Some analyses of pro-attitudes

Robert van Rooy*
ILLC/University of Amsterdam
vanrooy@philo.uva.nl

Abstract

According to the pragmatic or functional conception of attitudes, we can say that John desires A iff John behaves such that he tends to bring it about that the actual world is an A-world, if his beliefs are true. This puts certain constraints on how to analyse desire attributions, but it leaves open a number of alternative analyses. Several alternatives will be discussed and compared in this paper. It will be suggested that for the *semantic* analysis of desire attributions it is useful to look at recent analyses of belief revision and theories of action and rational choice.

1 Introduction

According to the pragmatic or functional conception of attitudes, we can say that John desires A iff John behaves such that he tends to bring it about that the actual world is an A-world, if his beliefs are true. This puts certain constraints on how to analyse desire attributions, but it leaves open, I believe, a number of alternative analyses. Several alternatives will be discussed and compared in this paper.

2 A Hintikka-style analysis

According to the most straightforward way to account for desires, we can assume that just like there exists an primitive accessibility relation for *belief*, there also exists for our agent John a primitive accessibility relation for *desire*, Bul_j. Some have argued, however, that in distinction with belief, for desire this set of possible worlds should not be thought of as being primitive; it should rather be defined in terms of the propositions desired. Let's say that the propositions one desire might be mutually inconsistent, and that we follow Van Fraassen (1973)

*This paper was written as part of the 'Sources and Streams of Information' project, sponsored by the Dutch Organisation for Scientific Research (NWO), which is gratefully acknowledged.

and Kratzer (1981) determining an ordering relation on worlds by looking at the number of desirable propositions that worlds make true. Thus, let $G(j, w)$ be the set of propositions that John finds desirable in w. Then we say that u is at least as desirable as v with respect to $G(j, w)$, $u \leq_{G(j,w)} v$, iff $\{A \in G(j, w) |\ v \in A\} \subseteq \{A \in G(j, w) |\ u \in A\}$.[1,2] World u can now be said to be strictly desirable to v with respect to $G(j, w)$, $u <_{G(j,w)} v$, iff $u \leq_{G(j,w)} v$, but not $v \leq_{G(j,w)} u$. On the basis of this ordering relation, we can define a function, $Bul(j, w, X)$, that gives us the set of most desirable worlds in X with respect to the ordering relation determined by $G(j, w)$: $Bul(j, w, X) \stackrel{\text{def}}{=} \{w' \in X |\ \neg\exists w'' \in X :\ w'' <_{G(j,w)} w'\}$[3] On the basis of this function, we can now say that John desires A in w iff the set of most desirable worlds for John in w, $Bul(j, w, W)$, is a subset of A.

But this analysis gives immediately rise to a problem: it is predicted that desires are closed under logical implication, which does not seem to reflect the facts. As noted by a number of authors,[4] if John desires A, and B follows from A and is already believed by John, it doesn't have to be the case that he also desires B. If John hopes that his wife survived the accident, it doesn't follow that he hopes that his wife had the accident. According to Stalnaker (1984, pp 89-90), "the propositions one wants to be true (relative to a set of relevant possibilities) includes all the consequences of any proposition one wants to be true *which distinguish between the relevant alternatives*." What are the relevant alternatives to consider for the analysis of *desire* attributions? It is clear that to determine whether A is desired or not, we should look at a contextually given set that contains some A-worlds and some $\neg A$-worlds. Moreover, for the analysis of *want that* it seems that normally this contextually given set is the set of worlds compatible with what the agent *believes*.[5] As a result, we can interpret desire attributions of the form *John wants A* in the following way (where $K(j, w)$ represents the beliefs about past, present and future of John in w, and $[[A]](K(j, w))$ is the intersection of A with $K(j, w)$ if the presupposition of A is entailed by $K(j, w)$, and $\emptyset$ otherwise):

[1] In this way, $\leq_{G(j,w)}$ determines a partial ordering, but not a total one. Not all worlds have to be connected with each other.

[2] I will assume that capitals stand both for sentences and for the propositions that they express. I hope this never leads to confusion.

[3] You might wonder, should $Bul(j, w, X)$ be introspective? Yes, if desires are introspective. But desires are not introspective: My boss wants another cigarette, but he wished he didn't want that. Another introspection condition for desires seems reasonable to assume, however; if $K(j, w)$ is the belief state of John in w, it should not only be the case that for all $v \in K(j, w) : K(j, v) = K(j, w)$, but also that $G(j, v) = G(j, w)$.

[4] For instance, Stalnaker (1984), and Heim (1992).

[5] Normally, because (i) in some *want* attributions the context of interpretation for the embedded clause needs to be a *superset* of the belief state, as for Heim's (1992) example *(John hired a baby-sitter because) he wants to go to the movie tonight*, and (ii) sometimes the context of interpretation should be a *subset* of the belief state, as for desire attributions conditionally dependent on other desire attributions: *John's father hopes that his son never smoked before, and hopes that he just started smoking.*

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, K(j, w)) \subseteq [[A]](K(j, w)),$$

and presuppose that A is true in some but not all worlds of $K(j, w)$.

