How to change a person's mind:
Understanding the difference between the effects and consequences of speech acts

Debora Field and Allan Ramsay

Computer Science, Univ. of Liverpool, L69 3BX, UK
Informatics, Univ. of Manchester, PO Box 88, M60 1QD, UK
debora@csc.liv.ac.uk, allan.ramsay@manchester.ac.uk

Abstract

This paper discusses a planner of the semantics of utterances, whose essential design is an epistemic theorem prover. The planner was designed for the purpose of planning communicative actions, whose effects are famously unknowable and unobservable by the doer/speaker, and depend on the beliefs of and inferences made by the recipient/hearer. The fully implemented model can achieve goals that do not match action effects, but that are rather entailed by them, which it does by reasoning about how to act: state-space planning is interwoven with theorem proving in such a way that a theorem prover uses the effects of actions as hypotheses. The planner is able to model problematic conversational situations, including felicitous and infelicitous instances of bluffing, lying, sarcasm, and stating the obvious. [1]

1 Introduction

The motivation for this research was the problem of planning the semantics of communicative actions: given that I want you to believe P, how do I choose what meaning to express to you? The well-documented, considerable difficulties involved in this problem include this: a key player in the ensuing evolution of the post-utterance environment is the **hearer** of the utterance.

First, consider an imaginary robot Rob, designed not for communication, but for making tea. Whenever he is in use, Rob's top-level goal is to attain a state in which there is a certain configuration of cups, saucers, hot tea, cold milk, etc. Rob's plans for making tea are made on the strong assumption that at plan execution time, the cups (and other items) will have no desires and opinions of their own concerning which positions they should take up—Rob expects to be the author of the effects of his actions. [2]

[1] Initially funded by the EPSRC. Recent partial funding under EU-grant FP6/IST No. 507019 (*PIPS*: Personalised Information Platform for Health and Life Services).

[2] notwithstanding impeding concurrent events, sensor failures, motor failures, *etc.*

In contrast, consider the human John, designed for doing all sorts of things besides making tea, including communicating messages to other humans. Imagine John's current goal is to get human Sally to believe the proposition *John is kind*. In some respects, John has a harder problem than Rob. Unlike Rob, John has no direct access to the environment he wishes to affect—he cannot simply implant *John is kind* into Sally's belief state. John knows that Sally has desires and opinions of her own, and that he will have to plan something that he considers might well lead Sally to infer *John is kind*. This means that when John is planning his action—whether to give her some chocolate, pay her a compliment, tell her he is kind, lend her his credit card—he has to consider the many different messages Sally might infer from the one thing John chooses to say or do. Unfortunately, there is no STRIPS operator [13] John can choose that will have his desired effect; he has to plan an action that he expects will **entail** the state he desires.

We considered 'reasoning-centred' planning of actions that entailed goals to be an approach that would enable this difficult predicament to be managed, and implemented a model accordingly. Our planner is, in essence, an epistemic theorem prover that hypothesises desirable actions, and is able to plan to achieve goals that do not match action effects, but that are entailed by the final state. Like John, the planner can have particular communicative goals in mind, and knows that the execution of any single plan could have a myriad different effects on H's belief state, depending on what H chooses to infer.

1.1 Bucking the trend

The main focus of current research in AI planning is on how to reduce the search space required for making plans, and thus, for example, to get Rob the tea-making robot to be able to make his plans fast enough to be of practical use in the real world. Many planners use heuristics, either to constrain the generation of a search space, or to prune and guide the search through the state space for a solution, or both [4,5,18,25]. All such planners succeed by relying on the static effects of actions—on the fact that you can tell by inspection what the effects of an action will be in any situation—which limits their scope in a particular way [4, p. 299]:

> "...if one of the actions allows the planner to dig a hole of an arbitrary integral depth, then there are potentially infinitely many objects that can be created... The effect of this action cannot be determined statically ..."

The class of problems that these planners do not attempt to solve—the ability to plan actions whose effects are not determined statically—was the class that particularly interested us.

