Synthese (2018) 195:433–440
https://doi.org/10.1007/s11229-016-1229-3

The Little Nell Problem: reasonable and resolute maintenance of agent intentions

Richmond H. Thomason[1]

Received: 7 May 2016 / Accepted: 18 September 2016 / Published online: 8 October 2016

Abstract The Little Nell Problem was formulated by Drew McDermott in 1982. It reveals unexpected complexities in the interaction of the beliefs and intentions of a planning agent. This paper discusses the problem and proposes a solution.

Keywords Intention · Belief · Planning · Practical reasoning

1 Little Nell

The Little Nell Problem features a brief drama with two characters: Little Nell, who has been tied to the railroad tracks, and Dudley Doright, who has just learned of Nell's plight. A train is on the way, but Dudley believes he has enough time to rescue Nell, by first going to her location and then untying her. He forms an intention to do this. Dudley is *resolute*: he is confident that he will achieve things that he believes he can achieve and intends to achieve. So, in forming this new intention he also forms the belief that he will rescue Nell. But a moment later, as Dudley reviews his plans, he drops the intention, because now he believes that Nell will be rescued, and because there is no point in forming intentions to achieve things that you believe will happen. Nell is mashed.

This scenario was formulated (by a computer scientist) in Mcdermott (1982). Since then, it has been discussed by computer scientists interested in the formation and exe-

✉ Richmond H. Thomason
rthomaso@umich.edu

[1] Philosophy Department, University of Michigan, Ann Arbor, MI, USA

cution of plans, and several solutions have been proposed.[1] But (as far as I know) the problem is unknown to philosophers. In this paper I hope to show that it is philosophically interesting—not because it is especially difficult to think of solutions, but because it reveals unexpected complexities about the relation between belief and intention in practical reasoning.

2 Intention, action, and belief

From Bratman (1984), philosophers are aware of the distinction between *present-directed* (or *immediate*) and *future-directed* intentions. I myself would prefer to say that all intentions are future-directed, and involve a commitment (which may be more or less firm) to the performance of actions in pursuit of a goal,[2] and to speak of *intentional actions* rather than immediate intentions. In any case, the Little Nell scenario has to do only with future-directed intentions, which in this paper will simply be called "intentions."

Rational attitudes, including intention and belief, need to be *maintained*: to be updated, revised, and discarded. The Little Nell Problem has to do with the simultaneous rational maintenance of intention and belief. The problem arises from three characteristics of this interaction:

(1) If an agent intends to bring about some future goal state (at a certain time) then the agent believes that this state will happen (at that time).
(2) There are many rational grounds for abandoning intentions. An intention should be dropped (i) when its goal has been achieved, (ii) when the goal is deemed impossible to achieve, (iii) when a competing goal is rendered preferable because of a new opportunity, (iv) when new information removes the reason for desiring the goal, (v) when achieving the goal is deemed impossible without sacrificing some more important goal, and (vi) when a belief crucially supporting the intention is undermined. But also it should be dropped when (vii) it is deemed unnecessary to achieve the goal, because the agent learns that this goal will happen anyway.
(3) It is not entirely easy to formulate condition (vii) properly. Dudley went wrong by adopting a simple-minded version of this condition: *Where I is A's intention to achieve goal G, I is to be dropped if A believes that G will occur.*

Before motivating and illustrating these characteristics, we need to clarify the phenomenon of intention, and its relation to goals and plans. Intention begins with a future-directed desire, which can be represented as a proposition or state of the world.

[1] I know of two papers that propose solutions to the Little Nell Problem: Haas (1985) and Cohen and Levesque (1990).

[2] Of course, an intention can be based on multiple goals. But if we allow goals to be conjunctive, we can assume without loss of generality that intentions are based on a single goal. That is what I will do. Also, I will assume that goals are dated propositions—propositions to the effect that a state obtains at one or more specified times. When I say that "a goal will occur," this means only that its associated times are later than some reference time, which is specified by the context.

But it doesn't end there: many things we would like to happen don't result in goals or intentions.

Under the right conditions, a desire about the future can become a tentative goal and induce a process of planning. Planning (or means-end reasoning) produces action descriptions (more or less coarse, and temporally ordered) from premisses that consist on the one hand of tentative goals, and on the other of beliefs about the causal consequences of actions.

Planning is more like abduction than deduction. In general, there will be many hypotheses that explain a given datum, and abductive reasoning aims at a satisfactory explanation. And in general, there will be many plans to achieve a given tentative goal, and planning aims at a satisfactory plan. Just as any satisfactory explanation, together with appropriate beliefs, should entail an explanandum, any satisfactory plan, together with appropriate beliefs, should entail the goal.