3 Desire as *ceteris paribus* preference

According to the above analysis of desire attributions, we have assumed that desires are closed under implication. To account for the obvious problem that such an analysis gives rise to, we have assumed that we should only look at implication with respect to the relevant alternatives. Another problem cannot be solved in this way, however.

I have argued above that desires might be mutually inconsistent. But our analysis does not predict that if A and B are mutually inconsistent, one can both desire A and B. The reason is that it is predicted that conjunction introduction is valid. The fact that it is possible that John wants to be with his wife, and that he wants to be with his mistress, although for obvious reasons he doesn't want to be with both,[6] suggests that the set of desires that one has need not be consistent, and thus that for the analysis of desire attributions we should not only look at the most desirable worlds consistent with what is believed, but rather base the analysis of desire attributions more directly on the preference order

Indeed, this is what Heim (1992) argued for. She proposes that an attribution like *John wants A* is true iff John *prefers A* above $\neg A$. In this way, she gets rid of the closure condition for rational desires. The simplest possible analysis of this form would demand that preferring A above $\neg A$ means that *all* A-worlds consistent with what one believes, are better than all $\neg A$-belief worlds. This would give rise to a very strong notion of desire. To weaken it,[7] Heim assumes a *ceteris paribus* analysis of preference: A is preferred to B, if for every situation compatible with what is believed, its closest world in which A but not B is true is preferred to its most similar world where B but not A is true.[8] If we assume that f is a similarity function known from the Lewis/Stalnaker analysis of counterfactuals,[9] and that in w John prefers proposition X to proposition $Y, X \leq_{j,w} Y$, iff $\forall w' \in X : \ \forall w'' \in Y : \ w' \leq_{G(j,w)} w''$ & $(Y = \emptyset \Rightarrow X \leq Y)$, Heim's interpretation rule for *want that* goes as follows:

[6] Some might add *at the same time.*

[7] Other ways to weaken this are to say that X is preferred to Y iff (i) *some* X-world is better than *all* Y-worlds, or (ii) *each* X-worlds is better than *some* Y-worlds. But both options are well known to be unplausible.

[8] For a defence of this *ceteris paribus* analysis of preference, see Von Wright (1963) and especially Hansson (1989).

[9] Such a selection function is a function in $[(W \times \wp(W)) \rightarrow \wp(W)]$ and satisfies the following conditions: (i) $f_w(A) \subseteq A$, (ii) $f_w(A) = \{w\}$, if $w \in A$, and (iii) if $f_w(A) \subseteq B$ and $f_w(B) \subseteq A$, then $f_w(A) = f_w(B)$.

$$[[Want(j, A)]]^w = 1 \quad \text{iff} \quad \forall w' \in K(j, w): \\ f_{w'}([[A]](K(j, w))) \leq_{j,w} f_{w'}([[\neg A]](K(j, w)))$$

Note that according to this interpretation rule not only the most preferable worlds in a set count, and that rational desires are not predicted to be closed under logical implication.

Although the analysis of preference implicitly used by Heim (1992) verifies the principle that if A is at least as preferable to B, A is also at least as preferable to $A \vee B$, which in turn is at least as preferable to B, it still doesn't verify the stronger principle that says that if A is strictly preferred to B, and A and B are both compatible with what is believed, A is also strictly preferred to $A \vee B$, which in turn is also strictly preferred to B. This principle comes out valid if we have a logic that gives $A \vee B$ a preference value somewhere *in between* the preference values of A and B. That this is needed is suggested by the following example due to Rescher (1967):

Suppose we have four relevant worlds, $\{w_1, w_2, w_3, w_4\}$, where the propositions A and B differ in truth-value such that A is true in w_1 and w_2 and false in the other worlds, while the opposite is true for B. Suppose now that the ordering relation between possible worlds is such that w_1 is strongly preferred to w_4 which is just a bit better than w_2, which in turn is strongly preferred to w_3. Suppose now that except for A and B, w_1 is closest to w_3, and w_2 closest to w_4. In this situation, the *ceteris paribus* preference analysis would predict that A is not preferred to B and so that A is not wanted, which seems counterintuitive.

For instance, let us consider the preference ordering of a German general who wants to know whether he should attack France via Belgium, A, or directly via the German-French border, B. The worlds w_1 and w_3 are very close to each other because in those worlds the French only expect a German attack directly via the German-French border. In worlds w_2 and w_4, on the other hand, the French are well prepared for a German attack both via Belgium and via the direct border. If A is true in w_1 and w_2, and B in w_3 and w_4, clearly w_1 is strongly preferred above w_2, and w_4 is strongly preferred above w_3. Obviously, w_1 is strongly preferred to w_3: w_1 means victory and w_3 means defeat, because it is assumed that the French army is equally good as the German army. It also seems reasonable to assume that if the French are prepared for an attack at both places, it is better to attack directly via the German-French border, because of limiting transport problems. So, w_4 looks a bit better to the German general than w_2. But although there is a B-world, w_4, that is strictly preferred to an A-world, w_2, the German general is advised to attack the French via Belgium, and has the chance of an easy victory in battle. But according to the *ceteris paribus* analysis of preference, we should not advice the general to go via Belgium.