2 Planning the semantics of utterances

With our attention firmly fixed on the myriad different effects a single communicative act can have on a hearer's belief state, we concentrated on a (logically) very simple utterance:

“There’s a/an [some object]!”

We devised situations culminating in this utterance which illustrate **sarcasm**, **stating the obvious**, **bluffing**, and **lying**, and developed a planner which could use these tactics. Here is a much-shortened example of a scenario from the model, which leads to the planning of an instance of sarcasm:[3] [4]

Initial state John has been bird-watching with Sally for hours, and so far, they have only seen pigeons. John thinks Sally is feeling bored and fed up. John has some chocolate in his bag. John thinks Sally likes chocolate. John knows lots of rules about how conversation works, and what one can expect a hearer to infer under given conditions.

Goal condition John wants to cheer Sally up.

Solutions John is just thinking about getting out some chocolate to give her, when yet another pigeon lands in a nearby tree. John sees an opportunity to make Sally laugh by means of a bit of sarcasm, and so plans to say to her,

“There’s an albatross!”

John plans (the semantics of) his utterance, expecting that the utterance will have particular ‘effects’ on Sally’s belief state; if John were to perform the utterance, he would not be certain that it had achieved his intention, but he would expect that it probably had. Whether John’s intention would be achieved by this utterance depends on Sally having the ‘right’ set of beliefs (the ones John thinks she has) and making the ‘right’ inferences (the ones John expects her to make).

For example, if John’s utterance “There’s an albatross!” is to be felicitous, the following must happen. Sally must first believe that John has said something that Sally thinks John and Sally mutually believe is false. From this, she must infer that John has flouted a conversational maxim, and consequently that John has attempted to implicate a meaning which is not expressed by the semantics of “There’s an albatross!”. Sally must then infer that the implicature John intends is of humour. Whether or not any of this happens depends on Sally’s beliefs, which John cannot observe, but about which he has beliefs. The formal version of this example contains all the necessary information about the beliefs of John and Sally in this situation for the planner: (i) to be able to plan John’s utterance; and (ii) to additionally deduce whether John’s utterance would be felicitous or infelicitous, if he performed it.

[3] The example is an English paraphrase of a task, written in the model in Prolog code.

[4] An albatross (*Diomedea exulans*) is a huge sea-faring bird, rarely seen from the land.

2.1 Linguistic motivations

Our approach to planning the semantics of utterances was to build on seminal work in speech acts [3,28] and pragmatics [29,15,21]. In contrast to the 'speech acts with STRIPS' approach [6,11,1,2], which is fraught with well-documented difficulties [9,16,26,7,27], we aimed to develop a small set of linguistic acts that were unambiguously identifiable purely by surface linguistic form (after [7]), including 'declare', 'request', and perhaps others—a set of acts with negligible effects (after [27]), and minimal preconditions. We in fact developed a single linguistic act for all contexts.

2.2 Planner design

The planner is essentially an epistemic theorem prover which employs some planning search. The development process we undertook is helpful in understanding the planner's design:

- A state-space search was implemented that searches backwards in hypothetical time from the goal via STRIPS operators (based on foundational work in classical planning [23,24,14,22,13]);
- A theorem prover for FOL was implemented that constructively proves conjunctions, disjunctions, implications, and negations, and employs modus ponens and unit resolution;
- State-space search and theorem proving were interwoven in such a way that:
  - · not only can disjunctions, implications and negations be **proved** true, they can also be **achieved**;
  - · not only can a goal Q be **proved** true by proving $(P \Rightarrow Q) \wedge P$, but Q can also be **achieved** by proving $P \Rightarrow Q$ and achieving P;
  - · a goal can be achieved by reasoning with **recursive** domain-specific rules—thus the planner is able to plan to 'dig holes of arbitrary depths'.
- The theorem prover was transformed into an epistemic theorem prover by incorporating a theory of knowledge and belief suitable for human reasoning about action, so agents make plans according to their beliefs about the world, including their beliefs about others' beliefs.

A goal is proved by assuming the effect of some action is true, on the grounds that the goal would be true in the situation that resulted from performing that action. Hence, a set of actions is computed that might be useful for achieving a goal by carrying out hypothetical proofs, where the hypotheses are the actions whose effects have been exploited.

