Reasoning about Ignorance: A Note on the Bush–Gorbachev Problem

Sarit Kraus[1,4]
Donald Perlis[2,4]
John F. Horty[3,4]

[1]Department of Mathematics
and Computer Science
Bar-Ilan University
Ramat Gan 52900
ISRAEL

[2]Computer Science Department
[3]Philosophy Department
[4]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
USA

1 Introduction

Autocircumscription was introduced in [3] as a way of modularizing circumscription into a consistency part and a default part. One problem that was solved there using autocircumscription is the "Bush problem," due to McCarthy: what formal means of commonsense reasoning could enable us to reach the conclusion, ceteris paribus, that we do not at this moment know whether George Bush is standing up or sitting down? The problem involves reaching a correct judgment of our ignorance about certain facts, and so it lends itself nicely to autocircumscription, which was designed to test ignorance in the agent's own reasoning.

The partner problem to the Bush problem is the "Bush–Gorbachev problem," also due to McCarthy.[1] Here the difficulty is to discover a means of commonsense reasoning enabling us to reach the conclusion that, ceteris paribus, Mikhail Gorbachev does not at this moment know whether Bush is standing or sitting. This problem places us in a different arena, because we are trying to assess not our own but rather another's knowledge or ignorance. Clearly, this is something we do all the time in commonsense reasoning.

Originally we thought that autocircumscription would have nothing to offer on the Bush–Gorbachev problem, since it employs a mechanism tied in an essential way to the reasoning agent's own knowledge. However, we have since found that it can serve as the basis for assessing ignorance on the part of others after all.

This paper is organized as follows. We first describe in more detail the kind of behavior that should be expected from a system capable of solving Bush–Gorbachev style problems. We then present autocircumscription, and a related notion of introspective circumscription, due to Lifschitz. Following that, we show how the Bush–Gorbachev problems might be solved using these two formalisms, and we examine an attempt to construct a solution using ordinary circumscription. We conclude with some general points, and a brief discussion of some related literature.

[1] These two problems are not as recent as their names might suggest; they were originally presented as the Johnson problem, and the Johnson-Brezhnev problem.

2 Behavioral specification

We present here four variants on the standard Bush–Gorbachev problem; the point of this is to establish some rough benchmark specifications against which to evaluate a formal system designed to yield appropriate conclusions about the knowledge and ignorance of others.

(I) As a default, in the absence of special information of the kind found in cases (II)–(IV) below, we should conclude that an agent does not know whether a proposition P holds—in particular, that he does not know whether Bush is sitting. (This default should be designed so as not to apply, of course, to the ordinary kinds of knowledge that people are commonly expected to have; see II and III below.)

(II) If we know or can conclude, even by default, that an agent does know whether P holds, then we should refrain from concluding by default that he does not; that is, we should preserve consistency. For example, if we are told that Gorbachev and Bush are at this moment shaking hands, and we conclude from this by default that Gorbachev therefore knows whether Bush is sitting, we should refrain from concluding also that he does not.

(III) If something *suggests* to us that an agent *might* know whether P, then we should in general refrain from concluding by default that he does not. For instance, if we are told that Gorbachev and Bush are in the same room, this gives us a reason to believe that Gorbachev might know whether Bush is sitting, and so the default about his ignorance should not apply.

(IV) Suppose we can prove directly that Gorbachev does not know whether P. (In such a case it might also be that something suggests to us that Gorbachev is likely—or unlikely—to know, or merely that nothing suggests to us that he knows. In all cases we will conclude that he doesn't know.) For instance if we know or can prove that Gorbachev is asleep, then we conclude directly (i.e., not by default) that he doesn't know about current transient facts, even though we might have defaults as in (III) above suggesting otherwise.

3 Autocircumscription

We recall here some details from [3], with some minor 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 $W[P, y]$. We let $K(x, y)$ be a predicate expression that can be read "x knows that y".[2] And throughout, we assume that ι is a constant referring to the individual who is performing the reasoning.

The autocircumscription schema is:

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

This can be presented in more suggestive terms as

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

where now we are testing possibility of V in the form of $\neg W$ and interpreting possibility as "not known to me to be false"— i.e., $\mathit{Possible}(\iota, w) \equiv \neg K(\iota, \neg w)$. The idea, much

[2] The second argument of K is actually the name of a formula as we will review below.

as in ordinary circumscription, is 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, then 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:

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

Notice that in autocircumscription, unlike the minimizing versions of circumscription, all formulas 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 formulas, so that they can appear as terms in other 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 a constant symbol if w has no free variables). For simplicity, and since context can be used to disambiguate, we will simply use w for t_w; thus $K(w)$ is really $K(t_w)$. We also assume that L has the dyadic predicate letter SMK, where a formula of the form $SMK(x, y)$ carries the intuitive meaning "something suggests that x might know that y.

4 Introspective Circumscription

In [2] Lifschitz presents a variation of circumscription similar to autocircumscription, but with an added feature, namely positive introspective (as well as negative introspective which is already present in autocircumscription).