How can we get rid of this problem? The answer is simple: by using a more fine-grained preference logic. The most suitable logic for our purposes seems to be (a variant of) Jeffrey's (1965) preference theory, to which I will turn now.

4 Desire as quantitative preference

Nice from our point of view is that Jeffrey's theory of preference, in distinction to some other quantitative preference logics, is compatible with the Boolean analysis of the connectives common in semantics. Let us assume that $P_{j,w}$ is the probability function that assigns to each world its probability according to j in w, and $d_{j,w}$ a function which assigns to each possible world a number in $\mathbf{R}$, measuring its desirability according to j in w. The *probability* that j assigns to A in w, $P_{j,w}(A)$, is simply the sum of the probabilities of the cases (worlds) in which it is true, $P_{j,w}(A) = \sum_{v \in A} P_{j,w}(v)$. The desirability of a proposition A for j in w, $d_{j,w}(A)$, is a weighted average of the desirabilities of the worlds in which it is true, where the weights are proportional to the probabilities of the worlds,

$$d_{j,w}(A) \quad = \quad \frac{1}{P_{j,w}(A)} \times \sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v).^{10}$$

Given Jeffrey's preference theory, the simplest idea would be to say that the desire attribution *John desires that A* is true if the desirability for John of the embedded clause is larger than the desirability of a tautology:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad d_{j,w}(A) > d_{j,w}(\top)$$

It is easily seen that this doesn't predict desires to be closed under logical consequence, that it doesn't make conjunction introduction valid anymore, that it predicts that if $Desire(j, A)$ is true, and John prefers B to A, also $Desire(j, B)$ is true, and that it can account for Rescher's problem.[11] In distinction to the analysis of bulletic predicates by Heim, it doesn't make use of the *ceteris paribus* condition, but in this case it's not needed to get a very weak system.

Let's consider our model again with four worlds, where w_1 and w_3 are most similar to each other, and the same holds for w_2 and w_4. Let us also assume that $A = \{w_1, w_2\}$, and $B = \neg A = \{w_3, w_4\}$, and that all four worlds are equally likely true. In that case, the *ceteris paribus* analysis of preference demands that for A to be desired, both w_1 must be preferred to w_3, and that w_2 must preferred to w_4. Jeffrey's preference theory, on the other hand, only demands that if we can give a cardinal valuation to the four worlds, that the average valuation of w_1 and w_2 is higher than the average valuation of w_3 and w_4. As this example illustrates, the quantitative approach *weakens* Heim's qualitative approach. In the quantitative approach, we don't compare possible worlds that are most similar to each other, but instead we compare whole information states.

[10] The given formulae are for simplicity based on the assumption that there are only finitely many possible worlds.

[11] Rescher's (1967) logic of preference can also handle those problems, but that is no big surprise; Rescher's logic is only a special case of Jeffrey's system in that all possible worlds have equal probability.

I am not sure whether the weakening is in general preferred to Heim's strong notion of preference, but as the above discussed example of the German general illustrates, it seems to be preferred in at least some cases.

5 A conditional analysis of desires

Until now we have discussed three kinds of analyses of desire attributions. The first was based on a classical all-or-nothing analysis of preference, the second was based on a *ceteris paribus* analysis of preference, and the third on a quantitative notion of preference. In this section I will discuss yet another analysis of preference.

Asher (1987) observed that desire attributions normally obey disjunction elimination, and it is also easily seen that indefinites in the scope of verbs of desire are normally interpreted "arbitrarily". Thus, we can normally infer (1b) from (1a), and (2a) is normally interpreted as something like (2b):

(1) a. Alexis hopes that she will have chicken or fish for dinner.

b. So she hopes that she will have chicken for dinner.

(2) a. John wants to catch a fish.

b. John wants to catch an arbitrary fish, any fish will do.

These facts are surprising for any of the above proposals. They can, however, be accounted for if we assume that desire attributions should be understood as implicit conditionals. Thus, if *John wants that A* means something like "If A is the case, John will be satisfied". Disjunction elimination now follows immediately, but unfortunately, also the more general *downward entailment* is predicted to be valid. It is predicted that if John wants A, and B entails A, it follows that John wants B, too. But this is obviously a wrong prediction: I want to have a holiday this summer, but do not want a holiday and bad weather. Still, the conditional interpretation of desire attributions can be rescued, if this conditional is not treated as an indicative conditional, but as a *subjunctive* conditional instead, that is, in terms of belief *revision*.

