人　工　知　能 79－2
（1991. 11. 26）

信念とコミットメントの変更に関する時間推論

磯崎 秀樹
NTT 基礎研究所

ロボットや人間などの知的行為者にとって，外界の情報とそれに基づくプラン生成は重要な役割を果たす．特に多数の行為者がお互いのことを考えている環境では，各行為者は他の行為者の持っている情報やプランなども考慮しなければならない．このような状況を扱うため，ここでは情報とプランをそれらに対応する信念とコミットメントの様相記号によって表し，その時間変化を扱うプログラミング言語の実現について考える．まず，どのようなことが記述できるかの例を示し，理論化あるいは実装する際の問題点を述べ，最後に簡単な推論システムの実現方法を説明する．

Temporal Reasoning about Revisions of Beliefs and Commitments

Hideki Isozaki
NTT Basic Research Laboratories

For intelligent agents such as robots and human beings, informations about the external world and plan generation based on them play important roles. In particular, if agents are thinking about each other, they should take other agents' informations and plans into account. Here we introduce modal operators for belief and commitment so as to represent informations and plans. First, we show what kind of things they can represent. Second, we describe their problems which we encounter when we intend to formalize/implement them. Finally, we show how to build a simple reasoning system.

1 はじめに

スタンフォード大学の Shoham の提案している「エージェント指向プログラミング (Agent-Oriented Programming, AOP)[1, 2]」では，エージェントのプログラムが心的状態に基づいて記述される．心的状態の記述法としては belief (確信，信念[†1]) と commitment (決意，約束，責務)[†2] という二つの様相記号 [3] にさらに時間を加えた論理が現在用いられている．

belief は $\mathrm{B}_a^t\varphi$ という記法で表され，エージェント a が時刻 t に φ を信じている (at time t agent a believes φ) ということを表す．一方 commitment は $\mathrm{CMT}_{a,b}^t\varphi$ という記法で表され，時刻 t にエージェント a がエージェント b に対して φ を義務づけられている，あるいは「決意」(約束) している (at time t agent a is obligated, or committed, to agent b about φ) ということを表す．ただしここでは，$\mathrm{CMT}_{a,b}^t\varphi$ の b を省略した $\mathrm{CMT}_a^t\varphi$ の形を用いる．このため，Shoham のものとはすこし論理体系が違ってくる．

ここでは，まずこれらの様相記号を用いて日常の様々な心理現象がどの程度表現できるか，その表現力を示し，次にその実現には様々な問題が伴うことを説明する．最後に簡単な推論機構を実現する方法について述べる．

なお，ここではいきなりある特定の論理体系の意味論や決定可能性や無矛盾性などを論じるのではなく，エージェント指向プログラミングの現在の記法を用いてどのようなことが記述可能であるかのイメージを膨らませ，今後の個々の理論構築の段階で生じるさまざまな問題点や選択肢や式と式との関係などについて一般的に考えることにする．

2 「信念＋コミットメント＋時間」の表現力

この節では，「信念」と「決意」(「コミットメント」では長いので，以下では表題を除いて「決意」という言葉を用いる．) で日常の心理現象の近似的な記述を試み，考察を加える．以下に列挙するさまざまな定義式は，いずれもごく粗い近似に過ぎず，特に時間関係については改良の余地があるが，表現能力の可能性を示すのが目的であるので，ここではあまり正確さを追求していない．なお「信念」や「決意」の内容を表すのに p^t や φ という表現を用いる．p^t は性質 p が時刻 t で成り立っているということであり，φ は p^t の時刻の部分を明示する必要がない時に用いる．たとえば，$\mathrm{B}_a^{t^{[1]}}(p_1^{t^{[2]}} \wedge \mathrm{B}_b^{t^{[2]}} p_2^{t^{[3]}})$ という具体的な式を抽象化し，$\varphi = p_1^{t^{[2]}} \wedge \mathrm{B}_b^{t^{[2]}} p_2^{t^{[3]}}$ とおいて $\mathrm{B}_a^{t^{[1]}}\varphi$ と表したり，$p = p_1 \wedge \mathrm{B}_b p_2^{t^{[3]}}$ とおいて $\mathrm{B}_a^{t^{[1]}} p^{t^{[2]}}$ と表したりする．

