Belief, Intention, and Practicality: Loosening up Agents and Their Propositional Attitudes[*]

Richmond H. Thomason
Philosophy Department
University of Michigan

rthomaso@umich.edu
http://www.eecs.umich.edu/~rthomaso/

December 4, 2010

Abstract

The beliefs of a single agent are typically treated in logic and philosophy as a single modality or epistemic attitude. I argue that it is better to treat belief as a family of loosely related modalities. This approach to belief, along with mechanisms for constructing modalities and for activating a modality that is appropriate for a specific reasoning situation, seems to provide a much better model of the relation of belief to intention in deliberative reasoning. I discuss this and other applications of this more flexible conception of belief and similar attitudes.

[*] Thanks for comments to Sarah Buss, Jason Konek, David Manley, Daniel Singer, Peter Railton,

1. Introduction: framing the problem

A lot has been written about the modularity of mind, but although I will be arguing for a modular account of the attitudes, I want to set aside what has been said about this topic in cognitive science and philosophy. In particular, I am not going to try to develop an account that could be fitted to the body of results obtained by cognitive psychologists. Nor do I want to articulate a formal theory of special-purpose reasoning and ground it in psychology by speculating about how it may correspond to a mental processing module.

Instead, I'm interested in general-purpose problem solving and its practical employment. This includes game-playing, planning, and calculation, as well as explicit, articulated reasoning about language, objects in space, and events in time. In fact, it includes just about anything that Allen Newell would put in the "intendedly rational band,"[1] or at what he calls "the knowledge level."[2] I believe that logic is the right tool for theorizing about the sort of reasoning that takes place at this level. As a first approximation, I take an analysis involving a first-order or even higher-order modal logic to be appropriate. Innovations from logical Artificial Intelligence, such as nonmonotonic consequence relations, may be useful additions to the logical framework.

For the purposes of this paper, I'm interested in practical reasoning: deliberation about what goals to pursue, how to pursue them, and the relation of goals and plans to what to do on a given occasion. I want a theoretical model that will engage this sort of reasoning—that, in particular, will provide a useful theoretical framework for practical reasoning, viewed at a fairly high level of abstraction.

The theory of programming languages and of program verification provides a good model for what I mean by "a fairly high level of abstraction." A programming language is a formalized medium for articulating complex imperatives for specialized agents. A semantics for the language makes it possible to either prove that the instructions are correct, showing that under certain conditions if the program is executed properly then the goals of the program will be satisfied. Often a failed attempt at such a proof will draw attention to a specific flaw in the program.[3]

This approach depends heavily on logical theory; but at the same time it engages the reasoning of an agent in significant and useful ways. At the same time, it abstracts away from many properties of the agent. It doesn't matter what operating system the computer uses, or how many processors the computer has, or (except for efficiency) to what extent parallelism is exploited in the program execution. It is irrelevant how the lower-level reasoning is implemented at the circuit level, after the program has been compiled.

In this work, we have a successful combination of logical sophistication,[4] a formal model of the reasoner, and an abstraction from many low-level details that still is able to engage significant and useful features of the reasoning, and that can deliver useful results such as correctness proofs.

[1] See [Newell, 1992, Chapter 7]. Newell's division of psychological theory into various levels is very helpful for methodological purposes.

[2] See [Newell, 1982].

[3] See, for instance, [Clarke *et al.*, 1999].

[4] Temporal logics are used in this area.

It is a bit of a stretch to extend this approach to cases that include human agents. With computers, we have specifications of correct behavior at the circuit level, and of the compiler that transforms a program into lower-level instructions. Despite the intense interest in recent years in neuropsychology, we have no such thing for human beings and (almost) no such thing for animals. Even so, I don't think that the extension is overly painful. Knowledge-level theories are appropriate and useful in robotics, and in this domain—even though we might have a specification of the agent as a computer running a particular operating system, the environments with which the agent has to deal are so rich and full of uncertainty that the importance of a specification at this level diminishes to little or nothing.[5]

When knowledge-level accounts of practical reasoning are used in robotics, they are usually incorporated in a "Belief-Intention-Desire" (or "BDI") model.[6] In the simplest case, we imagine that the agent has goals, in the form of a set of desired world-states.[7] The agent also has beliefs about the current world-state, as well as beliefs about the preconditions and immediate effects of actions in the agent's repertoire. Means-end reasoning then produces plans—sequences of actions that the agent believes will achieve the goal. One of these plans is selected and turned into an intention.[8] Intentions are then scheduled for execution, and normally the agent will then act on them.

Now, we can often reliably infer a (humanlike) agent's intentions, and can observe an agent's actions. The BDI model connects beliefs to intentions and actions. Even though there is some slippage in scheduling and execution, some uncertainty about desires, and alternative hypotheses about beliefs may be available to explain the observations, we do have evidential connections that can be pretty reliable. Developing a BDI-like architecture by filling in the details and elaborating the components could improve these connections.

In common sense BDI behavioral prediction one infers an intention (and hence, under the right circumstances, an action) from a belief in the presence of a supposed desire. For instance, if my wife is looking for the car keys and I tell her I left them on the kitchen counter, I'll expect her to look there. Conversely, in common sense behavioral explanation one infers a belief from an action, in the presence of a supposed intention. If I see my wife looking on the kitchen counter before going out to get in the car, I may infer that she believes that she doesn't have the car keys.