In the simplest variants of these belief revision frameworks (Harper (1975, 1976), Gärdenfors (1988)), an acceptance state is modelled by a set of possible worlds, K, and a selection, or belief revision function $*$. If we say that $\langle K, * \rangle$ is a belief state, and A any proposition, then K^*_A is called the revision of K by A, and this revision process is constrained by the following rules for minimal belief change:

(K^*1) For any proposition A, $K^*_A \subseteq A$

(K^*2) If $A \neq \emptyset$, then $K^*_A \neq \emptyset$

(K^*3) If $K \cap A \neq \emptyset$, then $K^*_A = K \cap A$

(K^*4) If $K^*_A \cap B \neq \emptyset$, then $K^*_{A \wedge B} = K^*_A \cap B$

Harper (1976) showed that instead of using a belief revision function, epistemic revision could be based on an ordering relation, $\preceq$, of possible worlds, too. It is quite easy to see what condition this ordering relation has to satisfy to implement the same belief revision policy as $*$ does: $v \preceq w$ iff $v \in K^*_{\{v,w\}}$. We can now check that this ordering relation is reflexive, transitive and connected. It can also be shown that if we take such an ordering relation $\preceq$ as primitive, we can define both K^*_A and K as follows: $K^*_A \stackrel{\text{def}}{=} \{w \in A |\ \forall v \in A : w \preceq v\}$ and $K \stackrel{\text{def}}{=} K^*_\top$, such that K^*_A satisfies $(K^*1) - (K^*4)$.

Obviously, when a belief state is represented by an ordering relation, or by a set of worlds *plus* a change function, such a belief state contains more information than a state just represented by the set of worlds alone. I will call such a belief state an *extended* belief state.

To analyse desire attributions in terms of revision, we can assume that $K(j, w)$ represents no longer the set of futures consistent with what John believes in w, but the possible ways the world might be *at this moment* according to John in w.[12] Thus, if we want to look at the future, we have to use already the more general revision rule. I will assume that if somebody wants A, he has a desire about the future and so does not believe it yet. Desire attributions can now be analysed in terms of revision as follows:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad K(j, w)^*_A \subseteq Bul(j, w, W)^{13}$$

Thus, John wants A in w is true iff $K(j, w)$ revised by A is a subset of the set of John's absolute favourites among the (what he considers to be) possible futures. Note that according to the above rule, neither upward entailment, nor downward entailment is valid. Moreover, disjunction elimination is allowed, but only if the complements of both disjuncts are equally strongly entrenched.[14] This seems exactly what we need. Normally disjunction elimination is valid, and normally indefinites get the arbitrary interpretation, but this is not always the case:

(3) John wants a beer, but not a warm one.

6 Buletic ordering

Still, a counterexample like (3) to the arbitrarily interpretation of the indefinite has intuitively nothing to do with *epistemic* entrenchment. This suggests that

[12] See section 8 for more on this.

[13] where $Bul(j, w, W)$ is defined as in section 2. The form of this interpretation rule was actually proposed by Price (1989) in his defence of the Desire-as-Belief thesis.

[14] If the revision function $*$ obeys $(K^*1) - (K^*4)$, $K^*_{A \vee B} = K^*_A \cup K^*_B$, only if $\neg A$ and $\neg B$ are equally strong entrenched in K.

the ordering relation by which we determine the relevant change function should not be induced by epistemic entrenchment, but by *desirability* instead. What we could do is to demand that *the best* A-worlds are among the most desirable belief-worlds. This suggests that we should use the following interpretation rule:

$$[[Desire(j, A)]]^w = 1 \quad \text{iff} \quad Bul(j, w, [[A]](K(j, w)) \subseteq Bul(j, w, K(j, w))$$

This interpretation rule has a number of desirable consequences. First, it predicts that disjunction elimination and the arbitrarily interpretation of indefinites used in desire attributions are not valid according to the above interpretation rule. From *John wants that A or B*, I can only conclude that John also wants A, if A is at least as desirable for John as B. Similarly, from *John wants an apple*, I can only conclude that John wants a green apple, if eating green apples is at least as desirable for John as eating apples of any other colour. And this is confirmed by (3). Second, in distinction with the conditional analysis of the previous section, it doesn't have to make use of revision in order not to predict that if B entails A, desiring A does not entail desiring B. This is due to the fact that we only look at the best A-worlds compatible with what is believed. Third, it can account for the fact why sequences like (4) are out:

(4) John wants a cool beer, but he doesn't want a beer.

The reason is, according to this approach, that desires are closed under logical implication. It can easily be checked that according to the above interpretation rule the sentence *John wants* A is true in w for any A compatible with what is believed, $[[A]](K(j, w)) \cap Bul(j, w, K(j, w)) \neq \emptyset$. It then immediately follows that if $A \subseteq B$, also $[[B]](K(j, w)) \cap Bul(j, w, K(j, w)) \neq \emptyset$, and thus John wants B too.

7 Combining belief revision and desirability

According to the above analysis, all counterexamples to disjunction elimination are due to the fact that some disjuncts are *strictly preferred* to other disjuncts. In many cases this seems indeed the reason behind such counterexamples, but I don't believe it is the reason behind all of them.[15] Consider (1a)-(1b) again, repeated as (5a)-(5b):

(5) a. Alexis hopes that she will have chicken or fish for dinner.

b. So she hopes that she will have chicken for dinner.

Consider now the case where Alexis thinks that there is a tiny chance of getting chicken, A, and a good chance of getting fish, B. She prefers both to anything

[15] I am indepted to Ede Zimmermann (personal communication) for this.

else she considers possible, but has no preference for the one above the other, i.e. in w it holds that $A \approx_{a,w} B$. The above analysis, just like the quantitative analysis discussed earlier, would then predict that disjunction elimination is allowed. However, it seems that in such circumstances it is okay to assert (5a), but not to assert (5b).