2.1 「信念」だけからなるもの

- $\mathbf{know}_a^t\varphi \stackrel{\mathrm{def}}{=} \varphi \wedge \mathrm{B}_a^t\varphi$: 信じていることが成り立っている，つまり，「知っている」[†3]
- $\mathbf{nonforgetting}_a\varphi \stackrel{\mathrm{def}}{=} \forall t, t'\ s.t.\ t' < t.\mathrm{B}_a^{t'}\varphi \supset \mathrm{B}_a^t\mathrm{B}_a^{t'}\varphi$: 「昔あることを信じていた」ということをその後ずっと信じている，つまり，「忘れない」(なぜ右辺が二重の信念になっているか疑問に思われるだろうが，これは「昔信じていたが何らかの理由で今は信じていない」ようなことを考えてみればよい．信念を変更するということと「忘れる」ということが別のことであることが分かるであろう．

[†1] 「信念」という言葉は日常用語としても使われるのでまぎらわしいが，ほぼ定訳になってきているのでそのまま使った．

[†2] commit という言葉も訳しにくい．訳しにくい単語はそのままカタカナにするというのが一つの解決方法であるが，commit の場合はそれも難しい．“a is committed” は「a がコミットされている」となり，誰か他の人に責任があるような感じになってしまう．

[†3] 逆に知識 K から信念 B を定義する方法もある．[8]

「昔信じていた」ということを忘れてしまっては，「忘れない」とは言えない．）

2.2 「コミットメント」だけからなるもの

- $\mathbf{compete}^t_a\varphi \stackrel{\text{def}}{=} \mathrm{CMT}^t_a\varphi \wedge \mathrm{CMT}^t_b\neg\varphi$: お互いに矛盾することをしようとしている．つまり「競合」
- $\mathbf{succeed}_a p^t \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' < t.\mathrm{CMT}^{t'}_a p^t \wedge p^t$: あることが実現する直前に，それを実現することを「決意」していた．つまり「成功」
- $\mathbf{order}^t_{a,b}p^{t'} \stackrel{\text{def}}{=} \forall t''\ s.t.\ t < t'' < t'.\mathrm{CMT}^t_a\mathrm{CMT}^{t''}_b p^{t'}$: 他人があることをするようにする．つまり「依頼」
- $\mathbf{forbid}^t_{a,b}p^{t'} \stackrel{\text{def}}{=} \forall t''\ s.t.\ t < t'' < t'.\mathrm{CMT}^t_a\neg\mathrm{CMT}^{t''}_b p^{t'}$: 他人があることをしないようにする．つまり「禁止」

2.3 「信念」と「コミットメント」の両方を含むもの

- 通知
  - $\mathbf{persuade}^t_{a,b}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' > t.\mathrm{CMT}^t_a\mathrm{B}^{t'}_b\varphi$: 他人にあることを信じさせようとしている．つまり，「説得」（この場合，$\mathrm{B}^t_a\varphi$ である必要があるかどうかは微妙である．ここではこの条件をはずしておいた．）
  - $\mathbf{inform}^t_{a,b}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' > t.\mathrm{CMT}^t_a\mathrm{B}^{t'}_b\mathrm{B}^t_a\varphi$: 自分があることを信じていると他人に信じさせようとしている，つまり「通知」

  日常生活を考えてみればわかる通り，何かを他人に知らせることは容易であるが，他人を説得することはなかなか難しい．これは，$\mathrm{B}^{t'}_b\mathrm{B}^t_a\varphi$ の状態から $\mathrm{B}^{t'}_b\varphi$ の状態に移すのが b の心的操作であって，b が他に $\mathrm{B}^{t'}_b\mathrm{B}^t_c\neg\varphi$ のような信念を持っていないか，b が a をどれくらい信頼しているか，などに依存するためである．
- 不誠実な通知
  - $\mathbf{deceit}^t_{a,b}\varphi \stackrel{\text{def}}{=} \mathrm{B}^t_a\neg\varphi \wedge \mathbf{persuade}^t_{a,b}\varphi$: 自分が信じていないことを他人に信じさせようとしている．つまり「だます」（ただし，$\mathrm{B}^t_a\neg\varphi$ か $\neg\mathrm{B}^t_a\varphi$ か $\neg\varphi$ かは微妙．）
  - $\mathbf{lie}^t_{a,b}\varphi \stackrel{\text{def}}{=} \mathrm{B}^t_a\neg\varphi \wedge \mathbf{inform}^t_{a,b}\varphi$: 自分が信じていないのに信じていると他人に思わせようとしている．つまり「嘘」
- 確認
  - $\mathbf{confirm}^t_{b,a}\varphi \stackrel{\text{def}}{=} \exists t', t''\ s.t.\ t' < t < t''.\mathbf{inform}^{t'}_{a,b}\varphi \wedge \mathbf{order}^t_{a,b}\mathbf{inform}^{t''}_{a,b}\varphi$: 知らせてもらったことをもう一度知らせてくれるよう依頼する．つまり「確認」
  - $\mathbf{observe}^t_a p^t \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' > t.\mathrm{CMT}^t_a((\mathbf{know}^{t'}_a p^t) \vee (\mathbf{know}^{t'}_a \neg p^t))$: あることの真偽を自分が将来知っている状態を「決意」する．自分で「観測」すればよい．
- 質問と回答
  - $\mathbf{clarifywhich}^t_{a,b}\varphi \stackrel{\text{def}}{=} \mathbf{inform}^t_{a,b}\varphi \vee \mathbf{inform}^t_{a,b}\neg\varphi$: φ を知らせるか $\neg\varphi$ を他の人に知らせる．つまり「真偽を明らかにする」
  - $\mathbf{inquirewhich}^t_{a,b}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' > t.\mathbf{order}^t_{a,b}\mathbf{clarifywhich}^{t'}_{b,a}\varphi$: 真偽を明らかにするよう他人に依頼する．つまり「質問」
- 要求に対する態度
  - $\mathbf{obey}^t_{b,a}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' < t.\mathbf{order}^{t'}_{a,b}\varphi \wedge \mathrm{CMT}^t_b\varphi$: 他人から依頼されたことを「決意」する．つまり「従う」