2. Methodological considerations

2.1. Applications and examples

Like other areas of contemporary philosophy, epistemology is too disengaged from challenging applications, and too driven by armchair examples, which often are far-fetched and unrealistic. This tendency is harmful in many ways. (1) Typically, realistic examples are more interesting and fruitful for philosophical purposes than contrived, unrealistic examples. (2)

[5] The literature in this area is extensive. See, however, [Doherty, 2004], [Nebel, 2002], [Reiter, 2001], and [Shanahan and Rundell, 2004],

[6] The original idea was proposed in [Bratman *et al.*, 1988]. See [Wooldridge, 2000] for a more extensive treatment of the topic, together with a formal language for reasoning about BDI agents.

[7] This account of desires is in fact much too simple, but it will do for our purposes here.

[8] Further desires, in the form of preferences for some plans over others, may play a role in the selection.

It is usually easier to produce systematic variations in realistic examples, providing evidence that can be connected with some confidence to generalizations, and eventually, to theories. (3) There is no reliable philosophical methodology for constructing purely imaginary examples, so philosophical inquiry that is driven by these examples tends to be capricious and unsystematic. (4) It is not as if we understand all the simple examples that are relevant to any area of philosophy, and so are forced to construct more complex cases in order to test our theories. It is easy to construct simple examples that challenge any area of philosophy. (5) Just as hard cases tend to make bad law, far-fetched examples tend to make bad philosophy, because we simply are not likely to have robust, reliable intuitions about bizarre examples.

The examples I use in this paper will, I hope, be realistic; in one instance (Section 4), I contrast a realistic example with a product of philosophical imagination.

2.2. Decision theory

Although I will be proposing an alternative to accounts of decision-making that use expected utility, what I will say is meant to be compatible with decision-theoretic approaches, as long as these are not applied generally, to all decisions whatsoever.

The model of decision-making developed in [Savage, 1972] requires an agent to bring a probability measure and a utility function to bear on every situation calling for a decision. The simplest way to achieve this would be to insist that the agent is equipped with an all-purpose probability measure and decision function, defined over a huge space including every hypothetical outcome with which the agent may have to deal.

This is clearly unworkable in many realistic deliberative situations. Even if we only require that in any decision situation the agent must be able to construct a probability measure and utility function that is appropriate for the situation, the probabilities and utilities are not always available. In fact, the ingredients we need for a decision-theoretic calculation can't be had except in cases with relatively few variables, with limited interactions between these variables, and with a relatively large amount of time for reflection. The use of computers has enlarged the cases where we can hope for such solutions, but even so such cases are relatively rare in practice.

That is why we need an alternative model of decision-making that appeals to reasons and reasoning, even if is not rational in the decision-theoretic sense. Such a model is also more faithful to the patterns we find in human decision-making.[9]

2.3. Intentions require beliefs

I will assume the following principle: ***intentions presuppose the appropriate beliefs.*** That is, there can be no intention without the appropriate beliefs. Suppose, for instance, that I approach a door that I closed an hour ago, leaving it unlocked, with the intention

[9] I am not saying that we should discard decision theory. It is fine in the cases where the deliberative situation can be modeled with global probabilities and utilities. But most deliberative situations simply can't be modeled this way. Perhaps some day we will learn how to combine decision theory with more flexible and qualitative forms of reasoning. That too, would be fine. But at the moment, we have to use many models of deliberation, if we want to be appropriately general.

to open it by simply turning the handle and pulling. Then I must believe that the door is unlocked, even if I don't have this belief explicitly in mind. If I didn't believe that the door was unlocked, I might well *try* to open it by turning the handle and pulling, hoping that it's unlocked. But under these circumstances, I can't intend to open it this way.

The idea that an action aiming at a desired outcome cannot take place without the belief that the outcome will be achieved is close to the principle that I just stated. But I do not accept this idea.

The difference between the two is best clarified using a decision that could be managed in two different ways by a deliberating agent: a probability-based style and a belief-based style. Suppose that I'm playing a game of five-card stud poker. The last card has been dealt. My four visible cards show a pair of jacks, and nothing better. In fact, all I have is a pair of jacks. One opponent is left in the game. The pot amounts to $500. Her hand shows a king, but no pairs. It is her turn to bet; she bets $250. Let's suppose that my choices are either to call her bet or to fold. If her down card is a king, she will win if I call her bet; otherwise, I win if I call her bet.

Case 1. I use decision theory. The utility is given by the amount of my stake in the outcome situation. Assuming I have a stake of $1000, the utility of folding is 1000. The utility of calling the bet if she doesn't have a pair of kings is 1750. The utility of calling the bet if she has a pair of kings is 750. Having counted the cards, I take the probability that she has a pair of kings to be .0571. The expected utility of folding, then, is 1000. The expected utility of calling the bet is $(.0571 \times 750) + (.9429 \times 1750)$, or about 1693. I maximize expected utility and call the bet.

Case 2. I have observed my opponent bluffing before. I know that the likelihood of her having a pair or kings is very low. Taking these to be reasons, I form the belief that she doesn't have a pair of kings. I call the bet, because according to my belief this will net me $750.

In Case 1, it would be wrong to say that I intend to win $750. I call the bet, hoping to win $750, and of course I'm trying to win $750, but I don't intend to win because the losing outcome is not ruled out by what I believe.

In Case 2, I do intend to win $750. I have the intention because I have formed the appropriate belief about what cards my opponent holds.