The outcome of successful planning is a plan, but is not yet an intention. The concluded plan is like a recommendation; only when an agent accepts the recommendation and commits to the plan is there an active goal and an intention to achieve it. In many cases, the adoption of a satisfactory plan to obtain a desired goal will be automatic, but occasionally there may be undesirable side effects and costs associated with even the best plan to achieve a goal, and the agent may decide to abandon a tentative goal rather than to adopt a plan to achieve it.

An intention involves not only one or more goals, but a plan or way to achieve these goals. A desire, no matter how strong, to get in the Guinness Book of Records is not an intention until it's accompanied by a way to achieve this goal.[3]

We can now explain why an agent will believe that the goal of its intended plans will be achieved. (a) To have an intention is (*inter alia*) to believe that you will act on that intention. If you come to believe that you won't act to carry out a plan, you have also abandoned the intention. Of course, intentions can be weak and infirm—but so can beliefs. To the extent that someone's belief that she will act on an intention is weak, so is her intention. (b) Likewise, to intend to act on a rationally produced plan is to have confidence in the soundness of the plan. This is inherent in the planning process, which ensures that a plan will be grounded on beliefs about what the future will be like if the plan is executed. Together, (a) and (b) imply that there will be a belief that the goal will be achieved.

Confidence in our intentions, in the form of beliefs about the future, serves to inform further planning. If I intend to spend August in London, I need to have the belief that I will be in London. For instance, this belief supports a further plan to go to a play at the Globe Theater in August. And it will rule out a plan to go to the San Francisco opera that month.

The motivation for Characteristic (1) is evidently quite strong. Characteristic (2)—and in particular, Part (vii), with which we are concerned, can be supported by examples. Suppose, for instance, that Alex believes he will have no beer for the

[3] Not everyone accepts this condition on intentions, and in fact in planning theory it is useful to think of goals as degenerate partial plans. (Here, I call these things 'tentative goals'.) By analogy, it could seem sensible also to allow degenerate intentions. Readers who feel uncomfortable with this way of framing the process of intention formation can think of this as a terminological decision.

weekend and accordingly forms an intention to buy beer at the grocery. He then discovers beer in the back of his refrigerator, or perhaps a visiting friend tells him she will bring beer for the weekend. Certainly, it is then reasonable for Alex to drop his intention. To take a less routine example, the Allies' intention to invade the Japanese home islands was dropped as soon as the Japanese offered to surrender in 1945.

3 Solutions to the problem

Alex's intention was *gratuitous*, because his goal would be secured even if he did nothing to achieve it. The Little Nell Problem raises the question of how to characterize the conditions under which an intention is gratuitous. Several solutions to the problem have been proposed in the AI literature on practical reasoning.

In Mcdermott (1982), where the problem is introduced, McDermott says that the logic he uses in the paper (a nonmonotonic first-order temporal logic for reasoning about actions and plans) can't formulate a satisfactory solution, because "We need to express, 'Nell is going to be mashed *unless* I save her', and *unless* is a nontrivial concept." (Here, he cites philosophical work on conditionals, Goodman 1947; Lewis 1973.)

McDermott declines to formalize a solution, apparently because he could not see how to implement it. At the time this paper was written, AI researchers tended to assume the relation between the machinery of the logic and the reasoning performed by an implemented system would be more or less straightforward. Later work tends to think of a logical formalization as a specification of the reasoning, and makes room for a looser relation between logic and implementation.

Haas (1985) develops a planning formalism with multiple histories; this provides a modal semantics in the usual way for historical necessity.[4] Haas' main point is that it can be useful in planning to reason about possible events. He uses the Little Nell Problem to illustrate the idea, proposing that the reason for Dudley's intention to save Nell is not a belief that Nell will be mashed, but a belief that she *might* be mashed—a belief that in some possible future history she is mashed. Once Dudley forms his plan, according to Haas, he also forms the belief that Nell will not be mashed; but this does not undermine his reason for saving her, which is the belief that she *might* be mashed.

This doesn't seem right. For one thing, it doesn't account for legitimate cases of intention revision. It's reasonable for Alex to drop his intention to buy beer when he learns that his friend will bring beer, even though he thinks there's a (remote) possibility that she will not bring it.

And if future *possibilities* served as reasons for intentions, agents would have far too many reasons for actions. If a stock might go up and might go down, I would have a reason to buy it and, it seems, should form a plan to buy. But equally, I should form a plan to sell. Or suppose that the villain in the Little Nell scenario is going to flip a coin, and will tie Nell to the tracks if it comes up heads and throw her in the river if comes up tails. On Haas' view, Dudley would have to adopt contradictory goals.