- $\mathbf{defy}^t_{b,a}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' < t.\mathbf{order}^{t'}_{a,b}\varphi \wedge \mathtt{CMT}^t_b\neg\varphi$: 他人の依頼と矛盾することを「決意」する．つまり「反抗」

- 通知に対する態度
  - $\mathbf{disagree}^t_{b,a}\varphi \stackrel{\text{def}}{=} \exists t'\ s.t.\ t' < t.\mathbf{inform}^{t'}_{a,b}\varphi \wedge \mathbf{inform}^t_{b,a}\neg\varphi$: 相手が知らせてくれたことと両立しないことを知らせる．つまり「意見の不一致」
  - $\mathbf{dispute}^t_{a,b}\varphi \stackrel{\text{def}}{=} \mathbf{persuade}^t_{a,b}\varphi \wedge \mathbf{persuade}^t_{b,a}\neg\varphi$: お互いに両立しないことを説得しようとしている．つまり「論争」

- 自分の信念に対する態度
  - $\mathbf{takemeasures}^t_a p^{t'} \stackrel{\text{def}}{=} \exists t''\ s.t.\ t'' < t < t'.\mathrm{B}^{t''}_a p^{t'} \wedge \mathtt{CMT}^t_a\neg p^{t'}$: 予測されることを避けるよう「決意」する．つまり「対策」

3 「信念＋コミットメント＋時間」の問題点

以上のように，「信念」と「決意」の組合せにより，実に多様な日常の心理現象を記述できる可能性が明らかになった．もちろん記述するだけでは不十分であって，無矛盾な公理系として構築することを考えたり，計算機上にいかにして実装するかということについて検討しなければならない．

しかし，この分野は不明なことがまだあまりに多く，一気に満足のいく論理体系を組み立てることは難しい．そこで実際に計算機上で実装し，さまざまな例題を扱ってみることによって，どのような理論体系が適当であるか検討していくという立場をとることにする．これはまた，AGENT0[2] のようなエージェント指向プログラミング言語をいかに実現するかということにもなる．

たとえば belief は，信念の様相記号と限量子の可換性の問題 (Barcan formulae) や，エージェント間の用語の不統一の問題や，内包 /外延の問題など様々な問題点を持っているが，我々が特に興味を持っているのは次の問題である．

- 整合性の問題．複数のルートから伝わってきた情報が矛盾している場合，自分としてはどの情報を信頼し，どの情報を捨てるか，あるいは態度を保留するか．

また，時間には次のような問題が伴う．

- 持続性の問題．持続性とは，たとえばある物体がある状態に置かれた場合に，何もしなければずっとそのままの状態でいることを表す．「コップが机の上にある」などの世の中の多くの物理現象は持続性を持っている．信念やコミットメントのような心的状態もその例外ではない．ある人が一旦あることを信じ始めれば，特にそれに反する情報を新たに入手しない限り，そのことを信じ続けるであろうし，一旦あることを「決意」(約束) すれば，それが不可能であることに気づくなどの特殊な事情でもない限り，その「決意」を取り消すことはないであろう．
  しかし，すべての事象が無限の持続性を備えているわけではなく，他からの働きかけがあれば変化するし，ある種の事象はもともと有限時間しか持続しない．効率良くどの時点で何が成り立っているかを求めるにはどうしたらよいか．
- 粒度の問題．数学的に物理現象を記述するには時間を実数と対応づけて考えるのが便利である．日常生活では時間は分や秒などのある時間粒度の整数倍に丸めて表す．実際，計算機のハードウエアのマシンサイクルレベルの話をしているような場合はそれでよいが，一般の事象を記述する場合には時間粒度が固定されていると不自由である．
  たとえば性質 p の持続性を $p^t \wedge Mp^{t+1} \supset p^{t+1}$ のような規則を含む非単調論理やデフォールト推論に