It is the same if we contrapose. Suppose that in Case 2 you take me aside and persuade me that my opponent might have a pair of kings. Having given up the belief that she doesn't have a pair of kings, I have to give up my intention to win by calling the bet. But the discarded intention may not prevent me from calling the bet. I can perfectly well say "Yes, she might have that pair; but I still intend to call her," elaborating by saying she probably is bluffing. In effect, I fall back on a qualitative version of Case 1. What I *can't* coherently say is "Yes, she might have that pair; but I still intend to win the pot." Again, we see that the intention requires the appropriate belief.

There may be cases where an agent acts on both sorts of deliberations, and cases where it is hard to tell whether an intention on a hope is in play. But there are also clear cases of both sorts of deliberate action. The distinction between acting with the intention to achieve a goal and acting in the hope that a desired outcome will be achieved is well grounded in common sense, and intuitions about the clear cases are very robust. In fact, the principle that intentions presuppose beliefs is, I think, entirely plausible.

This means, among other things, that situations that call for us to form intentions can act as inducements to provide the requirements for the intentions by acquiring appropriate beliefs. Suppose, for instance, that I have a standing goal not to overspend, and that my immediate problem is to decide whether to buy a new computer that I want. I need to effect this decision by forming an intention to buy the computer while not overspending or and intention to refrain from buying it.

In this situation, I need an appropriate belief. Suppose that there is just one pivotal issue: whether I can afford to buy the computer. Then somehow, I have to either form a belief that I can afford it or form a belief that I can't afford it. In resolving the issue, of course, I might gather information about my finances. But if this information doesn't suffice to produce a belief as to whether the computer is affordable, I can also adjust what counts as affordable in one direction or another. Otherwise, like an epistemic Buridan's Ass, I will be stuck.

An agent that must, in some cases, form intentions in order to make decisions, may find itself in situations in which a decision must be made, but the available information does not suffice to precipitate a belief. The need for mechanisms to deal with wich quandaries has consequences for how beliefs must function in the overall cognitive architecture, motivating a modular picture of an agent's beliefs. In Section 5, we will see how this plays out.

3. A proposal about belief

Later in this paper I'll present some logical theories. At the moment, I just want to present the general idea.

First, stipulate that we are concentrating on belief as a practical attitude: the "B" attitude of a BDI agent.

On a monolithic picture of practical belief, there is a single, ideally consistent general-purpose pro-attitude, "belief," that applies generally across the various practical situations that an agent faces in life. An agent has a single "belief base" that is applied to whatever decisions may come its way. Of course, the beliefs are updated—perhaps nonmonotonically—in light of experience. But on any single occasion when an agent is bringing beliefs to bear on several independent decisions, it will be drawing on the same, general-purpose attitude. And monolithic belief is dynamically inflexible: it can only be updated by rational revision in the light of new evidence. On some idealizations this update may be nonmonotonic, so beliefs could be lost as the result of observations. But the beliefs cannot change without new information.

On this view, an agent's beliefs are like the goods in a ready-made clothing store. There is a procedure for updating the inventory. Independently, a customer can go to the store and find clothing. The supplies of clothing are unrelated to the needs of the customer; they depend only on the state of the inventory.

I propose to think of the belief shop as more like a gentleman's tailor. The tailor keeps materials and tools for making clothing. A customer goes to the tailor, is measured, and orders custom-make clothing.

I want to say that appropriate beliefs for a particular practical purpose are manufactured for the occasion, and that the manufacturing process may involve reasoning. Instead of a single belief attitude, we have an open-ended and loosely organized family of belief-like attitudes. The family is open-ended because there are mechanisms for constructing these attitudes. And a new belief-like attitude may be constructed for a particular occasion.[10]

A constraint on the belief-producing reasoning that enforces joint consistency—in effect, requiring that all the beliefs that the reasoning produces should be part of a single consistent theory—would make the modular account of belief equivalent to the monolithic one. But (for reasons I'll get to) we do not want to impose such a constraint. The beliefs that are appropriate for one practical occasion may be inconsistent with those that are appropriate for another.

Let's assume the view of beliefs as modalities, characterized semantically by relations over possible worlds. Although it makes many idealizations,[11] this picture of epistemic attitudes has been successfully used in many applications having to do with reasoning about knowledge and belief.[12] This makes belief a modality. Belief is realized syntactically as an operator $\Box$ taking formulas into formulas, Where ϕ is a formula and a denotes an agent, $\texttt{[}a\texttt{]}\phi$ is a formula, expressing the proposition that the agent denoted by a believes the proposition expressed by ϕ. The usual interpretation of $\texttt{[}a\texttt{]}$ associates it with modal frames that are euclidean and serial; this corresponds to the axiomatization called **KD45** in [Fagin *et al.*, 1995].

The logical model that I'm recommending is not a great departure from this approach—but instead of equipping each agent a with a single belief operator $\texttt{[}a\texttt{]}$, I give an agent a family of belief operators $\texttt{[}a, i\texttt{]}$. As before, these operators take formulas into formulas. Where ϕ is a formula, $\texttt{[}a, i\texttt{]}\phi$ is a formula representing the proposition that epistemic module i of agent a believes the proposition that ϕ. As before, each operator is interpreted using a euclidean, serial relation. The resulting logic looks at first like a multiagent modal logic of the familiar sort, but in fact intra-agent modality is different in some important ways from inter-agent modality. In the inter-agent case, agents reason about one another's attitudes in much the same way that they reason about any other feature of their worlds. In the intra-agent case, modules of the same agent access one another in transactions that transmit information directly. We might expect such an important difference to affect the logic.[13]

The indices representing epistemic modules needn't be unstructured. In fact, it is convenient to think of them as bundles of features representing the provenance and status of the information contained by the associated module. In fact, the main thing I want to do in the rest of this paper is to consider some features that could be used to organize these information modules, and to suggest how they might be used in reasoning.