Here is a simple, non-dialogue example to aid explanation. Consider the achievement of the goal *above(e,f) and on(e,d)*, where *above* is the transitive closure of *on*. First, it is not possible to judge whether the first goal *above(e,f)* is true by inspecting the current state (which contains *on(_,_)* facts but no *above(_,_)* facts), so reasoning is carried out to find out whether it is false. Secondly, in order to achieve *above(e,f)*, something different from an action with an *above(_,_)* expression in its add list is needed. Placing *e* onto *f*, for example, will make *above(e,f)* proveable, but it will also make the achievement of *on(e,d)* impossible. By reasoning with rules that describe the meaning of *above* as the transitive closure of *on*, the planner hypothesises that *on(d,f)* might enable the proof of *above(e,f)* to be completed, and also knows that *on(d,f)* is an effect of action *stack(d,f).* A proof of the preconditions of action *stack(d,f)* is carried out, and the process continues (with backtracking), until a solution is found.

The preference for a backwards planning search was motivated by a defining quality of the communication problem, as epitomised by utterance planning: there are too many applicable actions to make a forwards search feasible. People generally have the physical and mental capabilities to say whatever they want at any moment. This means that the answer to the question 'What **can** I say in the current state?' is something like '**Anything**, I just have to decide what I want to say'. A backwards search is far more suitable than a forwards search under conditions like these.

With this 'reasoning-centred' design, the planner is able to plan an utterance to achieve a goal, 'knowing' that the utterance may or may not achieve the desired effects on *H*, and that the same utterance can have many different effects, depending on *H*'s belief state.

3 Modelling problematic conversations

In the model, utterances are planned according to Grice's Cooperative Principle [15]. Here is an extract from the CP (*ibid* p. 308):

> "[**Quantity**]
> (i) Make your contribution as informative as is required (for the current purposes of the exchange).
> (ii) Do not make your contribution more informative than is required...
> [**Quality**]
> (i) Do not say what you believe to be false.
> (ii) Do not say that for which you lack adequate evidence."

Grice's maxims prescribe a standard for speaker behaviour which S can blatantly contravene ('flout'), thus signalling to H that there is an implicature to be recovered. For instance, in our 'sarcasm' scenario, John's utterance is planned using the following maxim, derived from Grice's first Quality maxim. [5] The first line means, 'If S addresses H by putting Q into the conversational minutes':

```
(1) minute([S], [H], Q)
        and believes(S, believes(H, mutuallybelieve(([H, S]), not(Q))))
    ==> believes(S, believes(H, griceuncoop(S, [H], Q)))
```

[5] The model embodies a 'deduction' model of belief [19], rather than a 'possible worlds' model [17,20]. Thus agents are not required to draw **all** logically possible inferences, and are therefore not required to infer an infinite number of propositions from a mutual belief.

Using this maxim, John reasons that he can get Sally to realise he is flouting a maxim in order to generate an implicature (that he is being 'Grice uncooperative with respect to Q'). But what is the nature of the implicature? This is dealt with by two additional rules: (2), which describes what John thinks Sally believes about the meaning of this kind of maxim-flouting; and (3), a 'general knowledge' rule:

```
(2) believes(john,
        believes(sally,
            (griceuncoop(PERSON2, _PERSON1, Q)
             and mutuallybelieve(([sally,john]), not(Q)))
                ==> funny(PERSON2, re(Q))))
```

```
(3) believes(john,
        believes(sally,
            (funny(PERSON2, re(Q))
                ==> happy(sally))))
```

With these three rules, John can reason that saying something he thinks he and Sally mutually disbelieve will make her laugh, and thus cheer her up, thus achieving his goal. Here is a second maxim from the model, also derived from Grice's CP:

```
(4) minute([S], [H], Q)
        and believes(S, believes(H, mutuallybelieve(([H, S]), Q)))
    ==> believes(S, believes(H, griceuncoop(S, [H], Q)))
```

Using this maxim, and some additional rules, John can plan to flout Quantity maxim 2, and generate an implicature by 'stating the obvious'.