従ってそのまま計算するのは，イベントの発生が時間粒度に対して疎な場合には無駄が多すぎるし，密な場合には複数のイベントが単位時間内に発生してしまいがちで，本来ある順序関係を表現できない．

- 実数の問題．時間を実数とすると，実数を論理体系の中に導入しなければならない．しかし，実数を一階の述語論理の対象であるかのように考えるのは，厳密にいえば問題がある．このため，できれば実数を導入せずにすませたい．
- 曖昧さの問題．時刻そのものは全順序であるとしても，様々な理由によってエージェントは時刻を厳密に知ることが出来ないので，時間を含む推論は必然的に曖昧にならざるをえない．

これらの時間の問題については，特にアルゴリズムやインプリメンテーションの立場からいくつかの研究 [5, 6] がなされている．これらは事象の発生時刻の順序関係を直接記述できるようにし，定量的な情報が入ってきてもある程度は扱えるように拡張している．

このようにいくつもの複雑な問題がからまりあった系を扱うので，理論の面においても実装の面においてもいかにこれらの諸問題を上手に処理するかが AOP の実現の鍵を握っている．

3.1 信念とコミットメントの多次元持続性と心的時間地図

すでに述べたように，「信念」も「決意」も持続性を持っている．これらは Shoham[1] の表現をまねれば，整数時間では $\mathrm{B}_a^t\varphi \wedge \neg\mathrm{LEARN}_a^t\neg\varphi \supset \mathrm{B}_a^{t+1}\varphi$, $\neg\mathrm{B}_a^t\varphi \wedge \neg\mathrm{LEARN}_a^t\varphi \supset \neg\mathrm{B}_a^{t+1}\varphi$, $\mathrm{CMT}_a^t\varphi \wedge \neg\mathrm{REVOKE}_a^t\varphi \supset \mathrm{CMT}_a^{t+1}\varphi$ と表わせる．

さて，$\mathrm{B}_a^{t^{[1]}}p^{t^{[2]}}$ が $t^{[1]} < t^{[2]}$ から $t^{[1]} = t^{[2]}$ を経て $t^{[1]} > t^{[2]}$ になった場合について考えよう．$t^{[1]} < t^{[2]}$ の場合，これは将来のことを予測していることになる．$t^{[1]} = t^{[2]}$ のところでは実際にこの予測があたっているか外れているかが決定される．この結果は $t^{[1]} < t^{[2]}$ において自分で直接観測できるかも知れないし，他のエージェントから教えられるかもしれない．そうして実際にどうだったかの情報が伝わってきた後でも $\mathrm{B}_a^{t^{[1]}}p^{t^{[2]}}$ が成り立っているとすると，それは予測通りだったのだし，$\mathrm{B}_a^{t^{[1]}}\neg p^{t^{[2]}}$ になったとすると予測が外れたということである．

同様に $\mathrm{CMT}_a^{t^{[1]}}p^{t^{[2]}}$ が $t^{[1]} < t^{[2]}$ から $t^{[1]} = t^{[2]}$ を経て $t^{[1]} > t^{[2]}$ になった場合について考えよう．$t^{[1]} < t^{[2]}$ の場合，これはある命題を将来実現しようと決意していることになる．$t^{[1]} = t^{[2]}$ のところでは実際にこの決意が達成されたかどうかが決定される．結果についてはBと同様である．

問題は $t^{[1]} > t^{[2]}$ の場合に $\mathrm{CMT}_a^{t^{[1]}}p^{t^{[2]}}$ がどういう意味を持っているかである．つぎのような立場が考えられる．

- $\forall t^{[1]} > t^{[2]}.\neg\mathrm{CMT}_a^{t^{[1]}}p^{t^{[2]}}$: 過去のことは「決意」できない．
- $\forall t^{[1]} > t^{[2]}.\mathrm{CMT}_a^{t^{[1]}}p^{t^{[2]}} \supset \mathrm{B}_a^{t^{[1]}}p^{t^{[2]}}$: 今でも信じていることだけ「決意」できる．

