修 士 論 文

マルチエージェントシステムにおける
知識状態の遷移観測機構の構築

北陸先端科学技術大学院大学
情報科学研究科

西山 功太郎

2014 年 3 月

修 士 論 文

マルチエージェントシステムにおける
知識状態の遷移観測機構の構築

指導教員　東条敏 教授

審査委員主査　東条敏 教授
審査委員　島津明 教授
審査委員　Nguyen Minh Le 准教授

北陸先端科学技術大学院大学
情報科学研究科

1210042 西山 功太郎

提出年月: 2014 年 2 月

概 要

近年SNSやその上で展開されるソーシャルゲームなど，インターネット上において，テキストを用いた多くのコミュニケーションが急速な発達を見せている．中でもここ数年，「汝は人狼なりや？」という元来はテーブルゲームの一種であるが，インターネット上で掲示板やチャットを用いて広く行われているトークゲームが流行の兆しを見せている．これは，8〜13名程度のプレイヤーが人間役と狼役に分かれて会話を行いながら，人間役は狼役を，狼役は人間役を根絶させる事を目標にゲームを行う．このゲームは，プレイヤー同士は誰が狼役なのかが分からないまま会話を繰り返し，その中に含まれている発言の矛盾や揺らぎなどを指摘しながら嘘つきを見つける，あるいは騙し続ける，という点が特徴である．

一方で，近年の様相論理の研究，特にダイナミック論理と認識論理の研究において，アナウンスメントやゲーム，通信などにおける共有知識の更新や信念変更に関する様相論理系としてDynamic Epistemic Logic(DEL)が提案されている．DELにおいては，公開的告知やプライベートな情報伝達など，多様な行為が当事者達の認識状態を変化させる出来事として解釈される．例えば，ある公開的告知をされた後では皆がその公開的告知で得られた知識を必ず知っているという事が成り立つ．その公開的告知で得られた知識が，元から自分が持ち合わせていた知識と矛盾する場合などは，公開的告知から得られた知識を棄却し，その告知をしたエージェントは嘘つきであると捉える場合と，告知を受け入れて現在の知識状態が更新されるという場合が考えられる．

また，DELの一種である公開告知論理の拡張としてエージェント告知論理がある．エージェント告知論理は，エージェントが他のエージェントに対して嘘やブラフを告知する事を定義する為に考案された様相論理系で，エージェントが他のエージェントに対して嘘の告知を行う事や，自分自身で真偽のついていない事についてブラフの告知を行う，といった告知を定義している例えば，エージェントがある論理式が真であると知っている時，他のエージェントに対してある論理式は偽であると告知を行う事で，他のエージェントを騙すといった告知を表現する事が可能である．

本研究では,「汝は人狼なりや？」のゲームシーンをマルチエージェントシステムのモデルとして扱い，インターネット上の掲示板やチャットで行われたプレイログをシステムの入力として用いる．また，エージェントの知識状態の更新や信念変更における様相論理系のDELないしはその改変形を用いて，発言(公開告知)が行われた時点での知識状態の遷移を観測する事が出来る機構を構築する．具体的には，公開告知が行われた際に，各エージェントが持つ知識状態と，それに準ずる，あるいは反する知識の候補を得て，それぞれのエージェントがその公開的告知を受け入れるか，あるいは棄却するかという事が確認できる機構を提案する．

目 次

図 目 次

第1章　はじめに

1.1　研究の背景

近年 SNS やその上で展開されるソーシャルゲームなど，インターネット上において，テキストを用いた多くのコミュニケーションが急速な発達を見せている．中でもここ数年，”Are You A Werewolf?”[1] という元来はテーブルゲームの一種であるが，インターネット上で掲示板やチャットを用いて広く行われているトークゲームが流行の兆しを見せている．これは，8〜13 名程度のプレイヤーが人間役と狼役に分かれて会話を行いながら，人間役は狼役を，狼役は人間役を根絶させる事を目標にゲームを行う．このゲームは，プレイヤー同士は誰が狼役なのかが分からないまま会話を繰り返し，その中に含まれている発言の矛盾や揺らぎなどを指摘しながら嘘つきを見つける，あるいは騙し続ける，という点が特徴である．

一方で，近年の様相論理の研究，特にダイナミック論理と認識論理の研究において，アナウンスメントやゲーム，通信などにおける共有知識の更新や信念変更に関する様相論理系として Dynamic Epistemic Logic(DEL) が提案されている [2]．DEL においては，公開的告知やプライベートな情報伝達など，多様な行為が当事者達の認識状態を変化させる出来事として解釈される．例えば，ある公開的告知をされた後では皆がその公開的告知で得られた知識を必ず知っているという事が成り立つ．その公開的告知で得られた知識が，元来自分が持ち合わせている知識と矛盾する場合などは，その告知から得られた知識が棄却され，告知をしたエージェントは嘘つきと捉える場合と，告知を受け入れ現在の知識状態が更新されるという場合が考えられる [3].