3.1 Modelling deception

Grice's CP seems an excellent formalism for planning and understanding utterances, so long as everyone is committed to obeying it. We know, however, that people **violate** the CP maxims—S contravenes maxims without wanting H to know. For example, lying violates Quality maxim (1) , bluffing violates Quality maxim (2) , and being economical with the truth violates Quantity maxim (1) . However, there is nothing in Grice's maxims to help H deal with the possibility that S may be trying to deceive her. Our solution is to give S and H some further maxims which legislate for the fact that speakers do not necessarily always adhere to the CP, and which enable S to plan to deceive, and H to detect intended deceptions.

3.1.1 Hearer violation maxims

Given that H admits the possibility that S might be trying to deceive her with his utterance, we consider that there are three strong predictors of how H's belief state will change in response to S's utterance of the proposition P:

(5) i **What is *H*'s view of the proposition *P*?**
ii **What is *H*'s view concerning the goodwill of *S*?**
iii **What is *H*'s view of the reliability of *S*'s testimony?**

Consider, for example, an attempt at bluffing:[6]

Initial state John has gone bird-watching with Sally. John is wearing a warm coat, and he thinks that Sally looks cold. John thinks Sally will be impressed by a chivalrous gesture. John thinks Sally is new to bird-watching, and that she is keen to learn about birds. John knows lots of rules about how conversation works, and what one can expect a hearer to infer under given conditions.

Goal condition John wants Sally to be impressed by him.

Solutions John is just thinking of offering Sally his coat to wear, when a huge bird lands in a nearby tree. John isn't quite sure what species the bird is, nevertheless, he decides to try and impress Sally with his bird expertise, and plans to say to her,

"There's a dodo!"

Let us imagine that Sally's answers to three above questions are as follows. Before John performed his utterance:

(6) i Sally believed that the proposition P ("There's a dodo!") was false (because she knew the bird was a buzzard).
Additionally, she did not believe that John thought that they mutually believed P was false.
ii She believed that John was well-disposed towards her.
iii She didn't know whether John was a reliable source of information or not.

After John has said "There's a dodo!", Sally derives the following new set of beliefs from the above set:

(7) i′ Sally still believes that the proposition P ("There's a dodo!") is false.
She **now** believes that John thinks that they mutually believe P is true.
ii′ She still believes that John is well-disposed towards her.
iii′ She **now** believes John is an unreliable source of information.

The mapping of belief set (6) into belief set (7) is determined in the model by a 'hearer violation (HV) maxim'. We call this maxim the 'infelicitous bluff' HV maxim. We have so far implemented eight HV maxims, however, there is clearly scope for many more permutations of all the different possible answers to (6). There are obvious additional refinements that should be made, for example, people do not normally consider others to be reliable sources of information on **all** subjects.

3.1.2 Speaker violation maxims

If S is to succeed in his attempt to deceive H, he will have to take into account how H is going to try and detect his deception. To represent this in the model, S has his own 'speaker violation (SV) maxims', which concern the same issues as the HV maxims, but from the other side of the table, as it were. What S plans to say will depend on which answer he selects from each of these four categories:

[6] A dodo is a large flightless bird that is famously extinct.

(8) i **What is *S*'s view of *H*'s view of various different propositions?**
ii **What is *S*'s own view of the same propositions?**
iii **What is *S*'s view of *H*'s view of the goodwill of *S*?**
iv **What is *S*'s view of *H*'s view of the reliability of *S* as a source?**

Here is an example of an SV maxim from the model:

```
(9) minute([S], [H], Q)
         and believes(S, believes(H, reliable(S)))
         and believes(S, believes(H, well_disposed_towards(S, [H])))
         and believes(S, believes(H, Q or not(Q)))
    ==> believes(S, believes(H, gricecoop(S, [H], Q)))
```

Using this maxim, John can reason that Sally will believe he is being Grice-cooperative, which means Sally will believe that what he is saying is true, even if John does not believe it himself. Thus John is able to plan to lie to Sally by using tactics he hopes will prevent Sally from detecting his attempt to deceive.