This certainly is an attractive addition, as long as inconsistency is not thereby introduced as pointed out in [4]: logics with both positive and negative introspection, and a modest amount of substitutive textual manipulations such as provided, for instance, by arithmetic, are always inconsistent. However, in the limited form employed by Lifschitz, consistency is preserved. Specifically, he does not have a wholesale positive introspection of all predicates including predicates which themselves are introspection predicates, but rather he isolates positive and negative introspection to a finite number of second order predicate constants specified in advance. That is, for each such specified constant P, there is a companion predicate constant LP (meaning P is believed); but in general there is no constant LLP, etc.

Whether unlimited nesting of predicate companions would lead to the inconsistencies shown in [4] is an open question. Note that autocircumscription does allow unlimited nesting of belief statements, but avoids contradiction by employing only negative, not positive, cases of introspection. Negative cases of introspection are, in general, the ones that give the green light to a pending nonmonotonic conclusion: if not LP (P is not believed) then go ahead and believe Q. Thus it is essential to the very *raison d'etre* of nonmonotonic reasoning, that formulas of the form not-LP be provable. And this is the feature that has been exploited here in the Bush-Gorbachev example.

Since introspective circumscription is—in its semantical sense of logical consequence—stronger than autocircumscription, it follows that it also can handle our four scenarios with the same results, as given in the next section.

5 The Solution

The idea is to supplement autocircumscription with a default axiom whose intuitive reading is that if we do not know anything that suggests person x might know a fact y, then we should assume that x does not know that y:

$$\neg K(\iota, SMK(x,y)) \rightarrow \neg K(x,y).$$

We will illustrate the application of this axiom application to the Bush–Gorbachev problem, showing that it allows us to solve the benchmark problems (I) through (IV) from Section 2. Note especially the constant ι, which reserved as a special symbol for the individual doing the reasoning, and not for reasoners in general. Suppose we find that we ourselves do not know of anything which suggests that x might know that y; then we should be willing to conclude by default that he does not. But we would not want to draw this default conclusion simply from the fact that *some* reasoning agent does not know of anything which suggests that x might know that y.

Ignorance is taken to be the normal state of affairs. An agent x comes to know that some fact P is true by interacting in a special way with things that lead him to know that P. These things together may constitute a sort of circumstance that, if we know about it, can be taken as 'suggesting to us x 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 potential suggestions that x might know whether P. Of course, this is not to say that we or anyone else knows about these suggestions. But if someone does, then she ought to refrain from an automatic use of the ignorance default, i.e., she ought not to conclude G is ignorant whether P at least on the basis of the default.[3]

Now suppose the agent accepts the axiom set

$$A[InSameRoom, Sit, SMK, K, Unconscious, ShakingHands],$$

representing:

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

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

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

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

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

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

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

[3] In private communication, John McCarthy mentioned the idea of assessing 'the kind of experience x 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.

In addition, we can assume that the agent has some axioms about the ability of other agents to make deductions, such as:

$$SMK(p, x \rightarrow y) \wedge SMK(p, y \rightarrow z) \rightarrow SMK(p, x \rightarrow z),$$

$$SMK(p, x) \wedge SMK(p, x \rightarrow y) \rightarrow SMK(p, y).$$

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

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

Proof: The only axioms that can be used in $A*$ in order to prove a formula of the form $K(x, w)$ are (4), (5) and the default rule (6). From (4) and (5) it is clear that $A* \vdash w$, where w is either $Sit(x)$ or $\neg Sit(x)$. If the default rule is used, then 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 $A* \not\vdash Unconscious(\iota)$. (Note here that this claim is correct only for ι, not for any agent x.) ▮

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

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

Using Lemmas 1 and 2 it is easy 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 ranges over 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 construction 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. ▮

We now proceed to demonstrate that the four desired formal behaviors set out above can be captured in this setting.

(I) The default case: nothing suggests to us that Gorbachev knows whether Bush is sitting. Here we simply use the above axioms of A.

We choose $W[P, y]$ to be $SMK(y, Sit(x))$, and SMK', K', $ShakingHands'$, $InSameRoom'$ and $Unconscious'$ to be false. The others stay as they are. It is then easy to see that

$$A[InSameRoom', Sit', SMK', K', Unconscious', ShakingHands']$$

holds. We may thus conclude that

$$\neg false \rightarrow \neg K(\iota, SMK(y, Sit(x))),$$

and so we may conclude $\neg K(\iota, SMK(Gorbachev, Sit(Bush)))$. Now, using the default rule we get $\neg K(Gorbachev, Sit(Bush))$.

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

(II) Now let B be A with the additional axiom

$$ShakingHands(Gorbachev, Bush).$$

Then we can conclude both

$$Sit(Bush) \rightarrow K(Gorbachev, Sit(Bush),$$

$$\neg Sit(Bush) \rightarrow K(Gorbachev, \neg Sit(Bush).$$

Now suppose we add to B (without loss of generality) $Sit(Bush)$; we can then conclude $K(Gorbachev, Sit(Bush)$ and also $SMK(Gorbachev, Sit(Bush)$. 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.