[10] Of course, there is such a thing as habit, and often the reasoning is minimal and routine.

[11] Logical omniscience is the hardest to swallow of of these idealizations.

[12] See [Fagin *et al.*, 1995].

[13] In unpublished work, I explore the use of a non-normal modal logic for distributed belief. That is because $\texttt{[}a, i\texttt{]}\texttt{[}a, j\texttt{]}\phi$ is peculiar if the contents of j are not accessible to i. The options are to treat it as false or as truthvalueless in this case; I choose the former option, which produces a logic like **S3**. But these details are not important for present purposes.

I will begin with a problem from the philosophical literature.

4. Kripke's Pierre puzzle

The Pierre puzzle is stated in [Kripke, 1979]. Briefly, Pierre grew up in France, where he learned about "Londre," always hearing charming things about Londre. "Londre est jolie," he says to himself, and continues to believe accordingly. Later he moves to an unpleasant part of London, learns English by immersion, and believes that London is not at all pleasant. He never realizes that Londre and London are the same city.

Although it doesn't constitute an entire solution, a modular account of belief seems to be a necessary condition for resolving this problem. (I am assuming it is out of the question to suppose that some beliefs are in English and some in French.) You can't begin to say anything very helpful about the puzzle unless you associate two belief modules with Pierre: one associated with his life in France and the other with his life in England. Certainly, there may be a lot of overlap between the two, but the overlap needn't be complete—in unusual cases, there may even be unresolved contradictions—and cases where Pierre's second language learning is imperfect may induce such discrepancies. When Pierre hears French, or speaks French, or even thinks to himself in French,[14] the life-in-France module is activated. When Pierre hears English, or speaks English, or even thinks to himself in English, the life-in-England module is activated.

As I said in Section 2.1, I believe that artificial examples tend to make for artificial philosophy. Whether or not you agree with me about this, I would also like to suggest that realistic, and if possible naturally occurring examples are more likely to be instructive than ones that are fantastic and contrived. Let me illustrate this point by contrasting the case of Drew McDermott's sink with the Pierre example.

I was told this as a true story, but haven't verified it. What makes it especially funny is the fact that McDermott is a computer scientist who at one time worked on planning.

> Once, McDermott's sink was so badly blocked that he had to remove the U-joint. He put a bucket under the sink, loosened the joint with a wrench, and the water in the sink ran into the bucket. Several minutes later he had to get the bucket out of the way, so he took it out and emptied it into the sink.

It seems plausible to me to say that, when he emptied the bucket, McDermott believed that the water would go on the floor, though perhaps the belief wasn't activated at the time.[15] But also, in a way, he must have believed at the same time that the water would go down the drain. According to the model of practical reasoning I subscribed to in Section 1, we can't explain his action of pouring the water into the sink without ascribing to him the belief that the water would go down the drain. Drew's probable reaction to this mishap is instructive; most likely he was startled. but not at all surprised. He was startled because he expected the water to go down the drain. He wasn't surprised because he knew it wouldn't.

[14] I only assume that some thinking is accompanied by subvocalization. I certainly do not assume that all thinking is associated with a language.

[15] I even think it's plausible even if emptying the bucket was automatic. Even automatic, habitual actions are intentional, and so have to be based on beliefs.

As with Pierre, we can come to grips with this example by supposing that belief is modular. To me, the realistic story is much more compelling, and I think it is likely to be more instructive. But more important for my present purposes—because it is more closely connected to reasoning than lapses of attention—is the interaction between belief and the appreciation of risk.

5. Risk

Consider a case where a probability measure and a utility function are not available in a situation calling for reasoned action. An agent that fell back on BDI reasoning in these cases would either be reckless or paralyzed if the beliefs weren't tailored to the occasion. If the standards for belief are overly relaxed for the decision situation, then hearsay evidence, as well as long chains of defeasible inference, could justify a belief. Then the monolithic agent would be reckless: eating a mushroom just because an inexpert friend has declared it to be safe, or passing on a hill because there was no oncoming traffic on the last several hills. Suppose, on the other hand, that the standards are stringent. Then the agent would be paralyzed: unable to eat a spinach salad because it might be contaminated, or unable to back up a car because someone might have just crawled behind it.

In fact, however, human beliefs are influenced by a sense of risk.[16] Without any change in the available evidence, a belief can disappear in the presence of risk, and can appear in the absence of risk.[17] Contrast the following two cases.

(5.1) Normally, when I park my car, I turn off the lights. I park my car downtown, near a service station, and leave it to do some errands. Ten minutes away from the car, it occurs to me that I don't remember turning off the lights.

(5.2) Normally, when I park my car, I turn off the lights. I park my car at a remote trailhead, 12 miles from the nearest highway, and set off on day hike. Ten minutes away from the car, it occurs to me that I don't remember turning off the lights.