また，DEL の一種である公開告知論理の拡張としてエージェント告知論理がある [12]. エージェント告知論理は，エージェントが他のエージェントに対して嘘やブラフを告知する事を定義する為に考案された様相論理系で，エージェントが他のエージェントに対して嘘の告知を行う事や，自分自身で真偽のついていない事についてブラフの告知を行う，といった告知を定義している例えば，エージェントがある論理式が真であると知っている時，他のエージェントに対してある論理式は偽であると告知を行う事で，他のエージェントを騙すといった告知を表現する事が可能である．

また，DEL の一種である公開告知論理を拡張した様相論理系としてエージェント告知論理がある [12]．エージェント告知論理は，エージェントが他のエージェントに対して嘘やブラフを告知する事を定義する為に考案された様相論理系で，エージェントが他のエージェントに対して嘘の告知を行う事や，自分自身で真偽のついていない事についてブラフ

の告知を行う，といった告知を定義している例えば，エージェントが φ であると知っている時，他のエージェントに対して $\neg\varphi$ であると告知を行う事で，他のエージェントを騙す，といった告知が可能である．

1.2 研究の目的

本研究の目的は，マルチエージェントシステムにおいて，発言によって各エージェントの知識がいかに遷移し，それがいかに目標の達成に影響するかを観測するシステムを構築することである．通常のマルチエージェントシステムの場合，全てのエージェントが互いに知識を持ち寄りながら協力して同一の目標を達成しようと行動することにより，単体では解決出来ない複雑な問題を解決することが出来る．

しかし，目標の達成を阻害する虚偽の発言を意図的に行う非協力的なエージェントが集団に混ざっている場合，目標の達成が困難になる事が予想される．本研究では，このような協力的エージェントと非協力的エージェントが混在する状況において，各エージェントが，正当性が保証されない発言を自分の持つ知識といかに照らし合わせながら受け入れ，知識状態がいかに推移するかを観測するシステムを実装し，発言が知識状態に与える影響を検証する．

1.3 本論文の構成

本論文では，2 章においてマルチエージェントシステムの概要を紹介する．また，本研究で取り上げるマルチエージェントシステムのモデルとして”Are You A Werewolf?”と”Muddy Children Puzzle”を取り上げる．

3 章では，エージェントモデルのための論理についての説明を行う．特に本研究で着目している公開告知論理とエージェント告知論理に関しての説明を詳しく行う．続く 4 章では，本論文で提案する機構に関する説明を行う．設計，実装環境などに触れた上で，どのような実装を行ったを説明する．

5 章においては，提案した機構に対して検証を行い，評価実験とそれに対する考察を行う．続く 6 章にて，本論文のまとめと今後の展望について示す．

第2章　マルチエージェントシステム

本章では，マルチエージェントシステムの概要と本論文がマルチエージェントシステムに対してどのようなアプローチを取るかを述べたのち，本論文で研究対象として取り上げる対話型パーティーゲーム”Are You A Werewolf?”と論理パズル”Muddy Children Puzzle”について説明する．

2.1　マルチエージェントシステムとは

マルチエージェントシステムとは，自律的・社会的といった特徴を持つソフトウェアやロボットを意味するエージェントが複数集まり，個々の問題を解決しながら，全体としての問題を解くシステムを指す．一般に，エージェント単体では解決出来ない問題も，マルチエージェントシステムのアーキテクチャを利用する事によってエージェント間で分業化を実現し，容易に解決出来るようになる場合がある [8]．現実的な応用例としては，複数のロボットが自律的に動作する惑星探査ロボットなどがある．このようなケースでは，逐一ロボットを操作しようとすると，膨大な遅延が生じる為，ロボットが自律的に意思決定を行い，知的に行動する事が期待される．近年はこういった自律的なロボットを複数用意し，自律的に協調させながら，より複雑な作業を行わせる試みがなされている [9]．

マルチエージェントの研究には，AI，複雑系，経済学など様々なアプローチが存在する．本研究では知的エージェントが複数集まった時に，全体としていかに問題解決を行うかという点に焦点を当てる．特に AI 分野におけるエージェントとは，自分自身で動く自律性，他のエージェントや人間と対話する社会性，自ら目的を持って行動する目的志向性，環境の変化に対して自ら反応する反射性などの特徴を持つソフトウェアやロボットを指す [10]．本研究においては，様相論理やそれを基に拡張された論理で規律や行動を定義する事で，社会的な規則や倫理に反することなく行動する社会的エージェントを実現する論理的アプローチを取る．

2.2　本論文にて取り上げる対象

2.2.1　Are You A Werewolf?

Are You A Werewolf?とは，日本語では「汝は人狼なりや？」あるいは単に「人狼」とも訳されるパーティーゲームで，元来は 1930 年頃からヨーロッパでプレイされていた伝

統的なゲームが基になっている．元々は自作のカードやトランプなどを用いてプレイされていたが，2001 年にアメリカのゲームメーカーである Looney Labs. がカードセット「Are You A Werewolf?」の販売を開始した．もともとが伝統的なゲームである為，権利上の問題もなく，以降は各社から商業版のカードゲームとして様々なゲームメーカーから同様のカードゲームが出版された．日本においてはイタリアの daVinici 社が発売した「タブラの狼」がカードセットとしては最も普及している．図 2.1 に「タブラの狼」のカードの一部を示す．

図 2.1: 「タブラの狼」カードセットの一部

また，インターネットの普及により 90 年台後半からインターネット掲示板やチャットなどでルールがまとめられたり，プレイされるようになり始める．カードゲームとしてプレイする場合よりも，掲示板やチャットを用いた方が常に対面している必要が無い等環境によらないプレイが可能であったり，プレイ人数が確保しやすいなどの理由から特に普及が進んだ．近年はその普及から，テレビでも芸能人が人狼をプレイする番組などが放映される事がある．

基本的なゲームの概要としては，プレイヤーが人間役と人間に化けた狼役とに分かれ，狼役は自分自身の正体がばれないように会話を行いながら，人間役は狼役を，狼役は人間役を根絶される事が目標及び勝利条件になる．ゲームは半日単位で進行し，昼ターンにはプレイヤー同士で会話を行い，最後にプレイヤーの投票により決定された人狼の容疑者を処刑する．夜ターンには人狼による村人の捕食が行われ，処刑や捕食されたプレイヤーは以降ゲームから除外される [1]．以下に，ゲームの流れを示す．