われわれはここでは後者の立場をとる．

このように両者とも持続性をもっているが，その中身の持続性に対する挙動はまったく違う．たとえば，「さっき入った部屋の入口のそばに本棚があった」ということをエージェントが信じていたとしよう．すると，そのエージェントはだれかが動かさない限りその本棚は動かないと思っているはずだから，「いまでもその場所にその本棚がある」と信じているであろう．このように，対象の持続性は信念の中でも有効である．

ところがこれに対して「決意」の様相演算子の中では，このようなデフォールトの持続性が使えない．たとえば，朝から夕方までの会議があるとして，「会議の開始時に会議に出席している」ということを「決意」していて，会議を途中で抜け出すことを特に「決意」していなくても，「会議中ずっと会議に出席している」ことを「決意」しているとは言えない．もちろん「会議中ずっと会議に出席してい

るだろう」という信念は，この「決意」と「決意」に関する内省と「信念」の持続性から導かれる．もし「昼食までずっと会議に出席している」つもりであれば「決意」も持続しているが，これはデフォールトによるものではなく，その持続そのものが明確な「決意」によるものなのである．この差は次のような場合に行動の差としてはっきり現れる．もしこの会議の開催場所が昼休みをはさんで第一会議室から第二会議室に変わるとしたら，「会議の開催中はずっと会議に出席している」ことを「決意」している人は第二会議室に移動するだろうが，「会議の開始時に出席している」ことしか決意していない人は，昼食までたまたま第一会議室にいたとしても，第二会議室には行かないであろう．このように「決意」の中ではデフォールトの持続性が使えない．[†4]

この様子をもっとはっきり明示するため，心的状態の時間変化を示す地図を導入する．欲求や意図などの他の心的状態も含め，一般的にそのような地図を「心的時間地図 (Temporal Mental State Map)」，特に信念に関するものを「信念時間地図 (Temporal Belief Map)」という．

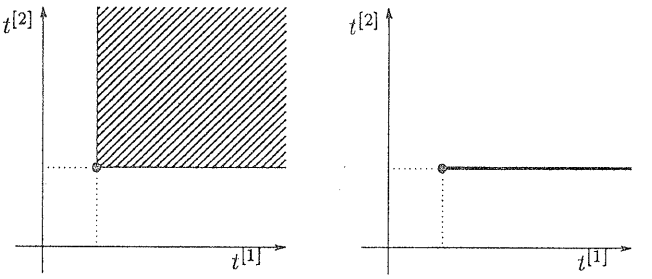

図 1: Persistence in Temporal Mental State Maps (left: $\mathrm{B}_a^{t^{[1]}} p^{t^{[2]}}$, right: $\mathrm{CMT}_a^{t^{[1]}} p^{t^{[2]}}$)

多重信念の場合，時間地図は 3 次元以上になって，紙の上に描くことは難しくなる．しかし $\mathrm{CMT}_a^{t^{[1]}} p^{t^{[2]}}$ の地図を見れば判るように，CMT が一次元の持続性しか持たないので，CMT を含む式は見かけよりもっと低い次元で議論できる場合がある．

たとえばエージェント b がコピー機のところにいるところをエージェント a が見かけたとする．そして「エージェント b はコピーをすることを「決意」しているのだ」と信じた場合の a の心的状態 $\mathrm{B}_a^{t^{[1]}} \mathrm{CMT}_b^{t^{[2]}} p^{t^{[3]}}$ を左に，エージェント c がエージェント d を説得しているときの c の心的状態 $\mathrm{CMT}_c^{t^{[1]}} \mathrm{B}_d^{t^{[2]}} q^{t^{[3]}}$ の地図を右に示す．

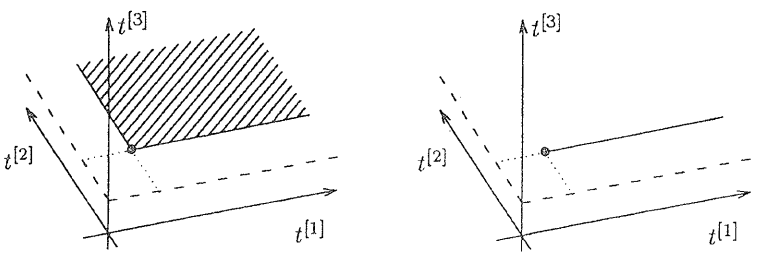

図 2: Persistence in Temporal Mental State Maps (left: $\mathrm{B}_a^{t^{[1]}} \mathrm{CMT}_b^{t^{[2]}} p^{t^{[3]}}$, right: $\mathrm{CMT}_a^{t^{[1]}} \mathrm{B}_b^{t^{[2]}} p^{t^{[3]}}$)