The only significant difference between Example 5.1 and Example 5.2 is risk. In the first case, I can easily produce the belief that I turned the lights off, based on the (defeasible) reason that I usually turn them off. In the second case, I can't produce it. If I'm a worrying type, I may even be able to produce the belief that I didn't turn them off.

This mechanism of adjusting beliefs to risk would not be possible with monolithic belief—in the absence of new information, there would be no adjustment to be made. But if beliefs are *ad hoc*, and if one criterion for choosing the beliefs that are appropriate for a reasoning situation is a qualitative measure of the expected utility of acting on them, we can begin to explain how such adjustments can occur. In this example, we can assume that the only relevant proposition is whether the lights are off, so—if we simplify and think of a modality as a set of propositions—the issue is whether to believe a unit set of propositions, and the credibility of the set is equal to the credibility of the proposition that the lights are off. In

[16] For related work, see [Armendt, 2010]. Armendt is working in the framework of subjective probability, but the ideas are very similar.

[17] I described cases like this in [Thomason, 1987].

both cases, this credibiity is significant, but lower than the highest level. Say it is .8 on a scale of 0 to 1. In Example 5.2, the badness of the outcome of acting on the belief is high. Say it is -7 on a scale of -10 to $+10$. This attaches a risk factor of $.8 \times -7 = -.56$ to acting on the belief. This high risk may prevent the supposition that the lights are on from being used as a belief in this practical situation.

In many deliberative situations, and especially when the risk is high, or there is emotional involvement, or there is social pressure to have reasons for decisions, we seem to be condemned to form intentions backed up by reasons. And these reasons will have to function as beliefs in the deliberative situation.

I am supposing that most hikers in the situation that I describe would have to deliberate in belief-based mode. Suppose that the rational thing to do in this case, according to the decision-theoretic model, were to flip a coin and then proceed with the hike or turn back to check the car lights, depending on the outcome of the coin toss. I would expect that few hikers could muster the detachment required to adopt this protocol and proceed with the hike, supposing this to be the recommended action. Without a belief that the lights are off, worry would prevent the hiker from following through.

We have seen that an intention to hike and then drive home requires a belief that the lights are off. On the monolithic model of beliefs, there would be no mechanism for forming the appropriate belief. There is no way to get new information about the car lights without walking back to the car. So an intention to hike and then drive home would be impossible on this model. But in fact some hikers in this situation, condemned to belief-based deliberation, will decide to continue with the hike.

In this case, and in many similar cases that will occur to you, an agent can't act without the appropriate beliefs. An agent in this predicament, who is inclined to adopt a risky course of action, is in a state of *belief hunger*; to continue with the hike, the agent needs the appropriate belief. In this case, and in fact typically, reasons for adopting the belief are easy to find. The hiker has a habit of turning off the car lights; this is the norm. It is most likely that in fact the lights are on. These are the resources that we usually appeal to in forming defeasible beliefs, and they apply in this case. Of course, belief hunger and its satisfaction has its pathologies, including beliefs formed in the face of compelling evidence to the contrary, and self-deceptive beliefs. But I'm not interested in the pathology here; I do not want to say that the hiker who adopts a belief that the lights are off and proceeds with the hike is epistemologically defective, or that the process of forming the belief is in any way unreasonable, even if it is somewhat risky.[18]

I hope it's clear that I have nothing against deliberation that appeals to calculated expected utility. I certainly don't want to do away with this method of deciding what to do. Often, tradeoffs between the desirability of outcomes and the likelihood of achieving them need to be reconciled in practical decision-making, and these tradeoffs call for such

[18] One way to solve the problem of repeated decisions that according to game theory would best be solved by randomizing, but that require a reasoned decision, is to enhance a randomizing method with social or even religious approval. The ancient Greeks and Romans used the flight of birds and the entrails of animal sacrifices to make decisions, some at least of which match this description. Plains indians apparently used the motions of insects to decide where they would hunt.

calculations. How, then, does the picture I'm painting differ from, say, Leonard Savage's?[19]

Well, I don't take expected utility to be the whole story about how even an ideally rational agent reasons in practical situations; in fact, I think it doesn't fit the reasoning in most cases. I differ from Savage in not wanting to postulate global, all-purpose utility and probability functions, and in denying that probability functions make beliefs unnecessary in practical reasoning. I think that sometimes people act on intentions, and sometimes they act on hopeful expectations. Intentions require beliefs, and these beliefs can be manufactured *ad hoc* for the decision-making situation at hand. Furthermore, I am willing to allow qualitative and approximate methods for calculating utility.[20]

6. An application to cooperative reasoning[21]

A great deal has been written since the publication of [Stalnaker, 1972] about the dynamics of the common ground in a conversation.[22] Far less has been said about how the common ground is initialized. Almost nothing is said about how it can be initialized so as to promote modal mutuality.[23] The requirement of mutuality for the common ground is strongly motivated by theoretical considerations. It is also supported by linguistic evidence; see [Clark and Marshall, 1981, 415–420].

But this requirement raises a problem: how can we account for the reasoning that gives rise to a sense of mutuality? How, for instance, when we begin a conversation with someone on an airplane, can we find a common ground?

Clark and Marshall, as well as many other authors, speak of the attitude associated with the common ground of a conversation as if it were a matter of the mutual beliefs or even the mutual knowledge of the participants. But neither alternative works. The attitude can't be knowledge, because conversations can easily presuppose what is false. It can't be belief, because the rules of conversation don't require the participants to go away from the talk exchange believing whatever they have assumed for the sake of their conversation. Situations can arise in which you don't entirely trust what someone is saying, but don't want to be rude or to disrupt the conversation with objections.[24] In these cases you simply

[19] See [Savage, 1972].