[4] See Thomason (1984).

Perhaps the right way to approach this (and similar) cases is to say that Dudley has conflicting reasons for actions. But then, since he can't adopt conflicting goals, our account of how reasons for action induce goals has to be much more complicated than Haas seems to suppose, involving conflict resolution,[5] and we would have to distinguish *prima facie* reasons for action from all-things-considered reasons.

Cases of this sort tempt us to rethink the entire process of intention maintenance in terms of probabilities rather than beliefs. This alternative raises many difficulties—in particular, questions about how to reconcile degrees of belief with means-end reasoning—and here I won't do more than mention it, because these matters don't bear directly on the Little Nell Problem.

Cohen and Levesque (1990) presents an elaborate, thoroughly formalized theory of intention. The theory is impressive, and the clarity of the proposal and its success in accounting for many aspects of a challenging topic is more important than the few details with which I disagree.

Cohen and Levesque's solution to the Little Nell Problem, for the case where Nell can only be rescued once (and agents believe this) is simply that Dudley will not abandon his intention until the goal is fulfilled or deemed impossible—this follows from their idea that intentions involve *persistent goals*—goals that an agent will maintain until the agent believes they are achieved or have become impossible.

In fact, their agents can be overly resolute, clinging to intentions more than seems reasonable. Their central idea—that intentions tend to persist—may work well for agents which, like most robots, are intended to function as intelligent servants, and whose high level goals are given to them in the form of commands. Goal maintenance may not be so important for agents of this sort. But agents that are capable of forming their own goals can't afford to be so inflexible about these matters. So the core approach of Cohen and Levesque (1990) doesn't apply very plausibly to the cases enumerated in (2) (iii–vii), in which it would be reasonable to abandon an intention. In fact, they provide only two mechanisms for dropping an intention: achieving the goal, or coming to believe that the goal will never be achieved.[6]

4 The conditional solution

Let's return, then, to McDermott's idea that the rule for abandoning gratuitous intentions involves a conditional. In retrospect, it may have been a mistake for McDermott and other AI researchers to think that this idea, however plausible, would be unworkable. Lent and Thomason (2015) shows that, if a logic of time and action like the nonmonotonic Situation Calculus is adopted,[7] this provides a basis for interpreting conditionals whose antecedents involve the performance of (perhaps counterfactual) actions.

[5] For an account of how this works with reasons for belief, see Horty (2012).

[6] It might well be possible to modify Cohen and Levesque's theory to accommodate more of the mechanisms for intention revision listed in (2), but I believe that a thoroughgoing attempt to do this would have to result in a very different theory, because of the importance they attach to the persistence of intentions.

[7] See, especially, Turner (1999) and the references in this paper.

There are several ways that the conditional proposal might be formulated.

(4) Abandon an intention based on goal G when: you believe that if you were to do nothing, G would (nevertheless) occur.
(5) Abandon an intention based on goal G and involving plan π if you believe that if you didn't act on π, G would (nevertheless) occur.
(6) Abandon an intention based on goal G if you believe that if you didn't act on a plan to achieve G, G would (nevertheless) occur.

Formulation (4) is a common and natural way of describing the condition we want, but is not useful for theoretical purposes. It raises the question of whether it's actually possible to do nothing, and whether refraining is an action. And even if there is such a thing as genuine inaction, this is not what is wanted in the counterfactual condition we seek. Suppose only two actions are available to Dudley in the Nell scenario: rescuing Nell and cooking breakfast. Then the conditional that Dudley needs to think about is 'If I were to cook breakfast Nell would be mashed'—but cooking breakfast is not the same as doing nothing. Clearly, there is a covert quantifier (over relevant actions) in (4), and it is this we need to clarify.

Formulation (5) is too specific. Returning again to the Little Nell scenario, suppose Dudley has two equally good routes to the tracks where Nell is tied, and chooses to rescue Nell by taking the first route to her location and then untying her. If he were not to act on this plan, he would act on the alternative plan to rescue her by taking Route 2, and in this case too she would be rescued. But clearly this is not a reason to abandon the first plan.

Formulation (6) is better: it repairs the defect in (4) by explicitly quantifying over all plans to achieve the relevant goal, saying that the goal would be achieved even if the agent were to act on none of these plans. Pretty clearly, under these conditions it would be gratuitous to adopt any of these plans in order to achieve the goal.

Let's adopt (6), then, as a correct version of the counterfactual condition for abandoning an intention.