[1]「タブラの狼」では処刑候補者の選抜にのみ除外されたプレイヤーも参加出来る

1. 夜ターンの開始

2. 人間側の特殊能力を持つプレイヤー達が能力を行使する

3. 人狼が人間を一人だけ襲撃する

4. 昼ターンの開始

5. 村人全体で議論を行い，その日の人狼容疑者を一人だけ処刑する

6. 人間の人数と人狼の人数が等しくなるか，人狼を全滅させるまで 1. に戻る

人間役のプレイヤーの一部は夜ターンにおいて特殊能力を行使する事ができ，特殊能力には，指定したプレイヤーが人狼であるか否かを判別出来る，処刑されたプレイヤーが人狼であったか否かを判別出来るなど様々な種類があり，そのような能力を持つプレイヤーはそれを他のプレイヤーに通達する事で議論のヒントを与える事が出来る．そして，プレイヤーの人数により，狼役や特殊能力を持つ人間役の数は増減する．例を挙げると,「タブラの狼」ではプレイ人数は 8 人から 24 人とされるが，8 人〜15 人でゲームをプレイする際には狼役が 2 人だが，16 人以上の場合は狼役が 3 人となる．また，8 人でプレイする際には役職を持つ人間役は 1 名だが，9 人以上でプレイする場合には増えた人数分だけ通常の人間役，もしくは特殊能力を持つ人間役を導入する事が勧められている．

また，一般的に対面でプレイする際にはゲームの進行役となるゲームマスター (「タブラの狼」では調停者と呼ぶ) が必要となるが，インターネット上でプレイする際には，ゲームマスターの役割を行うプログラム（自動的に役職を割り振る，設定された時間で昼夜の各ターンを進行させる等）を組み込んだ人狼をプレイする専用の掲示板やチャットなども多数存在し，ローカルルールで特殊な役職が存在したり，特別な遊び方なども数多く存在する．

このゲームは，狼役のプレイヤー以外は，誰が人狼役なのか分からないまま会話を繰り返す事になる [2]．その会話の中で役職を持った村人役は自分の能力を発揮し，周りに自分が得た情報を教え，その内容を信じさせる事が重要であるし，人狼役は自分の身を守る為に推理を混乱させるような発言をする，もしくは自分が役職を持った村人であるかのように振る舞うなどが重要になる．また，役職を持たない通常の村人役は人狼役の意図的な邪魔を回避しつつ正しい意見を信じる事で，正確な推理を行っていち早く人狼役を特定する事が重要である．

例として，図 2.2 と図 2.3 にチャット形式で行われた人狼ゲームの一場面 [3] を示す．なお，画像の上方に行くほど新しい発言である．まず，図 2.2 では赤枠で囲われた部分において，3 名のプレイヤーが自分が占い師という能力者であると宣言している [4]．その上の

[2] 人狼同士は誰が人狼であるかは把握出来る

[3] http://werewolf.ddo.jp/log/log13761.html「汝は人狼なりや？」るる鯖 No.13761「17A 普通」村 より一部抜粋

[4] CO：人狼チャットにおけるスラングの一種でカミングアウトの略記

黄枠では別の１名のプレイヤーが自分が霊能者という能力者であると宣言している．ルール上，占い師と霊能者は必ず１名ずつしか存在しないので，占い師を名乗った３名の内，少なくとも２名は嘘をついている事が分かり，霊能者を名乗ったプレイヤーはある程度信頼出来そうだ，という事が全プレイヤー間で推測する事が出来る．

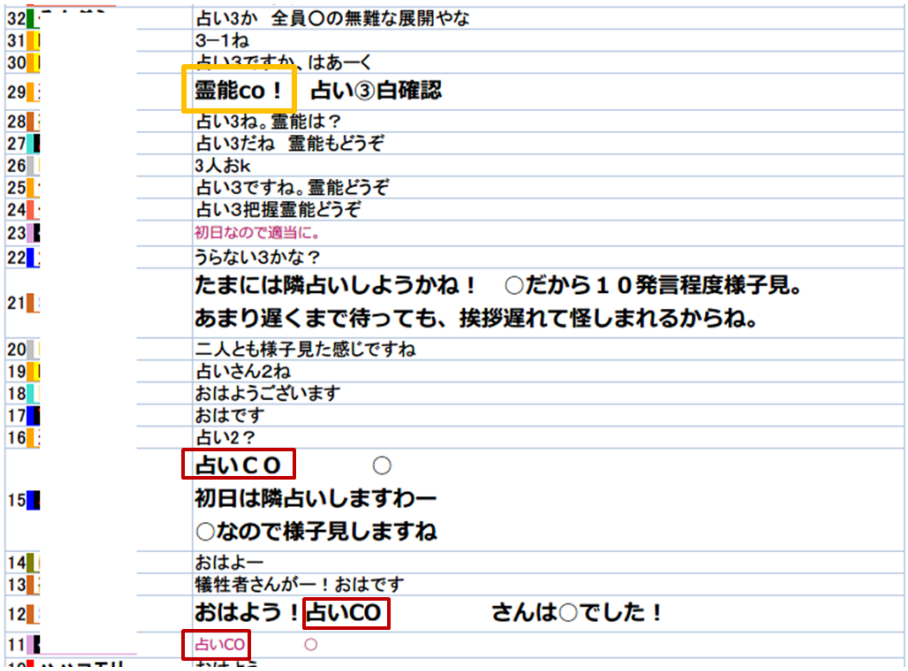

図 2.2: チャット人狼の会話の一例（1）

更にこの次の日，図 2.3 のような会話が行われた．こちらの画像も同様に下からターンが進んでいる為，上の発言がより新しい発言となっている．赤破線で区切られた一番下のブロックにおいて，先程の３名の占い師[5]だと名乗ったプレイヤーの内，一人が「（プレイヤー A）●」と発言している．「●」とは，このような人狼チャットや掲示板内でよく使用されるスラングの一つで，人狼を指す言葉である．即ち，この占い師は「自分の占いの結果，プレイヤー A は人狼である事が分かった」という意味の発言をしている．別の占い師を名乗るプレイヤー２名は，同一のプレイヤーを指して「（プレイヤー B）〇」と発言しているが，こちらは「自分の占いの結果，プレイヤー B は人間である事が分かった」という意味になる．