Perhaps the most obvious way to account for this is by making use of the quantitative framework. By using Jeffrey's theory of preference, we might say that instead of looking at the *desirability* of a proposition, we rather should look at its *expected value*. Where the desirability of a proposition is the *weighted* average of the desirabilities of the worlds in which it is true, and thus does not increase in case the probability of the proposition increases, the expected value of a proposition gets higher in case the probability increases. The expected value of A for John in w, $EV_{j,w}(A)$ is defined as follows:

$$EV_{j,w}(A) \quad = \quad \sum_{v \in A} P_{j,w}(v) \times d_{j,w}(v).$$

Then we might say that each desire attribution is interpreted w.r.t. a set of alternatives, C, and that John desires A, if the expected value of A is at least as high as the expected value of any of its alternatives:

$$[[Desire_C(j, A)]]^w = 1 \quad \text{iff} \quad \forall B \in C: \ EV_{j,w}(A) \geq EV_{j,w}(B)$$

Another way to account for our above problem within a quantitative framework is to make use of *revision* of probability functions. We can do this by making use of so-called *Popper functions*, also known as *extended probability functions*.[16] Popper functions are probability functions that take conditional probabilities as basic. In contrast to standard probability functions, for a Popper function Pr, $Pr(A/B)$ is also defined if $Pr(B) = 0$. As a result, a Popper function contains the extra information what would happen under revision. Harper (1975) showed that if we limit ourselves to probability 1, the minimal revision modelled by Popper functions satisfies $(K^*1) - (K^*4)$. Let us now say that $Pr_{j,w}(v/A)$ gives us the probability John assigns to v in w under the revision of A. In that case we can define the *desirability* of A, $d_{j,w}(A)$, with respect to probability function $Pr_{j,w}$ and desirability function $d_{j,w}$ as follows:

$$d_{j,w}(A) \quad = \quad \sum_{v} Pr_{j,w}(v/A) \times d_{j,w}(v).$$

Observe that this is similar to the *expected value* of A. Now that we have made use of revision, we can say that you desire A if the expected value of A is greater than the expected value of doing nothing, $d_{j,w}(\top)$.[17]

[16] See Stalnaker (1970), and Harper (1975).

[17] It should be obvious that the last solution to Zimmermann's problem also has its qualitative variants. I leave those to the reader's imagination, however.

8 Intention as stable desirable action

Until now I have assumed that all verbs of desire should be analysed in the same way. Thus that the emotive cognitive attitude *hope* should be analysed in the same way as a pro-attitude like *intend*. Intuitively, however, there are at least two differences between *intend* and *hope*: (i) whereas what you intend has typically something to do with your own activities, hopes are not so closely related with actions of the agent himself, and (ii) whereas *intend* is necessary future oriented, *hope* need not be, as in *I hope he survived the operation*. For *intend* it is normal to take as complement *to*-infinitives that seem to designate abilities, but the verb *hope* takes also that-clauses as complements. I don't want to suggest that the two verbs should be analysed in a completely different way, but it might be the case that we use two different concepts of desire, one concept about futures that the agent can influence himself, and one that is about circumstances he cannot influence. Moreover, that these two concepts are typically expressed by the words *intend* and *hope*, respectively.

One option to 'explain' this all is to say that the truth conditions for these constructions are identical and should be analysed as before, but that *appropriateness* conditions for asserting such sentences differ. For *hope* it should be the case that both the embedded clause and its negation should be consistent with what is believed about the present by the agent, but this need not be the case for *intend*.

Perhaps the intuitive difference between the two concepts can be accounted for in this way, but maybe we should take the notion of *action* more seriously than we have been doing until now. This can be done by following the lead of proponents of *causal decision theory* by analysing actions in terms of *imaging*.

8.1 Imaging

The Bayesian account of probability is purely epistemic in nature. So $P(C/A) > P(C)$ means that A is *evidentially* relevant for the acceptance of C. But some puzzles in Jeffrey's (1965) purely evidential decision theory make clear that evidential relevance should not be confused with *causal relevance*. According to Jeffrey's decision theory, actions are evaluated according to the probability the deliberator assigns to the desired state conditional on the proposition expressed by the action. The conditional probability $P(C/A)$ models the evidential relation the agent sees between A and C; if $P(C/A)$ is high, the agent would assign a high probability to C, if he would learn the news that A is the case. Obviously, if A causes $C, P(C/A)$ would be high, but the problem is that $P(C/A)$ might also be high in cases where A does not cause C, but where both are caused by a common cause. Suppose, for instance, that the correlation between smoking and long cancer was not due to the consequences of smoking through the lungs, but due to a common genetic factor that causes both the tendency to smoke and the tendency to develop lung cancer. In that case there is no reason for agents to

withdraw smoking in order to prevent lung cancer, although the probability of getting a lung cancer conditional on smoking is high. Proponents of *causal decision theory*, like Gibbard & Harper (1978), have concluded that causal relevance, the kind of relevance needed to evaluate one's actions in a deliberation, should not be modelled by the conditional probabilities of consequences with respect to actions, but rather by the probabilities of their counterfactuals expressed.