[20] Many such methods are discussed in the Artificial Intelligence literature. In fact, there is an extensive literature on qualitative preferences, on calculating qualitative preferences over plans, and on integrating these preferences with planning algorithms. See [Baier and McIlraith, 2008] for a recent survey. Most of this work does not yet consider cases where uncertainty, risk, and the consequent need for expected utility is present; but see [Fargier and Sabbadin, 2005].

[21] There is some overlap between what I say in this section and the motivating parts of [Thomason, 2000] and [Thomason, 2002].

[22] Stalnaker uses the term 'presuppositions'. Here, I use Herbert Clark's term. See [Clark and Marshall, 1981].

[23] Many others use the term 'common', referring, for instance, to 'common knowledge'. I prefer 'mutual', because it is less likely to be confused with other group attitudes. For details about the logic of mutuality, see [Fagin *et al.*, 1995, Chapter 6].

[24] This could be a matter of genuine distrust, as in a conversation with an overeager salesman. But it can also happen in story-telling. When we hear an entertaining story that is presented as recalled history, we may not be sure which parts of it are fact, which are enhanced, and which are entirely fictional. Usually, it isn't important to sort this out.

suspend disbelief. In effect, this means that you create an *ad hoc* attitude of acceptance-for-the-sake-of-the conversation.[25] Some of the things accepted in this way might serve as beliefs for certain purposes. We might be quite confident that other such things are false. Notice that much the same thing can happen in reading a work of fiction. You can learn a lot about the Napoleanic Wars by reading Dostoyevsky, but of course you shouldn't believe that all the people, places, and events in *War and Peace* are historical.

This idea of an *ad hoc* conversational modality goes a long way towards explaining the mutuality of the common ground. We initialize this modality by tailoring it to our interlocutors—by putting things in it that we have good reason to suppose they will put in the modality they are constructing for us. We can provide a mechanism for constructing the modality by supposing that learning a proposition is more complicated than simply adding the information it contains to a basket full of beliefs. We need to tag what we have learned with background information. How did we learn it? Did we learn it under circumstances that we would expect to apply to other people? What sort of people would these be? If what we have learned is enriched in this way, constructing an *ad hoc* attitude might just be a matter of selecting propositions with certain features.

You can find an informal version of this proposal in [Clark and Schober, 1989]. They put the idea in terms of speech communities.

> The common ground between two people—here, Alan and Barbara—can be divided conceptually into two parts. Their *communal common ground* represents all the knowledge, beliefs, and assumptions they take to be universally held in the communities to which they mutually believe they both belong. Their *personal common ground* represents all the mutual knowledge, beliefs, and assumptions they have inferred from personal experience with each other.
>
> Alan and Barbara belong to many of the same cultural communities ...
>
> 1. *Language*: American English, Dutch, Japanese
> 2. *Nationality*: American, German, Australian
> 3. *Education*: University, high school, grade school
> 4. *Place of Residence*: San Francisco, Edinburgh, Amsterdam ...

There is more about the problem of mutuality in [Thomason, 2000, Thomason, 2002], and in fact a full solution to the problem has other ingredients. But treating the belief-like attitudes associated with conversations as flexible, *ad hoc* modalities is an important component.

7. Time and social pressure[26]

In situations calling for a reasoned decision, time pressure can enhance belief hunger. Social pressure can have a similar effect.

[25] Stalnaker makes this suggestion in [Stalnaker, 1975].

[26] There is some overlap here with [Thomason, 2007].

Philosophers of practical reasoning have paid attention to many sorts of practical pathologies, some of them invented. But little attention has been paid to dithering.

Consider a nervous driver at a stop sign at a busy intersection on a dark night. He needs to drive across the intersection. He looks left. A car zooms by from that direction. He looks right. It's clear. He looks left, it's clear. But wait—he can't see what's going on to the right, and doesn't believe it's clear anymore. So he looks right. He repeats the process until he realizes that he'll never get across this way. Time is pressing. But he can't move unless the road is clear. So he lowers his standards, saying to himself "If it was clear to the right a second ago it's clear now." And he hits the gas. Sometimes, of course, there may be no intention to cross the intersection, and no belief—just a sort of desperate hope. But I think that in this sort of case the need to act will sometimes induce a belief.

Jury duty can produce an enhanced and extreme case of pressure to believe. For responsible jurors, anyway, the duties call for a reasoned decision, and require certain beliefs about the facts of the case. In many cases, the risk factor (at least, the moral risk factor) can be high. But a holdout member of a jury can be under severe time pressure, as well as social pressure, to reach a decision. Most often, I suspect, this pressure induces a belief that might not otherwise have come into being.

8. The hypothetical dimension

Supposing blends into entertaining, entertaining blends into positing, positing blends into occasional belief, and occasional belief blends into entrenched, global belief. On the model that I'm advocating, there is no real need to draw a line at a particular place in order to separate genuine from pseudo beliefs. But the practicalizing mechanism that assembles an attitude for a decision-making situation would need to take these distinctions into account. To take two extremes, we would not want something supposed purely for the sake of argument to ever be practicalized. On the other hand, a supposition about what my name is should be generally and freely available in just about any reasoning situation.