There is at least one remaining puzzle. Suppose Dudley has a colleague, Dick Daring. Dudley sincerely and convincingly promises Dick that he will rescue Nell if Dick doesn't. Dick sincerely and convincingly promises Dudley that he will rescue Nell if Dudley doesn't.

If both Dudley and Dick reason in accordance with (6), and take only each others' promises into account, then neither will rescue Nell, who will then be mashed. But they could be more sophisticated, and take one step of each others' reasoning into account, in which case they both will act to rescue Nell. The reasoning oscillates between joint action and joint inaction with each new layer of sophistication, with Dick and Dudley always acting similarly, because of symmetry. (This, of course, assumes that the two are unable to communicate after they make their promises.)

So our account of intention maintenance doesn't produce a determinate answer in this case. Since common sense doesn't seem to yield an answer either, this is not a problem for the account.[8]

[8] But, as a referee points out, it points in the direction of a different problem, having to do with commitment cycles in multiagent reasoning.

5 Feasibility considerations

There is no reason to believe that reasoning that would incorporate a counterfactual condition like Formulation (6) in intention maintenance would be particularly cumbersome and unworkable.

We are looking for an implementable approximation of (6) rather than a perfect rendition. This can be found by showing, for a representative set of plans that are available to the agent, that all of them, if executed, would actualize the goal.

Now, a restricted set of plans, representing the alternative courses of action that the agent might choose at a given time, should be available to any well designed planning agent. In fact, this agent will in general have already formed many intentions. Often, in fact, there will be only one plan in place for the relevant stretch of time. Dudley, for instance, just before he received the news about Nell, might have intended to make breakfast, eat breakfast, wash the dishes, and read a novel. In that case, this is what he would do if he didn't act to rescue Nell. To see whether an intention to rescue Nell is gratuitous, he only has to check whether Nell would be mashed if he were to pursue his standing plan for the morning. And Lent and Thomason (2015) shows that a process of contingent planning, no more complex than what is involved in ordinary planning, can establish that.

Another method of approximating (6) would resort to storing the *reasons* for beliefs about the future. An agent's beliefs about the future will depend on many things: assumptions about natural causal mechanisms, as well as about the actions that agents will take. Caching the reasons for beliefs is useful for many purposes, including belief revision and learning.[9] If this is done, an agent can test whether a goal is gratuitous by seeing whether the belief that the goal will be achieved depends on any beliefs about the performance of any of the agent's intended actions.[10]

I conclude that the doubts that McDermott and others have had about implementing a conditional-based approach to intention maintenance were misplaced. Implementing this approach may be difficult, but would not be much more, if any, difficult than implementing a well-performing autonomous planning agent.

6 Conclusion

I have no very sweeping conclusion to draw from this exercise, but I do wish to make two recommendations. (1) AI researchers do not need to be shy about accounts that involve conditionals. (2) Philosophers who are interested in practical reasoning may have much to learn from systematic attempts to think through the reasoning requirements for autonomous agents. These attempts can lead to new and interesting philosophical problems about intention and agency.

Acknowledgements I am grateful to two referees of this paper for helpful comments.

[9] See Kleer (1986) and Mitchell (1986).

[10] As a referee pointed out, a similar idea is incorporated in causal-link planners, a popular type of planning algorithm. See Barrett and Weld (1994).

References

Barrett, A., & Weld, D. S. (1994). Partial-order planning: Evaluating possible efficiency gains. *Artificial Intelligence*, *67*(1), 71–112.

Bratman, M. E. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375–405.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, *42*(3), 213–261.

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, *28*(1), 127–162.

Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy*, *44*, 113–118.

Haas, A. (1985). Possible events, actual events, and robots. *Computational Intelligence*, *1*(2), 59–70.

Horty, J. F. (2012). *Reasons as defaults*. Oxford: Oxford University Press.

Lent, J., & Thomason, R. H. (2015). Action models for conditionals. *Journal of Logic, Language, and Information*, *24*(2), 211–231.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Mcdermott, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science*, *6*, 101–155.

Mitchell, T. M. (1986). *Version space: An approach to concept learning*. Ph.D. dissertation, Computer Science Department, Stanford University, Stanford, CA.

Thomason, R. H. (1984). Combinations of tense and modality. In D. Gabbay & F. Günthner (Eds.), *Handbook of philosophical logic, Volume II: Extensions of classical logic* (pp. 135–165). Dordrecht: D. Reidel.

Turner, H. (1999). A logic of universal causation. *Artificial Intelligence*, *113*(1–2), 87–123.