[†4] 知識 $\mathrm{K}_a^{t^{[1]}} p^{t^{[2]}}$ など多くの心的状態を表す様相演算子が $\mathrm{CMT}_a^{t^{[1]}} p^{t^{[2]}}$ と同様にその内容の持続性を消してしまうことに注意．知識と信念のこの違いは，NTT コミュニケーション科学研究所の赤埴淳一氏に指摘していただいた．

3.2 心的時間地図の例

このような地図を想定することにより，エージェントの心的状態に関する議論を，以下の二つの段階に分割できる．[†5]

- 入力データからそのエージェントの現状認識を表現した心的時間地図を構成する段階．
- その心的時間地図にもとづいて行動を決定する段階．

計画を立てると心的時間地図が変更を受けるので，計画は心的時間地図に対する操作ともみなせる．

前の段階で問題となるのは，外界に関する情報の整合性である．もし，入力データが整合的であれば，すべてをそのまま信じても構わない．もし矛盾が見つかれば，確認 (inquirewhich/confirm/...) など，なんらかの方法でこれを解決しなければならない．矛盾したデータについては態度を保留し，どちらを信じていても大丈夫なように計画を立てることも時にはできようが，いつもそうとはかぎらない．さまざまな事情を総合して判断していくしかないであろう．これについては機会を改めて説明することにする．

さて，後の段階，つまり行動計画の方をもう少し詳しく考えてみよう．自分がある行動を起こすと，それ以降にその影響が及ぶ．そしてそのことを自分は計画した時点で予測するので，この行動は信念についても外界についてもそれ以降の時間 (二次元信念時間地図ならその点を頂点とする右上四半平面) に影響を与える．

たとえば次の図は一度目の「決意」を中断して二度目の「決意」で成功した様子を示す．

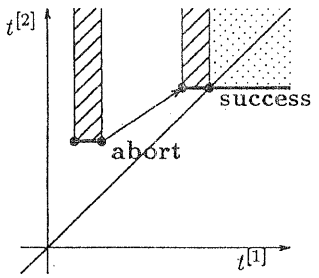

図 3: Retry (shaded regions: $\neg \mathrm{B}_a^{t^{[1]}} \neg p^{t^{[2]}}$, dotted region: $\mathrm{B}_a^{t^{[1]}} p^{t^{[2]}}$, thick lines: $\mathrm{CMT}_a^{t^{[1]}} p^{t^{[2]}}$)

なお図中の矢印は，「信念」による「決意」の「正当化」である．この「正当化」は AGENT0 の「決意規則」(commitment rule)[1, 2] に当たる．「決意規則」は次のような形をしている．

COMMIT(メッセージ条件，心的条件，相手のエージェント，行為)

これはあるメッセージを受けとった時に，そのメッセージがある条件を満たしていて，その時の自分の心的状態がある条件を満たしていれば，あるエージェントのためにある行為を決意せよ，ということを指定したルールである．

この「決意規則」を使えば，「地下鉄の駅を知っていますか」という yes/no 質問を発したエージェントに対して「地下鉄の駅への道順」を通知したり，「12 月 29 日の飛行機を予約したい」という希望を述べたエージェントに対して「12 月 28 日ではどうですか」と提案したりすることが正当化できる．

[†5] 動画像認識におけるオプティカルフローの役割に似ている．動画像認識は入力画像からオプティカルフローを求める段階とオプティカルフローから三次元情報を復元する段階に分割できる．

この図の例の場合，中止したという「信念」とそれによって正当化される再試行の「決意」は $t^{[2]}$ 軸方向について順方向であるが，予測される事態を回避するような場合は $t^{[2]}$ 軸の向きと逆方向になることもある．このように，ある予測に基づいてそれを回避するような行動の例として，McDermott[4] の「Nell の救助」を取り上げよう．McDermott[4] は同じ問題を分岐時間論理で表現しているが，我々はいわば「多次元時間論理」で表すことになる．

Nell というエージェントが線路に縛り付けられて (tied) いて，電車に轢かれ (mashed) そうな状態にあるとする．これを Dudley というエージェントが発見して (found)，駆けつけ (rush)，Nell を解放し (release)，一緒に逃げる (runaway) という状況を考える．この場合の Dudley の様々な命題 $\{p_i\}$ に対する心的状態 ($\{\mathrm{B}^{t^{[1]}}_{\mathrm{Dudley}} p_i^{t^{[2]}}\}$, $\{\mathrm{CMT}^{t^{[1]}}_{\mathrm{Dudley}} p_i^{t^{[2]}}\}$) は，まとめて次の左図のように表せる．