今のこの場面では，このゲームをプレイしているプレイヤーの多くは，誰が本物の占い師であるか判別がついていない．しかし，霊能者[6]を名乗ったプレイヤーは一人しか名乗りを上げていない為，３名居る占い師よりも信憑性は高いと推測される．そこで，そのターンでのプレイヤー間で行われた協議の結果，この１名の占い師から人狼であると判定されたプレイヤー A を処刑し，霊能者の結果と一致するか否かを確認する事になった．

[5]占い師：毎夜１人のプレイヤーを選択し，そのプレイヤーが人狼であるか否かを判定出来る特殊能力
[6]霊能者：毎夜，その日処刑されたプレイヤーが人狼であったか否かを判定出来る特殊能力

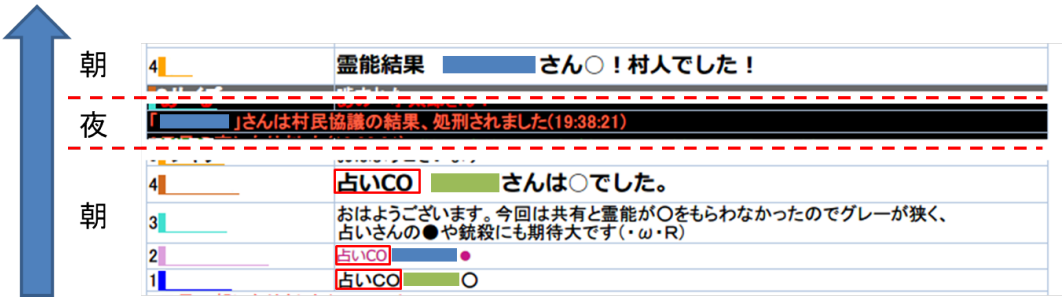

図 2.3: チャット人狼の会話の一例（2）

なお，このゲームでは狩人という毎夜プレイヤーを 1 名だけ人狼の襲撃から守る事が出来る役職が導入されているのだが，基本的なセオリーとして霊能者よりも占い師を守る，という事がある．それは，生きているプレイヤーの中から人狼であるか否かを判別出来る占い師の方が，死んだプレイヤーが人狼であったか否かを判別出来る霊能者よりもプレイヤーの推理に有益である，とされているからである．しかし，このシーンにおいては，霊能者を守る事が優先されると実際の狩人役のプレイヤーは判断した．

なぜならば，もし本物の占い師がその夜に人狼に殺害された場合，それは残りの 2 名がほぼ間違いなく人狼陣営のプレイヤーであると推論する事が出来る．逆に，その夜に霊能者が殺害されると，どの占い師が本物の占い師であったのかいつまでも知る事ができず，3 人の占い師候補の内誰を信じていいのか分からなくなるからである．故に，このシーンだけを切り取って考えると，狩人役はセオリーとされている占い師（しかも誰が本物なのかはまだ分からない）を守るよりも，信憑性が高く，翌日得られる情報の重要さから，霊能者を守った方が良い，と推理する事が出来るのである．

そしてその次の日，霊能者の発言は「（プレイヤー A）○」であった．即ち，「プレイヤー A は人間であった」という結果が得られたという旨の発言をしている．これにより，霊能者の出した結果と矛盾を発生させた占い師は，嘘つき（＝人狼役，ないしはその協力者[7]）である，という事が推測される．そして同様に，他の占い師 2 人から「人間である」と判定されたプレイヤー B は人間役である事がほぼ確定的である，という事も同時に推測される．

そのように会話の中に含まれている発言の矛盾や意見の揺らぎなどを指摘しながら人間役は人狼役の嘘を見つける，それに対して人狼役はなるべく矛盾を生じさせないように騙し続ける，といった点がゲームの勝敗を決定させる重要な要素であり，大きな特徴である．

[7]人間役でありながらも，人狼役に協力する役職も存在する

2.2.2 Muddy Children Puzzle

Muddy Children Puzzle とは，マルチエージェントシステムを論理的アプローチで解く場合の適用例としてよく挙げられる古典的な論理パズルの一つである [6]．Muddy Children Puzzle の概要としては以下の通りである．

> 3 人の子供が外で遊んで帰ってきた．
> 父親は子供達に「君たちの内，一人以上は額に泥をつけているよ」と言った．
> 続いて父親は「自分の額に泥がついているかどうか分かるかな？」と尋ねた．
> 3 人はそれに対して正直に答える．
> 父親はもう一度「自分の額に泥がついているかどうか分かるかな？」と聞き，
> 子供達も同様に正直に答える．
> この問答を何度続ければ子供達の全員が自分の額に泥がついているかどうか
> 分かるか.