4 Epistemic theorem prover

The planner's theorem prover embodies a constructive/intuitionist logic and it proves theorems by natural deduction, chosen in preference to classical logic and its inferencing methods. The way humans do every-day inferencing is, we consider, quite different from the way inferencing is handled under classical logic. In classical logic, for example, and using our general knowledge, we judge the following formulae to be true:

(10) Earth has one moon ⇒ Elvis is dead

(11) Earth has two moons ⇒ Elvis is alive

(12) Earth has two moons ⇒ Elvis is dead

(10) is true simply because antecedent and consequent are both true formulae. We find this truth odd, however, because of the absence of any discernible relationship between antecedent and consequent. (11) and (12) are true simply because the antecedent is false, which seems very counter-intuitive. Even more peculiarly, the following formula is provable in classical logic in all circumstances:

(13) (Earth has one moon ⇒ Elvis is dead) or
(Elvis is dead ⇒ Earth has one moon)

but it feels very uncomfortable to say that it must be the case that one of these implies the other.

In order to avoid having to admit proofs like this, and to be able to do reasoning in a more human-like way, we opted for constructive logic and natural deduction. In order to prove $P \Rightarrow Q$ by natural deduction, one must show that Q is true **when** P is true; if P is not true, constructive logic does not infer $P \Rightarrow Q$. This treatment of implication hints at a relationship between P and Q which is absent from material implication.

4.1 Constructive logic and belief

Taking a constructive view allows us to simplify our reasoning about when the hearer believes something of the form $P \Rightarrow Q$, and hence (because of the constructive interpretation of $\neg P$ as $P \Rightarrow \bot$) about whether she believes $\neg P$. We will assume that $believes(H, P)$ means that H **could** infer P on the basis of her belief set, not that she already does believe P, and we will examine the relationship between $believes(H, P \Rightarrow Q)$ and $believes(H, P) \Rightarrow believes(H, Q)$.

Consider first $believes(H, P) \Rightarrow believes(H, Q)$. Under what circumstances could you convince yourself that this held?

For a constructive proof, you would have to assume that $believes(H, P)$ held, and try to prove $believes(H, Q)$. So you would say to yourself 'Suppose I were H, and I believed P. Would I believe Q?' The obvious way to answer this would be to try to prove Q, using what you believe to be H's rules of inference. If you could do this, you could assume that H could construct a proof of $P \Rightarrow Q$, and hence it would be reasonable to conclude $believes(H, P \Rightarrow Q)$.

Suppose, on the other hand, that you believed $believes(H, P \Rightarrow Q)$, and that you also believed $believes(H, P)$. This would mean that you thought that H had both $P \Rightarrow Q$ and P available to her. But if you had these two available to you, you would be able to infer Q, so since H is very similar to you she should also be able to infer Q. So from $believes(H, P \Rightarrow Q)$ and $believes(H, P)$ we can infer $believes(H, Q)$, or in other words $(believes(H, P \Rightarrow Q)) \Rightarrow (believes(H, P) \Rightarrow believes(H, Q))$.

We thus see that if we take $believes(H, P)$ to mean 'If I were H I would be able to prove P', then $(believes(H, P \Rightarrow Q))$ and $(believes(H, P) \Rightarrow believes(H, Q))$ are equivalent. This has considerable advantages in terms of theorem proving, since it means that much of the time we can do our reasoning by switching to the believer's point of view and doing perfectly ordinary first-order reasoning. If, in addition, we treat $\neg P$ as a shorthand for $P \Rightarrow \bot$, we see that $believes(H, \neg P)$ is equivalent to $believes(H, P) \Rightarrow believes(H, \bot)$. If we take the further step of assuming that nobody believes $\bot$, we can see that $believes(H, \neg P) \Rightarrow \neg believes(H, P)$ (though not $\neg believes(H, P) \Rightarrow believes(H, \neg P)$). We cannot, however, always assume that everyone's beliefs are consistent, so we may not always want to take this further step (note that in possible worlds treatments, we are **forced** to assume that everyone's beliefs are consistent), but it is useful to be able to use it as a default rule, particularly once we understand the assumptions that lie behind it.