(III) Now let C be A with the addition of the axiom

$$InSameRoom(Bush, Gorbachev).$$

We now can prove that

$$Sit(Bush) \rightarrow SMK(Gorbachev, Sit(Bush)),$$

$$\neg Sit(Bush) \rightarrow SMK(Gorbachev, \neg Sit(Bush)).$$

Suppose we add $Sit(Bush)$ to C. We can then conclude that

$$SMK(Gorbachev, Sit(Bush)).$$

On the other hand from AUTO[C] it is not possible to prove that

$$K(Gorbachev, Sit(Bush)).$$

It is left to see whether $\neg K(Gorbachev, Sit(Bush))$ can be proved from $AUTO[C]$. The only axiom that could be used to prove this assumption is $\neg K(\iota, SMK(x, y)) \rightarrow \neg K(x, y)$. Therefore if

$$AUTO[C] \vdash \neg K(Gorbachev, Sit(Bush)),$$

then

$$AUTO[C] \vdash \neg K(\iota, SMK(Gorbachev, Sit(Bush))).$$

Such a formula can be proved only by using $AUTO$ schema where W is

$$SMK(Gorbachev, Sit(Bush))$$

and

$$\neg SMK'(Gorbachev, Sit'(Bush))$$

holds. But

$$C \vdash SMK(Gorbachev, Sit(Bush))$$

and

$$C[InSameRoom', Sit', SMK', K', Ab'].$$

Therefore

$$SMK'(Gorbachev, Sit'(Bush))$$

also holds, which is a contradiction.

Thus there is no conclusion as to whether Gorbachev knows about Bush's posture here.

(IV) Let D extend A by adding the axiom $Unconscious(Gorbachev)$. Then we get $\neg K(Gorbachev, Sit(Bush))$ directly from from ordinary logic, without any default reasoning at all.

6 Ordinary Circumscription

We will show here that three of the above four behaviors can also be achieved directly with standard (e.g., formula) circumscription.

Cases II and IV are straightforward since no defaults are involved and no circumscription at all is called for to get the indicated results. Case I can be handled as follows:

We simply postulate that others' knowledge is abnormal: $ab(K(x,y))$ for any agent x other than ourselves (ι).[4] Then for all sorts of everyday scenarios, we will have beliefs about the many abnormal cases in which knowledge does (tend to) get into others' heads.

7 Who Is Doing the Reasoning?

But ordinary circumscription will not do for the third behavior, in which something suggests *to us* that Gorbachev might know whether P; and for interesting reasons: information about the identity of the individual who is doing the (circumscriptive and other) reasoning seems essential to the third behavior, and is the hallmark of autocircumscription.

It is commonplace in much of AI to remain silent on who the reasoner is: is it a robot or human actually on the scene? Or a God's-eye reasoner who merely surveys a distant planet? Or merely a program going through its paces? This may not always matter, but in this case it apparently does. Note that related and interesting work of E. Davis [1] on perception and ignorance, for instance, does not make clear who the reasoner is, i.e., what is the perspective of the axioms being proposed? A remote scientist studying others only (and not herself)? A mere program with "interesting" behavior and no particular role in the world (robot or otherwise)?

[4] It is tempting to apply the same reasoning to ourselves, but this amounts to a kind of positive introspection: if we know something then something must suggest it to us and so we know that we know it and we cannot be sanguine about consistency.

Moreover, other work, such as [5], also shows that the first person pronoun requires special treatment in commonsense reasoning.

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 proposition P (or even just ordinary circumscription in three of the four cases considered). 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 such situations as Gorbachev shaking hands with Bush and yet, for special reasons (e.g., obscured vision), not knowing about Bush's current posture. This would mean making our other axioms into defaults as well, and then using some method for prioritizing defaults. Since the issues involved in selecting the correct prioritizations among competing defaults are themselves not currently well understood, we leave the matter for future work.

Acknowledgments

This work has been supported by the National Science Foundation under Grant No. IRI-8907122.

References

[1] E. Davis. Solution to a paradox of perception with limited acuity. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning* (KR-89), Morgan Kaufmann Publishers (1989), pp. 79–82.

[2] V. Lifschitz. Between circumscription and autoepistemic logic. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning* (KR-89), Morgan Kaufmann Publishers (1989), pp. 235–244.

[3] D. Perlis. Autocircumscription. *Artificial Intelligence*, vol. 36 (1988), pp. 223-236.

[4] D. Perlis. Languages with self reference II: knowledge, belief, and modality. *Artificial Intelligence*, vol. 34 (1988), pp. 301–322.

[5] M. Miller and D. Perlis. Proving self-utterances. *Journal of Automated Reasoning*, vol. 3 (1987), pp.329–338.