone(x, y) \rightarrow SMK(y, Sit(x))$ but we will concentrate on the simple case, and assume the only axioms are those given above.

Before we demostrate that the four desired formal behaviors can be captured in this setting we will prove some lemmas and a theorem about A. Let $A*$ is a consistent extention of A such that the new axioms are K-free and do not include ι.

Lemma 1 *$A*$ is autoconsistent i.e. if $A* \vdash K(\alpha, w)$ then $A* \vdash w$.*

of experience α can have' as a critical element in this sort of problem. We agree, and view the SMK construct as an attempt to capture this idea.

Proof: The only axioms that can be used in $A*$ in order to prove a formula of the form $K(\alpha, w)$ are 4, 5 and the default rule (6). From 4, 5 it is clear that $A* \vdash w$. If the default rule is used than w is $SMK(x,y)$ where $A* \vdash K(x,y)$. But in such a case $A* \vdash SMK(x,y)$ by 7.

Lemma 2 *If $A* \vdash \neg K(\iota, w)$ then $A \not\vdash w$.*

Proof: The only axiom that can be used is 3, but $A* \not\vdash Unconcious(\iota)$ We want to remark that this claim is not valid for any α.

Theorem 1 *$AUTO[A*]$ is consistent.*

Proof: We shall construct a model M for $AUTO[A*]$. Let $D = Domain(M)$ contain all formulae and constants of L. We interpret $K(\iota, w)$ to be true in M iff $A* \vdash w$ If $\alpha \neq \iota$ then $K(\alpha, w)$ iff $A* \vdash K(\alpha, w)$. Others predicates and constants are interet as themselves.

Using lemmas 1 and 2 it is is to see that M is a model for $A*$. We shall prove now that M is also a model for $AUTO[A*]$. Any such axiom has the form: $A * [P'] \wedge \neg W[P', y] \rightarrow \neg K(\iota, W[P, y])$. where P stands for all the predicates of $A*$. Assume that this is false in M for some P'. $K(\iota, W[P, y])$ holds in M iff $A* \vdash W[P, y]$. Therefore we have to show that $W[P, y]$ is not a theorem of T. But if $W[P, y]$ were a theorem of $A*$, then $W[P', y]$ would be a consequence of $A[P']$, and so $W[P', y]$ would hold in M. This contradicts the antecedent $\neg W[P', y]$, which we are assuming to hold in M.

It is left to check that the default axiom still holds in M with $AUTO$. So, suppose using $AUTO$ one can prove $\neg K(\iota, SMK(x,y))$ i.e. W is SMK. We have already shown that in such a case $\neg K(\iota, SMK(x,y))$ holds in M. We have to show that $\neg K(x,y)$ also holds in M. Because of the building of M it is enough to show that $A* \not\vdash K(x,y)$. If $A* \vdash K(x,y)$ then $A* \vdash SMK(x,y)$

by 7. But therefore it can't be that $\neg SMK'(x,y)$ is true in M. contrudiction.

$\blacksquare$

Now we proceed to demonstrate that the four desired formal behaviors can be captured in this setting.

I. The default case: Nothing suggests to us that G knows whether P. Here we simply use the above axioms of A.

We choose W[P,y] to be SMK(y,Sit(x)) SMK' to be false, K' to be true and InSameRoom'and Unconscious' be false. The others stay as they are. It is easy to see that $A[InSameRoom', Sit', SMK', K', Unconscious, ShakingHands']$ holds. We may conclude: $\neg false \rightarrow \neg K(i, SMK(y, Sit(x)))$ and we may conclude $\neg K(i, SMK(Gorbachev, Sit(Reagan)))$. Now, using the default rule we get $\neg K(Gorbachev, Sit(Reagan))$.

Similarly we can prove that $\neg K(Gorbachev, \neg Sit(Reagan))$. Those results can be obtained also after adding to A either $Sit(Reagan)$ or $\neg Sit(Reagan)$. It is also easy to see that $AUTO[A]$ is consistent by theorem 1. This illustrates case (I).

II. Now let B be as in A except with the additional axiom $ShakingHands(Gorbachev, Reagan)$. Then we can conclude that $Sit(Reagan) \rightarrow K(Gorbachev, Sit(Reagan)$ and $\neg Sit(Reagan) \rightarrow K(Gorbachev, \neg Sit(Reagan)$ Now suppose we add to B (with out lost of generality) $Sit(Reagen)$, we can conclude $K(Gorbachev, Sit(Reagan)$ and also $SMK(Gorbachev, Sit(Reagan)$. Now using theorem 1 it is easy to check that the negative conclusion in $AUTO[A]$ above is blocked and that $AUTO[B]$ is consistent. This then illustrates case (II).

III. Now let C be an extension of A such that
$C = A + InSameRoom(Reagan, Gorbachev)$.

We now can prove that $Sit(Reagan) \rightarrow SMK(Gorbachev, Sit(Reagan))$ and $\neg Sit(Reagan) \rightarrow SMK(Gorbachev, \neg Sit(Reagan))$.

Suppose we add to C $Sit(Reagan)$. We can conclude that $SMK(Gorbachev, Sit(Reagan))$. On the other hand from AUTO[C] it is not possible to prove that $K(Gorbachev, Sit(Reagan))$.

It is left to check whether $\neg K(Gorbachev, Sit(Reagan))$ can be proved from $AUTO[C]$. The only axiom that could be used to prove this assumption is $\neg K(i, SMK(x,y)) \to \neg K(x,y)$. Therefore if $AUTO[C] \vdash \neg K(Gorbachev, Sit(Reagan))$ then $AUTO[C] \vdash \neg K(i, SMK(Gorbachev, Sit(Reagan)))$. Such a formula can be proved only by using $AUTO$ schema where W is $SMK(Gorbachev, Sit(Reagan))$ and $\neg SMK'(Gorbachev, sit'(Reagan))$ holds. But $C \vdash SMK(Gorbachev, Sit(Reagan))$ and $C[InSameRoom', Sit', SMK', K', Ab']$, therefore $SMK'(Gorbachev, sit'(Reagan))$ also holds, contradiction.

Thus there is no conclusion as to whether Gorbachev knows about Reagan's posture here. This illustrates case (III).

IV. Let D extend A by adding the axiom $Uncouncious(Gorbachev)$. Then we get directly (in A) $\neg K(G, sit(R))$

4 Conclusions

We have shown that assessing another's ignorance by default is possible, using autocircumscription to assess our own ignorance of anything that might suggest the other's knowing a given P. We have illustrated this in several versions, where the outcome is sensitive to details of what we know or can prove about the other person.

However, we point out that these results as they stand do not merge well with certain other desiderata, namely those relating to interacting defaults. That is, we have no device at present for handling situations as G shaking hands with R and yet, for special reasons (e.g., obscured vision) G does not know about R's current posture. This would mean making our other axioms into defaults as well, and then using a prioritized or other taxonomic scheme, perhaps as in [?]cc Lif Ether-Reiter. We also mention Rab-Halp as possible related here. this then is for future work.

References

[1] F. Lin. Circumscription in a modal logic. In *Proc. Theoretical Aspects of Reasoning About Knowledge-88*, pages 113–127, 1988.

[2] John McCarthy. Circumscription–a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[3] John McCarthy. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28:89–118, 1986.

[4] Donald Perlis. Autocircumscription. *Artificial Intelligence*, 36:223–236, 1988.