という論理パズルである．この論理パズルに登場する子供達は常に正直で，賢いと仮定されている．また，父親の言葉も常に真である．

この論理パズルを解く場合，例えば 3 人の子供のうち 2 人だけが額に泥を付けていると仮定する．すると，最初の問答では，子供達は全員「分からない」と答える．しかし，次の問答では，泥を付けている子供達は次のように考える事が出来る．

> 「もし自分が泥をつけていなかったとしたら，泥をつけている子供は『自分が
> 汚れている』という事が分かるはずだ．しかし先程の問答では分からないと
> 答えた．故に自分は汚れている.」

子供達全員が正直者で，かつ賢いと仮定するならば，このような推論が可能となる．

第3章　エージェントに関する論理

近年，アナウンスメント，ゲーム，通信などにおける共通知識の更新や信念変更に関する様相論理系として動的認識論理 (Dynamic Epistemic Logic) が提案されている．本章では最初に様相論理の概要を述べた後に，動的認識論理の一種である公開告知論理 (Public Announcement Logic) についてその概要を述べ，次いで更にそれを拡張して考案されたエージェント告知論理 (Agent Anouncement Logic) の概要を述べる．

3.1　様相論理

一般的な推論を考える時，その多くは時間の流れや可能性，知識状態など，推論が行われる様々な状況を踏まえたものである．ここではそのような推論形式を分析する事を可能とする為に様相論理を導入する [11]．様相論理は一般的な古典論理に加え,「必然的に φ である」「φ である可能性がある」といった事象の可能性や必然性に関わる命題を扱う論理体系である．ここで,「必然的に φ である」「φ であると知っている」「φ であると信じている」ことを記号$\Box$を用いて $\Box\varphi$ と表す．この時，$\neg\Box\varphi$ は「φ は必然的ではない」「φ であると知らない（信じていない)」と，$\Box\neg\varphi$ は「φ でない事は必然的にである（φ である可能性はない)」や「φ でないと知っている（信じている)」などと解釈する事が出来る．また，$\Box\neg\varphi$ の否定，すなわち $\neg\Box\neg\varphi$ は $\Diamond\varphi$ と表し,「φ である可能性がある」と解釈される．これらの$\Box$や$\Diamond$をまとめて様相演算子 (modal operator) と呼び，$\Box$を必然演算子，$\Diamond$を可能演算子と呼ぶ．

3.1.1　構文論

様相論理の構文論は以下のように定義される．

1. 命題変数は論理式である
2. 命題定数 $\top$，$\bot$ は論理式である．ここで $\top$ は真を，$\bot$ は偽を表す
3. いまある論理式を φ,ψ と仮定する．この時，$\varphi\vee\psi$,$\varphi\wedge\psi$,$\varphi\to\psi$,$\neg\varphi$，$\Box\varphi$ はいずれも論理式である．また，$\Diamond\varphi$ は$\neg\Box\neg\varphi$ の略記であるとする
4. 以上によるもののみが論理式である

また，命題変数 p を仮定したとき，$\neg p$ と p をリテラルと呼ぶ．

3.1.2　クリプキ意味論

様相論理の意味論は以下に述べるクリプキフレームによって定められる．空でない集合 S と S 上の二項関係 R の対 (S, R) を（様相論理に対する）クリプキフレーム（もしくは単にフレーム）と呼ぶ．また，S 及び R をそれぞれこのクリプキフレームの可能世界（possible world）の集合及び到達可能関係（accessibility relation）と呼ぶ．

今，(S, R) をフレームとし，V を各可能世界 sに対して $V(p) \subseteq S$（$p \in P$ は命題変数）となる写像とする．この時，V をフレーム (S, R) 上の付値と呼ぶ．そしてこの三重対 (S, R, V) をクリプキモデルと呼ぶ．与えられたクリプキモデル (S, R, V) に対して M の要素と論理式の間の二項関係 $\models$ を以下のように帰納的に定義する．

1. $M, s \models p \Longleftrightarrow s \in V(p)$
2. $M, s \models \varphi \wedge \psi \Longleftrightarrow, M, s \models \varphi$ かつ $M, s \models \psi$
3. $M, s \models \varphi \vee \psi \Longleftrightarrow M, s \models \varphi$ または $M, s \models \psi$
4. $M, s \models \varphi \supset \psi \Longleftrightarrow M, s \models \varphi$ でないか，または $M, s \models \psi$
5. $M, s \models \neg\varphi \Longleftrightarrow M, s \models \varphi$ でない
6. $M, s \models \Box\varphi \Longleftrightarrow sRt$ となる，すべての t に対し $M, t \models \varphi$
7. $M, s \models \Diamond\varphi \Longleftrightarrow sRt$ となる，ある t に対し $M, t \models \varphi$

$M, s \models \varphi$ の時，論理式φ は s において真であるという．関係 $\models$ は付値 V から一意に定まるので，V と $\models$ を同一視して，$\models$ を付値といったり，$(S, R, \models)$ のことをクリプキモデルと呼ぶ場合もある．

クリプキモデルはより直感的に理解する為にしばしば図を用いて表現される．例えば，以下の通りのクリプキモデル (S, R, V) を考える．

- 可能世界の集合　$S = \{s_0, s_1, s_2\}$
- 到達可能性関係　$R = \{(s_0, s_0), (s_0, s_1), (s_0, s_2)\}$
- 付値　$V(p) = \{s_0, s_2\}$

このクリプキモデルを図に表すと図 3.1 となる．今，s_0 に着目すると，s_0 と s_1，そして s_2 の 3 つの可能世界への到達可能性が存在する．ここで，p と$\neg p$ が成り立つ可能世界に同時に到達可能性が存在している．即ち，s_0 においては，$\Diamond p$，$\Diamond\neg p$ が真になっている．$\Diamond$は「〜である可能性がある」と解釈される為，図 3.1 で考えると p が成り立つ世界への到達可能性 s_0Rs_0と s_0Rs_2 とが存在している．これは「p である可能性がある」事を意味している為，$\Diamond p$ が真となる．同様に，$\neg p$ が成り立つ世界への到達可能性 s_0Rs_1 が存在