With the distinction between evidential and causal decision theory, there corresponds a distinction between two ways of changing one's belief state. Conditionalisation, or epistemic revision satisfying $(K^*1) - (K^*4)$, is supposed to mirror the way a rational agent would change his belief state if he would learn new information, while *imaging* is supposed to mirror the way a rational agent would change his belief state if he, or somebody else, would do a certain *action*. Imaging is a function of minimal belief change which uses not primarily the information available in the information state ordered by epistemic entrenchment, but the similarity relation between *individual* possible worlds.[18] Let f be a Stalnaker selection function mirroring this similarity relation, then probability function P_A, the image of P by A, can be defined as follows:[19]

$$P_A(w') \quad = \quad \sum_w P(w) \times \begin{cases} 1, \text{ if } f_w(A) = \{w'\}, \\ 0, \text{ otherwise} \end{cases}$$

Imaging can obviously also be defined in a qualitative framework; if K is a belief state, and f a selection function, the new belief state will simply be the image of K under $f(A)$, where $f(A)$ is a function from worlds to their most similar A-worlds:

$$C_K(A) \quad = \quad \bigcup \{f_w(A) : \ w \in K\}$$

The main difference between imaging and epistemic revision is that belief change by imaging need not be *preservative*, while belief change by epistemic revision is. That is, whereas it will hold that for any K and A such that $K \cap A \neq \emptyset : K^*_A \subseteq K$, it need not hold that $C_K(A) \subseteq K$.

Although imaging need not be preservative, if we want to use it to analyse actions it should satisfy another constraint; if we go from w to w' by imaging w.r.t. A, w' must be a possible *future* of w. To account for this formally, we can use Thomason's (1970) framework of *branching time*.

Let $\langle M, < \rangle$ be a modal structure, where M is a nonempty set of *moments* and $<$ a relation on M. The relation should not only be transitive, asymmetric and irreflexive, but should also reflect the intuition that the *past*, in contrast to the future, is *settled*. The constraint that accounts for this says that for any

[18] In Katsuno & Mendelzon (1991) the qualitative version of conditionalisation is called the *revision* of a belief state, and by the *update* of a belief state is meant the qualitative version of imaging.

[19] See Lewis, 1975. Although in this rule I make use of Stalnaker's uniqueness assumption, it is easy to generalise imaging by giving up this constraint.

$n, n', m \in M$: if $n < m$ and $n' < m$, then $n < n'$ or $n' < n$ or $n = n'$. In terms of $<$, we can define $\leq$ in the usual way: $n \leq m$ iff $n < m$ or $n = m$. A *history* h on this model structure is a maximal chain through this structure. If H_m denotes the set of all histories through m, we might say that H_m represents the set of scenarios open at m. A world is everything that is the case, and should therefore be represented by a moment-history pair. We will say that $\langle h, n\rangle \leq \langle h', m\rangle$ iff $n \leq m$.

Let us now represent John's belief state at world w, $K(j, w)$, by a set of worlds, or moment-history pairs. From $K(j, w)$ we can in our framework of branching time also define the set of possible future worlds that our agent considers possible, $K^*(j, w) \stackrel{\text{def}}{=} \{\langle h', n\rangle : \exists\langle h, m\rangle \in K(j, w)\ \&\ \langle h, m\rangle \leq \langle h', m\rangle\}$.

To account for the intuition that an action done at w picks out a possible future of w, we can now demand that for any $\langle h, m\rangle$ and A : $f_{\langle h,m\rangle}(A) \subseteq \{\langle h', n\rangle \in A : \langle h, m\rangle \leq \langle h', n\rangle\}$. Note that by means of this constraint we assure that for all A and $K(a, w)$: $C_{K(j,w)}(A) \subseteq K^*(j, w)$, just like we wanted.

It is quite easy now to reformulate the analyses of desire discussed in the previous sections in terms of imaging. For instance, we could say that you desire A, or intend to make A true, if the *utility* of A is higher than doing nothing, or higher than any other relevant alternative action/proposition, where the utility of A, $u(A)$ is defined as $\sum_w P_A(w) \times d(w)$. Alternatively, we might use the conditional analysis for intention, and say that John intends A in w, iff doing A assures John to fulfill his goals:

$$[[intend(j, A)]]^w = 1 \quad \text{iff} \quad C_{K(j,w)}(A) \subseteq Bul(j, w, K^*(j, w))$$

Notice that neither of the two analyses predicts that intentions are closed under believed consequences, or side effects. Bratman (1987), followed by Cohen & Leveque (1990), argued that, indeed, intentions should not be closed under believed consequences. Consider Susan, who has a toothache. She intends to get rid of the tootache by getting her tooth filled. She believes, however, that getting her tooth filled will cause her much pain, because she is not informed about anaesthetics. Still, it seems reasonable to assume, she does not have the intention to be in pain.