In fact, if we don't draw a sharp line at any point in the continuum between supposing and believing, there is still no difficulty in preventing imagination and fiction from contaminating serious deliberation. The same mechanisms that apply when we gather information from external sources can and do apply when we assemble information from our own attitudes in order to construct an *ad hoc*, practical belief attitude. In Section 5, I included estimated credibility, adjusted for risk, among these factors. If things assumed for the sake of argument or for the sake or a conversation, or for following a work of fiction, were assigned credibility 0, this should suffice to keep them at bay in practical situations.

But in fact I don't think that credibility is the only disbelief-inducing mechanism. Temporary suppositions—assumptions for the sake of argument, or for contingency planning—can be forgotten once they have served their purpose. There is no point in maintaining a supposition that is of no future use. And somehow it doesn't seem plausible that we don't practicalize whatever we have understood in a conversation or read in a work of fiction simply because these things have low credibility. Perhaps assumptions can be labeled as impractical or hypothetical in various ways.

9. Activated belief and interactions between modules

Thinking of pro-attitudes as modular, as resources that can be marshaled and brought to bear on a particular problem, would allow agents to follow a more relaxed approach to storing and maintaining declarative information. This information could be stored in modules devoted to specific topics; these modules could be organized along taxonomic lines. Although consistency is always desirable, it would not be vital to check ensure global consistency across modules, as long as consistency is monitored when information is gathered from different modules for some specific purpose.[27]

This way of organizing things has turned out to be important in managing large-scale knowledge bases. If the task is actually more a matter of constructing a knowledge base than of acquiring large amounts discrete, unrelated information—that is, if the task is to decide how to formalize a topic and provide axioms and a reasoning mechanism—then it is very difficult to make progress on large scale repositories without modularizing the task. For a description of how this works out in the context of a specific knowledge representation project, see [Guha, 1991].

This approach, of course, will not work without procedures for collecting and organizing information from different modules. The architecture that suggested by this idea would involve more or less independent loci for representing, storing, and managing information, with general mechanisms for transferring the information. Some attention (but not enough) has been given to providing a logic for this sort of architecture.[28]

10. Minsky, the society of mind and the emotions

Although I think that something like the sort of epistemology I advocate here is pretty inevitable if you take seriously the idea that epistemology should have something to do with the sort of reasoning that is used in problem solving, I suspect that it will seem pretty radical to traditional epistemologists. I would like to mention briefly what Marvin Minsky has to say about the architecture of human thought, if only to differentiate what I am doing from his views, and to point out that modular epistemology can be far more radical.

Minsky's published work on this topic goes back to [Minsky, 1985], but the most recent and comprehensive statement of the ideas is [Minsky, 2006].

If societies can be said to reason, the reasoning would have to be distributed, involving separate modules that may communicate seldom or never. And societies can be anarchic, and to the extent that anyone can be said to be in charge, the leadership can change frequently, and change in ways that are disorderly. Minsky wants to transfer these features of societies to the mind.

Minsky has surprisingly little (surprisingly, because Minsky's background, after all, is in Artificial Intelligence) to say about how this idea would play out in terms of reasoning, and

[27] This purpose could be to produce the active beliefs to be directed at a specific problem. But modules may need from time to time to acquire information from one another for internal maintenance purposes.

[28] For many years, John McCarthy has stressed the need for a "logic of context" and made suggestions about what such a logic should look like. As far as I know, [McCarthy and Buvač, 1998] is the latest of his papers on the topic. Also see [Thomason, 2005].

especially in relation to problem solving. Although [Minsky, 2006] contains many suggestive comments that could be applied to reasoning, and especially to the interactions between emotions and reasoning, there is no systematic or detailed account of the reasoning mechanisms. However, it is clear that he imagines that humans can invoke a variety of reasoning styles, that these styles are related to cognitive resources that can be activated to a greater or lesser extent, that the emotions play a role in the activation, and that this process is more or less unruly.

The account of belief that I have proposed is committed to none of these things. In fact, it is confined to rational or at least reasonable epistemology. On the model that I have proposed, beliefs are in fact resources that can be activated and deactivated. But I think of this process as rule-governed, driven by the needs of a problem situation, and reasonable, even if not rational in a strictly decision-theoretic sense.

I don't doubt that there are interactions—in both directions—between human emotions and human beliefs and other truth-directed attitudes, and that here Minsky has many insights to offer, even if he seems to be unwilling to develop these insights. In fact, a weakness of BDI models is that they have little or nothing to say about the origin or maintenance of desires, and any account of this would have to take the emotions into consideration. But I don't think that such an account would need to be as unruly as Minsky seems to think.

11. Conclusions

Agents who reason in the way I have suggested we in fact reason will have a way to abuse the process of deliberation. But human beings seem to be such agents. Others have noticed this; some have taken a perverse sort of pride in it.[29] But we can admit that the mechanism is available to us, without suggesting that its abuse is a good thing. In fact, mature, thoughtful people will tend to avoid such abuse; this is part of what it is to be a mature, thoughtful person.

I don't doubt that for some combinations of deliberating agents and deliberative problems, belief-based intentions are not the best way of reasoning.[30] But belief-based planning and intention formation may well be a good general-purpose method for agents with human cognitive capacities. In any case, we seem to be stuck with this method—condemned to it, as I said, in many of the deliberative situations we have to deal with.

In this paper, I have tried to suggest what belief would need to be like for agents in this position. It turns out, if I'm right, that belief would have to be different from what many philosophers and decision scientists have imagined it to be.