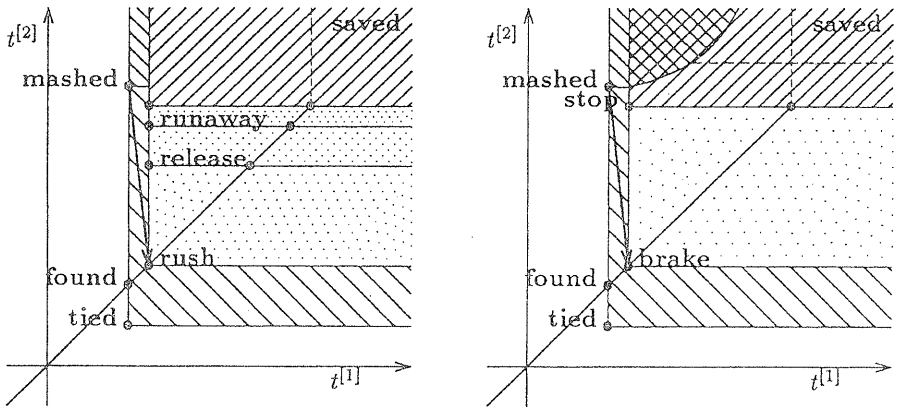

図 4: How to save Nell (left: by removing her, right: by braking)

ただし，この図では，線路に縛り付けられていることに気づいてから電車に轢かれることを予測するまでの遅延時間を無限小とし，「決意」した通りの時間に仕事をこなしたものとして単純化してある．

右の図は Dudley の救助がない代わりに，近付いてきた電車の運転士が Nell を発見し，ブレーキをかけて電車を止め，未然に事故を防いだという場合を想定して，その運転士の心的状態を描いたものである．

このように予測した自体を回避するような行動の場合は，「決意」の正当化を表す矢印が $t^{[2]}$ 軸の未来から過去へ向かう．しかしどんな場合でも，正当化の矢印が $t^{[1]}$ 軸の未来から過去へ向かうことはない (因果律)．

右図ではブレーキをかけても電車がすぐに止まらないことを考慮に入れたので，図がやや複雑になっている．ブレーキを少しかけただけでは，電車は減速するだけで止まらないので，t_{mashed} は延びるが轢かれてしまうことには変わりがない．しかしずっとブレーキをかけ続けていると，そのうち電車は止まるので，轢かれることはなくなる．このように電車が停止するまでブレーキをかけ続けることによって Nell の死ぬ時刻が双曲線的に遅れていく．右図に現れる t_{mashed} の $t^{[1]}$ による変化を表す右上がりの曲線はその事情を表したものである．もし，電車の運転手が突然ある時点でブレーキをかけるのをやめると，t_{mashed} の時間はその $t^{[1]}$ の時点での値に固定されるだろう．そのときの様子を $t^{[1]}$ に平行な破線で示す．†6

†6 ブレーキをかけた時の曲線はたとえば次のようにして求められる．電車は運転手がブレーキをかけ始めた時点 (t_{brake}) で $x_0 > 0$ 離れた場所にいて，速度 $v_0 > 0$ で近づいているとする．ブレーキをかけると電車が一定加速度 $-\alpha < 0$ で減速すると

4　「信念＋コミットメント＋時間」の実現方法

実際にエージェントのプログラムを書く場合には，前節の t_{mashed} の曲線のような定量的な情報をなるべく使わないで済むようにプログラミング言語を設計し，応用プログラムを書けるようにしないと，融通がきかず不便である．

そこで実際のアルゴリズムやインプリメンテーションを考えるため，状態は何らかの瞬間的事象（行為）によって中断されない限り永遠に持続しする仮定をおき，すべての領域は軸に平行な直線（多次元の場合は超平面）によって囲まれていると考える．このような領域は二次元なら矩形，三次元なら直方体，n 次元なら n 個の時区間の直積によって定義される集合の集合和で表現できる．

これらの n 次元の矩形領域は，おのおのその対角線の両端の頂点の座標の対によって表せる．このようにすれば，有限個の境界線（境界面，境界超平面）しか含まない心的時間地図を，有限個の点で表現できる．もしこれらの点の座標の定量的な値が少々曖昧だったり途中で多少の変更があったとしても，その全順序関係が変わらない限りは，その順序関係だけから導かれた定性的な結論をそのまま使える．