Although our analyses predict that intentions are not closed under believed consequences, it is predicted that for any A and B that are believed to causally entail each other, i.e. $C_{K(j,w)}(A) = C_{K(j,w)}(B)$, it follows that by intending the one you automatically also intend the other. However, it seems that even this is too strong a prediction: Suppose John intends to become rich, A, and believes that the only way for him to become rich is to work very hard, B. Thus, John believes that $A \rightsquigarrow B$ is true, where $\rightsquigarrow$ is our non-backtracking counterfactual connective. But John also has a lot of faith in himself, and believes that if he would work hard, he would also become rich. So he also believes $B \rightsquigarrow A$. In other words, the condition $C_{K(j,w)}(A) = C_{K(j,w)}(B)$ is satisfied. Still, in at least one sense of the word, I can imagine John intending to become rich, but

not intend to work very hard; if John *would* find out that he could become rich without working hard, he *would* go for that option.

What this argument suggests is that you can only intend something, if your desire or goal to do it is relatively immune, or *stable*, *under belief revision*. According to Bratman (1987), it is this stability of intentions that make them so useful for agents; we don't have to deliberate at each moment whether we should do a certain action or not.

If intention is an attitude that is relatively stable under belief revision, it shares a lot with another attitude, the attitude of *knowledge*. In fact, it seems even plausible to analyse intention partly in terms of knowledge. But before we will come to that, let us sketch how knowledge attributions might be analysed.

8.2 Knowledge

Just like other evidential verbs, also knowledge is normally analysed as something like *belief* plus something extra. One extra thing is obviously that what is known also has to be *true*. But true belief cannot be enough, as can be illustrated by the following examples of Russell:[20]

> It is clear that knowledge is a subclass of true beliefs. [...] There is a man who looks at a clock when it is not going, though he thinks that it is, and who happens to look at it at the moment when it is right; this man acquires a true belief as to the time of day, but cannot be said to have knowledge. There is the man who believes, truly, that the last name of the prime minister in 1906 began with a B, but who believes this because he thinks that Balfour was prime minister then, whereas in fact it was Campbell Bannerman. (Russell, 1948, pp. 170-171)

What should the extra condition be that turns a true belief into knowledge? Traditionally it was assumed that this extra condition should be a notion of *justification*. It seems reasonable to assume that the notion of justified belief should be analysed in terms of extended belief states that represents also the agent's belief revision policies. Moreover, if we want to analyse knowledge as justified true belief, where justified belief is analysed in terms of extended belief states, the 'internal' notion of justified belief and the 'external' notion of truth should somehow be related with each other. The question is *how?* I would like to propose simply to follow Hintikka's prime intuition about knowledge:

> It may be useful to remember that for us the primary sense of "I know that p" is the one in which it is roughly equivalent to "p, and no amount of further information would have made any difference to my saying so". (Hintikka, 1962, p. 52)

[20] See also Gettier (1962) for some similar examples.

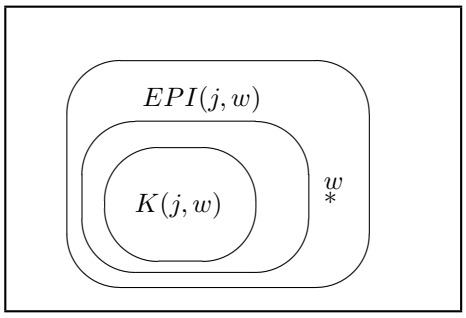
What this quotation suggests is that an item is known iff the item is believed, and it would not be given up by the acceptance of any new proposition that is true. Thus, an item of belief counts as knowledge, iff it is *robust* with respect to the truth (cf. Stalnaker, 1996). We can formalise this idea as follows:[21]

$$[[know(j, A)]]^w = 1 \quad \text{iff} \quad \bigcup\{K(j, w)^*_B : \ B \subseteq W \ \& \ w \in B\} \subseteq A$$

It is easy to see that this interpretation rule accounts for the *factivity* of knowledge. The reason is that one of the propositions B will be the maximal proposition that is only true in w, the proposition $\{w\}$. It will obviously be the case that for any K and w it holds that $K^*_{\{w\}} = \{w\}$.

Let us see whether our analysis can account for Russell's problems. We have already suggested that our analysis can account for the first problem; if the agent hears that the instrument on which he based his belief was not reliable, he probably wouldn't believe anymore the item he actually believed. The second problem is also straightforwardly accounted for; if the man is informed that the late prime minister is not Mr. Balfour, he probably wouldn't believe anymore that the name of the later prime minister began with a 'B'.

It turns out that our analysis of knowledge gives rise to an epistemic accessibility function $EPI(a, w)$ that can be defined as $\{v \in W : \ v \preceq w\}$. Obviously, if we want to say that John knows A in w iff $EPI(j, w) \subseteq A$, it has to be the case that $EPI(j, w)$ can also be defined as $\bigcup\{K(j, w)^*_B : \ B \subseteq W \ \& \ w \in B\}$.[22] It is easy to see that this is indeed the case.[23] In a picture, our analysis of knowledge would look as follows:



[21] I should note that my analysis of knowledge is the same as the analysis given by Stalnaker (1996), although some of the main ideas were developed independently.

[22] Our analysis of *knowledge* gives rise to the logic $S4.3$, characterised by an accessibility relation that is reflexive, transitive and connected, thus *not* euclidean. As a result, it is predicted that EPI does not satisfy negative introspection, just like most philosophers advised us.