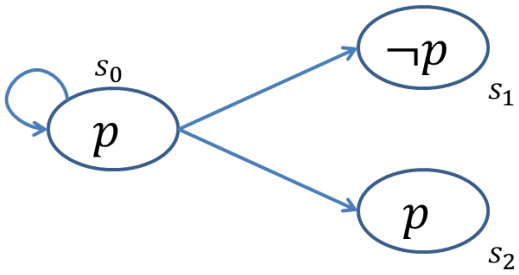

図 3.1: クリプキモデルの例（1）

している為，$\Diamond\neg p$ もまた真である．逆に，s_0 において「必ず p である」や「必ず $\neg p$ である」は成立していない為，$\neg\Box p$ が成立する．

例えば，s_0 において $\Box p$ が成り立つようなクリプキモデルとしては以下のように到達可能性を変更した図 3.2 などが考えられる．

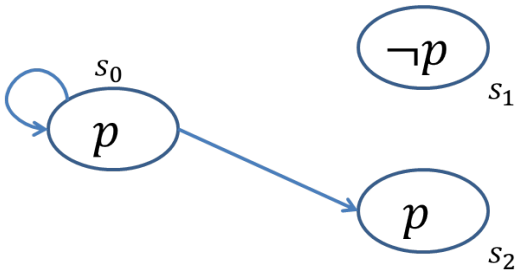

図 3.2: クリプキモデルの例（2）

図 3.2 における s_0 は到達可能性が s_0Rs_0 と s_0Rs_2 しか存在しない．即ち，p である可能世界への到達可能性しか存在しない為，$\Box p$ が成り立つ．図 3.2 のクリプキモデル (S, R, V) はこのような定義となる．

- 可能世界の集合　$S = \{s_0, s_1, s_2\}$
- 到達可能性関係　$R = \{(s_0, s_0), (s_0, s_2)\}$
- 付値　$V(p) = \{s_0, s_2\}$

3.1.3　到達可能性と様相論理式の対応

到達可能性の性質は様相論理式によって表現出来る．例えば，ある可能世界 s, t, u を仮定すると，次のような事が表現出来る．代表的なものとしてまず反射律 T がある．これは s から s へ到達可能であるかどうかを表している．次に推移律 4 は，s から t の到達可能性と t から u の到達可能性が存在する時，s から u へ到達可能であるかどうかを表す．その他にも，全ての可能世界は必ず一つ以上の到達可能な別の可能世界を持つ事を表す継続律

D，s から t の到達可能性が存在する時，t から s へ到達可能であるかどうかを表す対称律 B などがある．表 3.1 に代表的な論理式と対応する到達可能性関係を示す．

表 3.1: 様相論理の論理式と到達可能性関係

論理式	到達可能性関係	
$\Box\varphi \to \varphi$	反射的	s に対しても sRs
$\Box\varphi \to \Box\Box\varphi$	推移的	sRt かつ tRu ならば, sRu
$\Box\varphi \to \Diamond\varphi$	継続的	どの s に対しても，ある t が存在して sRt
$\varphi \to \Box\Diamond\varphi$	対称的	sRt ならば, tRs
$\Diamond\varphi \to \Box\Diamond\varphi$	ユークリッド的	sRt かつ sRu ならば, tRu

これらのフレームの構造によってその演算子の特性が規定され，様相論理で導入されている様相演算子も同様にこのフレームの構造によって特性が大きく変化する．以下に代表的な論理式を挙げる．

反射律 T

$T : \Box\varphi \to \varphi$

これは，反射律を認めるフレームを仮定すると，全ての可能世界においてそれぞれの可能世界がそれぞれの可能世界自身に到達する事が出来る事を示している．このとき，ある可能世界で「その先全ての可能世界において φ が成立する場合」を考えると，それぞれの可能世界は自分自身へその到達可能性関係を持つ為，その可能世界自身でも□演算子のつく命題 φ が成り立っている必要がある．

推移律 4

$4 : \Box\varphi \to \Box\Box\varphi$

これは，推移律を認めるフレームにおいては，ある可能世界 s から到達出来る可能世界 t に，そこから到達可能性のある別の可能世界 u が存在する時，可能世界 s から可能世界 u へ到達出来る事を示している．この時，可能世界 t において全ての到達可能性で φ が成立する可能世界があるとき、可能世界 t においては□ φ が成立すると言え，その可能世界 t に到達可能性を持つ可能世界 s においては□□ φ が成り立っている必要がある．

継続律 D

$D : \Box\varphi \to \Diamond\varphi$

これは，継続律を認めるフレームにおいては，全ての可能世界において到達可能性存在する事を示している．この時，ある可能世界において$\Box\varphi$が成り立っているとき，その世界ではある到達可能性でφが成立する必要がある．

対称律 B

B :$\varphi \to \Box\Diamond\varphi$

これは，対称律を認めるフレームにおいては，ある可能世界sから到達出来る可能世界tに，逆方向にも到達可能性が存在する事を示している．この時，φが成り立っている世界において，あるφが成り立つ可能世界に常に到達可能性がある必要がある．

ユークリッド律 5

$5 : \Diamond\varphi \to \Box\Diamond\varphi$

これは，推移律を認めるフレームにおいては，ある可能世界sから到達出来る可能世界tと，sから到達可能性のある別の可能世界uが存在する時，可能世界tから可能世界uへ到達出来る事を示している．

今，あるクリプキモデル$M = (S, R, V)$を考えたとき，

- 全ての可能世界$s \in S$において，ある可能世界$t \in S$への到達可能性が存在しているならば，Mは継続的であると言える．
- 全ての可能世界$s \in S$において，sRsの到達可能性が存在しているならば，Mは反射的であると言える．
- 全ての可能世界$s, t \in S$において，sRtという到達可能性があるときtRsが成立するならば，Mは対称的であると言える．
- 全ての可能世界$s, t, u \in S$において，到達可能性sRtとtRuが存在する時，sRuが成立するならば，Mは推移的であると言える．
- 全ての可能世界$s, t, u \in S$において，到達可能性sRtとsRuが存在する時，tRuが成立するならば，Mはユークリッド的であると言える．