そしてこの有限個の点の座標の順序の情報が与えられれば，その点で分割された領域では心的状態の真偽が同じなので，その値に関する質問には容易に答えられる．この質問回答機能をプリミティブな述語とした Pure Prolog のようなインタプリタを Prolog 上で作成することは容易である．もう少し難しい質問としては，どれくらいの範囲である心的状態が成り立っているか極大な時間領域を答えよという質問がある．これはどれくらいの範囲なら仕事をずらしても影響が出ないかを調べる時に必要な機能である．この種類の質問については，Sripada の時間演繹データベースの考え方 [9] が利用できる．

このような考え方に基づき，これまでに以下のものを実装した．

- 二次元の場合の信念時間地図の有限個のデータによる表現を入力データから簡単なアルゴリズムによって決定し，それに従って描画するプログラム．これは C で書かれている．
- 多次元の信念時間地図上の特定の点における命題の真偽を問う質問に対し，信念時間地図によって得られる結果と等価な結果を，地図の有限データ表現を構成せずに入力データから直接求める Quintus Prolog のプログラム．

  以上の二つの項目の原理については本稿で説明しなかったので，別の機会に述べる．また「決意」も含めた一般の心的時間地図のインプリメントも行なう予定である．
- 自分のできる私的行為（右手を挙げるなど）と通信行為（inform など）だけで直接達成できる命題に関する未達成の「決意」を調べ，すぐに実行できそうなものから処理していくインタプリタ．これは Quintus Prolog で書かれている．このインタプリタは Sripada の PhD 論文を入手する前にインプリメントした．したがって Sripada の手法は用いていない．かわりに Pure Prolog のインタプリタのプリミティブに相当する部分を前項のプログラムに質問する単純な構造になっている．

  まだ試行的なプログラムなので，通信機能は不十分であるし，「決意」を一般の命題として扱っており，「決意」と「信念」のネストもまだ扱っていない．

  この項についてもさらに検討を加えて別の機会に詳述する予定である．

仮定し，ある時点 $t^{[1]}$ で減速するのをやめ，その時点の速度 v_1 のままで走り続けたら，電車はいつ Nell の位置に達するかを求めて t_{mashed} とする．もちろん一般的にどのような減速の仕方をしても，t_{mashed} は単調増加し，Nell の手前で止まるなら，有限時間で（つまり速度がゼロになるときに）t_{mashed} は無限大に達する．ブレーキをかけているのに Nell を轢いてしまうのは，t_{mashed} の曲線が $t^{[1]} = t^{[2]}$ と交差する時である．

5 おわりに

単一エージェントの知識の変更に関して個々の理論が満たすべき性質を定性的に記述した「Gärdenfors の公準」[7] というものが知られている．この公準は知識の変更に関する様々な理論の共通の性質を明らかにし，同一の枠組でそれらの際を比較し論じるのに役立つ．

我々はこれに相当するものをエージェント指向プログラミングの枠組において構築しようとしている．そのためインプリメンテーションによって検証しながらいろいろな理論的可能性についても検討している．

ここでは，信念とコミットメント (決意) の様相記号と時間の意義と問題点，そしてその実現方法について概要説明を行なった．

なお，本稿は著者がスタンフォード大学ロボティクス研究所滞在中に行なった研究が元になっている．この研究に関しては，Y. Shoham 助教授をはじめ，ICOT の古川康一博士，東京理科大の溝口文雄教授，HP の Kave Eshgi 博士，NTT CS 研の赤埴淳一氏，NTT 基礎研の後藤滋樹博士，勝野裕文氏など多数の方々のご協力をいただいた．感謝したい．

参考文献

[1] Y. Shoham: Agent-Oriented Programming, Stanford Technical Report, STAN-CS-1335-90, 1990.

[2] Y. Shoham: AGENT0: A simple agent language and its interpreter, Proc. of AAAI, 1991.

[3] B. F. Chellas: Modal logic, an introduction, Cambridge University Press, 1980.

[4] D. McDermott: A Temporal Logic for Reasoning About Processes and Plans, Cognitive Science, Vol.6, pp.101–155, 1982.

[5] T. L. Dean, D. V. McDermott: Temporal data base management, Artificial Intelligence, Vol. 32, pp.1–55, 1987.

[6] R. Kowalski, M. Sergot: A logic-based calculus of events, New Generation Computing, Vol.4, pp.67–95, 1986.

[7] P. Gärdenfors: Knowledge in Flux, MIT Press, 1988.

[8] Y. Shoham, Y. Moses: Belief as Defeasible Knowledge, Proc. of IJCAI, pp.1168–1172, 1989.

[9] S. M. Sripada: Temporal Reasoning in Deductive Database, PhD thesis of Univ. of London 1991.