[29] Ralph W. Emerson, with his "A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines," is an example. Emerson apparently is thinking of consistency in beliefs from one occasion to another, and feels that self-reliant men are above such things.

[30] Special-purpose computers and chess playing may be such a combination, for instance.

Bibliography

[Armendt, 2010] Brad Armendt. Stakes and beliefs. *Philosophical Studies*, 147(1):71–87, 2010.

[Baier and McIlraith, 2008] Jorge A. Baier and Sheila A. McIlraith. Planning with preferences. *The AI Magazine*, 29(4):25–36, 2008.

[Bratman *et al.*, 1988] Michael E. Bratman, David Israel, and Martha Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[Clark and Marshall, 1981] Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Arivind Joshi, Bonnie Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge, England, 1981.

[Clark and Schober, 1989] Herbert H. Clark and Michael Schober. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232, 1989.

[Clarke *et al.*, 1999] Edmund M. Clarke, Orna Grumberg, and Doron A. Peled. *Model Checking*. The MIT Press, Cambridge, Massachusetts, 1999.

[Doherty, 2004] Patrick Doherty. Advanced research with autonomous unmanned aerial vehicles. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR2004: Principles of Knowledge Representation and Reasoning*, pages 731–732. AAAI Press, Menlo Park, California, 2004.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge, Massachusetts, 1995.

[Fargier and Sabbadin, 2005] Hélène Fargier and Régis Sabbadin. Qualitative decision under uncertainty: Back to expected utility. *Artificial Intelligence*, 164(1–2):245–280, 2005.

[Guha, 1991] Ramanathan V. Guha. Contexts: a formalization and some applications. Technical Report STAN-CS-91-1399, Stanford Computer Science Department, Stanford, California, 1991.

[Kripke, 1979] Saul A. Kripke. A puzzle about belief. In Avishai Margalit, editor, *Meaning and Use: Papers Presented at the Second Jerusalem Philosophy Encounter*, pages 239–288. D. Reidel Publishing Co., Dordrecht, 1979.

[McCarthy and Buvač, 1998] John McCarthy and Saša Buvač. Formalizing context (expanded notes). In Atocha Aliseda, Rob van Glabbeek, and Dag Westerståhl, editors, *Computing Natural Language*, pages 13–50. CSLI Publications, Stanford, California, 1998.

[Minsky, 1985] Marvin Minsky. *The Society of Mind*. Simon and Schuster, New York, 1985.

[Minsky, 2006] Marvin Minsky. *The Emotion Machine*. Simon & Schuster, New York, 2006.

[Nebel, 2002] Bernhard Nebel. The philosophical soccer player. In Dieter Fensel, Fausto Giunchiglia, Deborah L. McGuinness, and Mary-Anne Williams, editors, *KR2002: Principles of Knowledge Representation and Reasoning*, page 631. Morgan Kaufmann, San Francisco, California, 2002.

[Newell, 1982] Allen Newell. The knowledge level. *Artificial Intelligence*, 18(1):82–127, 1982.

[Newell, 1992] Allen Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1992.

[Reiter, 2001] Raymond Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems.* The MIT Press, Cambridge, Massachusetts, 2001.

[Savage, 1972] Leonard Savage. *The Foundations of Statistics.* Dover, New York, 2 edition, 1972.

[Shanahan and Rundell, 2004] Murray Shanahan and David Rundell. A logic-based formulation of active visual perception. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR2004: Principles of Knowledge Representation and Reasoning*, pages 64–72. AAAI Press, Menlo Park, California, 2004.

[Stalnaker, 1972] Robert C. Stalnaker. Pragmatics. In Donald Davidson and Gilbert H. Harman, editors, *Semantics of Natural Language*, pages 380–397. D. Reidel Publishing Co., Dordrecht, 1972.

[Stalnaker, 1975] Robert C. Stalnaker. Pragmatic presuppositions. In Milton K. Munitz and Peter Unger, editors, *Semantics and Philosophy*, pages 197–213. Academic Press, New York, 1975.

[Thomason, 1987] Richmond H. Thomason. The multiplicity of belief and desire. In Michael P. Georgeff and Amy Lansky, editors, *Reasoning about Actions and Plans*, pages 341–360. Morgan Kaufmann, Los Altos, California, 1987.

[Thomason, 2000] Richmond H. Thomason. Modeling the beliefs of other agents. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 375–473. Kluwer Academic Publishers, Dordrecht, 2000.

[Thomason, 2002] Richmond H. Thomason. The beliefs of other agents. http://www.eecs.umich.edu/ rthomaso/documents/nmk/index.html, 2002.

[Thomason, 2005] Richmond H. Thomason. Making contextual intensional logic nonmonotonic. In Anind Dey, Boicho Kokinov, David Leake, and Roy Turner, editors, *Modeling and Using Context: 5th International and Interdisciplinary Conference*, pages 502–514. Springer-Verlag, Berlin, 2005.

[Thomason, 2007] Richmond H. Thomason. Three interactions between context and epistemic locutions. In Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure View, editors, *Modeling and Using Context: Sixth International and Interdisciplinary Conference, Context 2007*, pages 467–481, Berlin, 2007. Springer-Verlag.

[Wooldridge, 2000] Michael J. Wooldridge. *Reasoning about Rational Agents.* Cambridge University Press, Cambridge, England, 2000.