References

[1] Allen, J. F. and C. R. Perrault, Analyzing intention in utterances (1980), *AI* 15: 143–78.

[2] Appelt, D. E., *Planning English referring expressions* (1985), *AI* 26: 1–33.

[3] Austin, J. L., *How to do things with words* (1962), Oxford: OUP, 2nd edition.

[4] Blum, A. L. and M. L. Furst, Fast planning through planning graph analysis (1995), in *Proc. 14th IJCAI*, pp. 1636–1642.

[5] Bonet, B. and H. Geffner, Heuristic Search Planner (2000), *AI Magazine* 21(2).

[6] Bruce, B. C., Generation as a social action (1975), in B. L. Nash-Webber and R. C. Schank (eds), *Theoretical issues in natural language processing*, pp. 74–7. Cambridge, Massachusetts: ACL.

[7] Bunt, H., Dialogue pragmatics and context specification (2000), [8], pp. 81–150.

[8] Bunt, H. and W. Black, (eds), *Abduction, belief and context in dialogue: studies in computational pragmatics* (2000), Philadelphia: John Benjamins.

[9] Cohen, P. R. and H. J. Levesque, Rational interaction as the basis for communication (1990), [10], pp. 221–55.

[10] Cohen, P. R., J. Morgan and M. E. Pollack, (eds), *Intentions in communication* (1990), Cambridge, Massachusetts: MIT.

[11] Cohen, P. R. and C. R. Perrault, Elements of a plan-based theory of speech acts (1979), *Cognitive Science* 3: 177–212.

[12] Feigenbaum, E. A. and J. Feldman, Editors, *Computers and thought* (1995), Cambridge, Massachusetts: MIT Press. First published 1963 by McGraw-Hill.

[13] Fikes, R. E. and N. J. Nilsson, STRIPS: A new approach to the application of theorem proving to problem solving (1971), *AI* 2: 189–208.

[14] Green, C., Application of theorem proving to problem solving (1969), in *Proc. 1st IJCAI*, pp. 219–39.

[15] Grice, H. P., Logic and conversation (1975), in P. Cole and J. Morgan, (eds), *Syntax and semantics 3: Speech acts*, pp. 41–58. New York: Academic Press.

[16] Grosz, B. J. and C. L. Sidner, Plans for discourse (1990), [10], pp. 416–44.

[17] Hintikka, J., *Knowledge and belief: An introduction to the two notions* (1962), New York: Cornell University Press.

[18] Hoffmann, J. and B. Nebel, The FF planning system: Fast plan generation through heuristic search (2001), *Journal of AI Research* 14: 253–302.

[19] Konolige, K., *A deduction model of belief* (1986), London: Pitman.

[20] Kripke, S., Semantical considerations on modal logic (1963), in *Acta Philosophica Fennica* 16: 83–94.

[21] Lewis, D., Scorekeeping in a language game (1979), *J. Phil. Logic* 8: 339–59.

[22] McCarthy, J. and P. J. Hayes, Some philosophical problems from the standpoint of artificial intelligence (1969), *Machine Intelligence* 4: 463–502.

[23] Newell, A., J. C. Shaw and H. A. Simon, Empirical explorations with the logic theory machine (1957), *Proc. Western Joint Computer Conference*, 15: 218–239.

[24] Newell, A. and H. A. Simon, GPS, a program that simulates human thought (1963), [12], pp. 279–93.

[25] Nguyen, X. and S. Kambhampati, Reviving partial order planning (2001), in *Proc. IJCAI*, pp. 459–66.

[26] Pollack, M. E., Plans as complex mental attitudes (1990), [10], pp. 77–103.

[27] Ramsay, A., Speech act theory and epistemic planning (2000), [8], pp. 293–310.

[28] Searle, J. R., What is a speech act? (1965), in M. Black, (ed), *Philosophy in America*, pp. 221–39. Allen and Unwin.

[29] Stalnaker, R., Pragmatics (1972), in D. Davidson and G. Harman, (eds), *Semantics of natural language (Synthese Library, Vol. 40)*, pp. 380–97. Dordrecht, Holland: D. Reidel.