3.1.4 様相論理のヒルベルト流公理系

様相論理の体系 K をヒルベルト流で公理化した **HK** を以下に示す．

(A1) $\varphi \to (\psi \to \varphi)$

(A2) $(\varphi \to (\psi \to \varsigma)) \to ((\varphi \to \varsigma) \to (\varphi \to \psi))$

(A3) $(\neg\varphi \to\to \psi) \to (\psi \to \varphi)$

(K) $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$

(MP) φ かつ$\varphi \to \psi$から,ψが推論出来る

(Nec) φ から，$\Box\varphi$ が推論出来る

これらに様相論理の代表的な公理系を加える事で，新たな以下の公理系が定義出来る.

- **HKD** は (A1),(A2),(A3),(K),(D),(MP),(Nec) から成る.
- **HKT** は (A1),(A2),(A3),(K),(T),(MP),(Nec) から成る.
- **HKB** は (A1),(A2),(A3),(K),(B),(MP),(Nec) から成る.
- **HS4** は (A1),(A2),(A3),(K),(T),(4),(MP),(Nec) から成る.
- **HS5** は (A1),(A2),(A3),(K),(T),(5),(MP),(Nec) から成る.

ここで，証明の定義を以下に示す.

> **HK** における証明とは，以下の条件を満たす論理式のリスト $B_1, \cdots, B_n$ を考えた時
>
> - B_1は $(A1), (A2), (A3), (K)$ である
> - $B_i(1 < i \leq)$ は $(A1), (A2), (A3), (K)$ の一つであるか,2つの論理式 $B_j, B_k(j, k < i)$ から (MP) によって導かれたものであるか，論理式 $B_j(j < i)$ から (Nec) によって導かれたものであるか
>
> これらを満たす時，n は証明の長さと呼び，B は **HK** の定理において B_nから B が導かれる.

ヒルベルト流公理系の完全性と健全性

今，$M = (S, R, V)$ において全ての可能世界 $s \in S$ において $V(\varphi, s) = 1$ が成立する φ が存在する時，φ は妥当であるという.

- $\vdash_{\mathbf{HK}} \varphi \Longleftrightarrow$ 全てのクリプキモデルにおいて φ は妥当である
- $\vdash_{\mathbf{HKD}} \varphi \Longleftrightarrow$ 全ての継続的なクリプキモデルにおいて φ は妥当である
- $\vdash_{\mathbf{HKT}} \varphi \Longleftrightarrow$ 全ての反射的なクリプキモデルにおいて φ は妥当である

- $\vdash_{\mathbf{HKB}} \varphi \Longleftrightarrow$ 全ての対称的なクリプキモデルにおいて φ は妥当である
- $\vdash_{\mathbf{HS4}} \varphi \Longleftrightarrow$ 全ての反射的かつ推移的なクリプキモデルにおいて φ は妥当である
- $\vdash_{\mathbf{HS5}} \varphi \Longleftrightarrow$ 全ての反射的かつ推移的かつ対称的なクリプキモデルにおいて φ は妥当である

3.2 公開告知論理

公開告知論理とは，動的認識論理 (Dynamic Epistemic Logic) の一種で，エージェントに対して情報を伝達する認識行為を定義したものである．そもそも動的認識論理は，知識変化についての推論の形式化を可能にする動的様相演算子を持つ認識論理の様々な拡張の総称である．一般的な動的認識論理では，認識演算子 K（あるいは B）を定義し，$K_a\varphi$ と記述する事で,「エージェント a が φ と知っている（信じている）」と解釈される．公開告知論理においては，一対多のエージェント間の相互作用を公開告知 (public anouncement) と呼び，これを行為演算子で形式化し，公開告知が実行されるとモデルが更新され，各エージェントの信念が更新する事を推論可能な論理体系である [7].

3.2.1 構文論

公開告知論理の構文論は以下のように定義される．

$$\mathcal{L}(!) \ni \varphi ::= p|\neg\varphi|(\varphi \wedge \psi)|K_a\varphi|\,[!\varphi]\,\psi$$

ここで，$K_a\varphi$ は「エージェント a はφ という事を知っている」と解釈でき，$[!\varphi]\,\psi$ は,「φ という誠実な公開告知の後では，必ず ψ が成り立つ」と解釈される．

3.2.2 クリプキ意味論

公開告知論理の認識モデル $M = (S, R, V)$ は以下のように定義される．

1. $M, s \models p \Longleftrightarrow s \in V(p)$
2. $M, s \models \neg\varphi \Longleftrightarrow M, s \not\models \varphi$
3. $M, s \models \varphi \wedge \psi \Longleftrightarrow M, s \models \varphi$ かつ $M, s \models \psi$
4. $M, s \models K_a\varphi \Longleftrightarrow$ 全ての t について, sR_at ならば $M, s \models \varphi$
5. $M, s \models [!\varphi]\,\psi \Longleftrightarrow M, s \models \varphi$ ならば $M|\varphi, s \models \psi$

ここで，$M|\varphi = (S', R', V')$ は以下のように定める．

- $S' = \{s' \in S | M, s' \models \varphi\}$
- $R'_a = R_a \cap (S' \times S') = \{(s', t') \in S' \times S' | s' R t'\}$
- $V'(p) = V(p) \cap S'$

これは，φ が真である可能世界や到達可能性だけを対象にしており，言い換えると $\neg\varphi$ である可能世界とそれに繋がるリンクを削除している事が分かる．