[23] **Proof:** Because each true proposition, B, is a superset of $\{w\}$, it is obviously the case that $\forall v \in K^*_B : \ v \preceq w$. But this means that for each v in $\bigcup\{K^*_B |\ B \subseteq W \ \& \ w \in B\}$ it holds that $v \preceq w$, which shows the equation from right to left. For the other side, let v be a world such that $v \preceq w$. But then it will be the case that there is a true proposition B such that $v \in K^*_B$, namely $B = \{v, w\}$.

8.3 Back to intention

As we have seen above, Bratman (1987) has argued that intention is not closed under believed consequences, not even if the believed consequence is also believed to be a sufficient condition for the intention. I have suggested above that this can be accounted for by saying that what one intends is a desire or goal that is *stable* under belief revision. I want to suggest tentatively that *John intends* A iff (i) John desires to do A, and (ii) almost no amount of further information would change that desire. A crude way to implement this is to say that doing A satisfies his desires not only with respect to his *belief* alternatives, but also with respect to his *epistemic* alternatives:

$$[[intend(j, A)]]^w = 1 \quad \text{iff} \quad C_{EPI(j,w)}(A) \subseteq Bul(j, w, EPI^*(j, w))$$

Notice that it is predicted now that even if John believes that A and B are causal consequences of each other, he can still intend the one, without the other. The stronger condition that needs to be fulfilled now is that John must *know* that the two are causal consequences of each other: $C_{EPI(j,w)}(A) = C_{EPI(j,w)}(B)$, a condition that in our above example is probably not fulfilled.

References

[1] Asher, N. (1987), "A typology for attitude verbs and their anaphoric properties", *Linguistics and Philosophy*, 10, pp. 125-197.

[2] Bratman, M.E. (1987), *Intention, Plans, and Practical Reason*, Cambridge.

[3] Cohen, P.R. & H.J. Leveque (1990), "Intention is choice with commitment", *Artificial Intelligence*, 42, pp. 213-261.

[4] Fraassen, B.C. van (1973), "Values and heart's command", *Journal of Philosophy*, 70, 5-19.

[5] Gärdenfors, P. (1988), *Knowledge in Flux, Modeling the Dynamics of Epistemic States*, Cambridge Mass., MIT Press.

[6] Gettier, E. (1963), "Is justified true belief knowledge?", *Analysis*, 6, pp. 121-123.

[7] Gibbard, A. and W.L. Harper (1978), "Counterfactuals and two kinds of expected utility", In: C. Hooker et al. (eds.), *Foundations and Applications of Decision Theory*, Western Ontario Series in the Philosophy of Science, Vol. 1, Reidel, Dordrecht.

[8] Hansson, S. O. (1989), "A new semantical approach to the logic of preference", *Erkenntnis*, 31, 1-42.

[9] Harper, W.L. (1975), "Rational belief change, Popper functions, and counterfactuals", *Synthese*, 30, 221.

[10] Harper, W.L. (1976), "Ramsey test conditionals and iterated belief change", In: W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol I, Reidel, Dordrecht.

[11] Heim, I. (1992), "Presupposition projection and the semantics of attitude verbs", *Journal of Semantics*, 9, pp. 183-221.

[12] Hintikka, J. (1962), *Knowledge and Belief*, Ithaca, NY: Cornell University Press.

[13] Jeffrey, R. (1965), *The Logic of Decision*, McGraw-Hill, New York. University of Chicago Press, Chicago.

[14] Katsuno, H. and A. Mendelzon, (1991), "On the difference between updating a knowledge database and revising it", In: *Proceedings of the 2nd International conference on Principles of Knowledge Representation and Reasoning.*

[15] Kratzer, A. (1981), "Partition and revision: the semantics of counterfactuals", In: *Journal of Philosophical Logic*, 23, pp. 35-62.

[16] Lewis, D.K. (1975), "Probabilities of conditionals and conditional probabilities", *The Philosophical Review*, 3, pp. 297-315.

[17] Price, H. (1989), "Defending desire-as-belief", *Mind*, pp. 119-127.

[18] Rescher, N. (1967), "Semantic foundations for the logic of preference", In: N. Rescher (ed.), *The Logic of Decision and Action*, University of Pittsburgh Press.

[19] Russell, B. (1948), *Human Knowledge, its scope and limits*, London.

[20] Stalnaker, R.C. (1968), "A theory of conditionals", *Studies in Logical Theory, American Philosophical Quarterly Monograph Series, No. 2*, Blackwell Oxford.

[21] Stalnaker, R.C. (1970), "Probability and conditionals", *Philosophy of Science*, 37.

[22] Stalnaker, R.C. (1984), *Inquiry*, Cambridge, MA. MIT Press/Bradford Books.

[23] Stalnaker, R. C. (1996), "Knowledge, belief and counterfactual reasoning in games", *Economics and Philosophy*, 12, pp. 133-163.

[24] Thomason, R. H. (1970), "Indeterminist time and truth value gaps", *Theoria*, 36, pp. 264-281.

[25] Wright, H. von (1963), *The Logic of Preference*, Edinburgh.