クリプキモデル

図 3.3 に公開告知論理によってクリプキモデルがどのように変化するかを示す．今，クリプキモデルを以下の通りに仮定する．

- 可能世界の集合　$S = \{s_0, s_1, s_2\}$
- 到達可能性関係　$R = \{(s_0, s_0), (s_0, s_1), (s_0, s_2)\}$
- $V(p) = \{s_0, s_2\}$

ここで，このクリプキモデルに対して $!p$ という告知を行ったと考える．すると，定義より p が真である世界以外を取り除いている事が分かる．

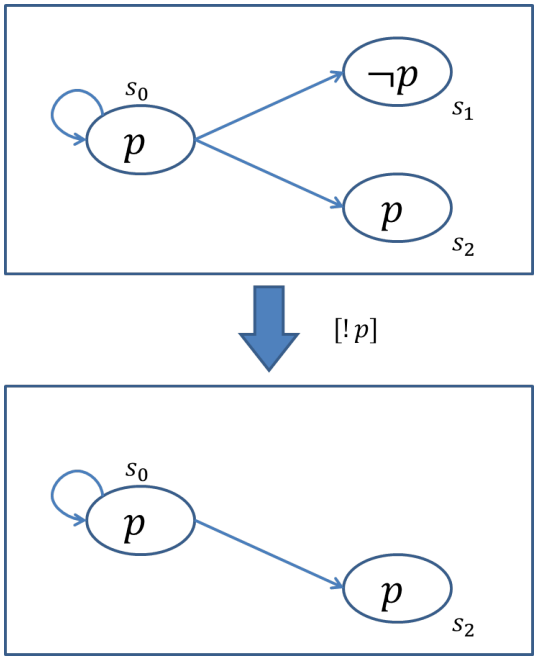

図 3.3: 公開告知によるクリプキモデルの変化

3.2.3 公開告知論理を用いたMuddy Children Puzzleの解法

本研究の解析対象として挙げたMuddy Children Puzzleは，本節で取り上げた公開告知論理を用いて解く事が出来る．以下に解法を示す．ここで，対象のモデルにおいては反射律Tと対称律B，推移律4が成り立つS5であるとする．

初期状態

まず，エージェントの知識状態の初期状態を図3.4に示す．紹介したMuddy Children Puzzleには3名の子供達が登場する．その子供達を公開告知を行うエージェントとして見做し，それぞれのエージェントが想像し得る可能世界の集合と，その間に存在する到達可能性のセットを示している．図中の円（ノード）一つ一つがそれぞれ可能世界を表しており，円の中に記されている数字は，左から順に子供A，B，Cが汚れている(=1)か，汚れていない(=0)かを表している．つまり，ノード000とは子供達全員が汚れていない世界を表しており，同様にノード111は子供A，B，Cと全員が汚れている世界，という事になる．なお，色が付いているノードが現在設定されている実世界を示している．即ち，今実際は子供A，Bが汚れていて，子供Cは汚れていない，という事になる

また，各ノードに接続されている青，赤，緑の実線は，それぞれ子供A，B，Cの到達可能性（リンク）を表しており，その子供は，その子供に対応した色のリンクで結ばれている世界間の区別が付かないという事を表している．例えば，子供Aは青のリンクで繋がっているノード000とノード100の区別が付かないという事である．これは，Muddy Children Puzzleにおいては，子供は自分の額が実際に汚れているかどうかは知る事が出来ないが，別の2人の子供が汚れているかどうかは観察から知る事が出来るからである．例えば，現在設定されている実世界はノード110だが，子供Aはノード010と，子供Bはノード100と，子供Cはノード111との区別がついていない事が分かる．

父親の告知

図3.4の知識状態を持つエージェントらに，まずは父親が「君達の内，一人以上が額に泥を付けている」と告知を行う．この時，エージェントAが汚れている事をd_aと表すとすると，父親の告知内容は論理式では$d_a \vee d_b \vee d_c$と表す事が出来る．これを公開告知論理における告知演算子を用いて表現すると，$[!(d_a \vee d_b \vee d_c]9$となる．これを満たす世界だけを残す，即ち，告知は実行される事で，これを満たさない世界を削除する．$d_a \vee d_b \vee d_c$を満たさない世界とは，ノード000しか存在しない．故に図3.4からノード000とそれに接続されているリンクを削除したエージェントらの知識状態を図3.5に示す．

図3.4から，父親の告知によってノード000とノード000へ繋がるリンクが子供達の知識状態から削除された事が確認出来る．

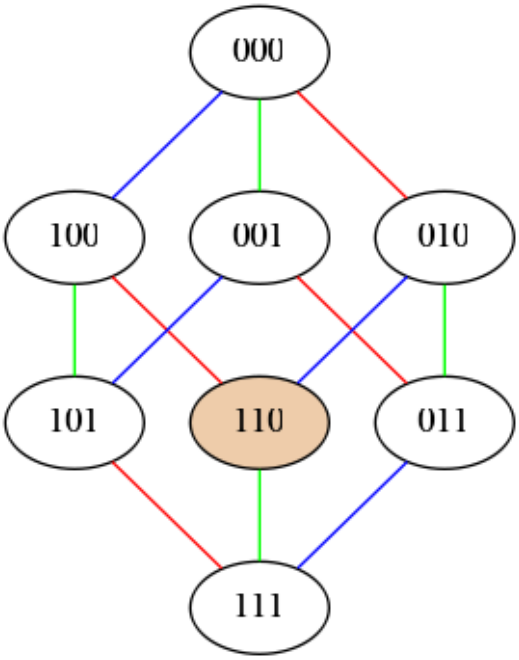

図 3.4: 公開告知論理：初期状態

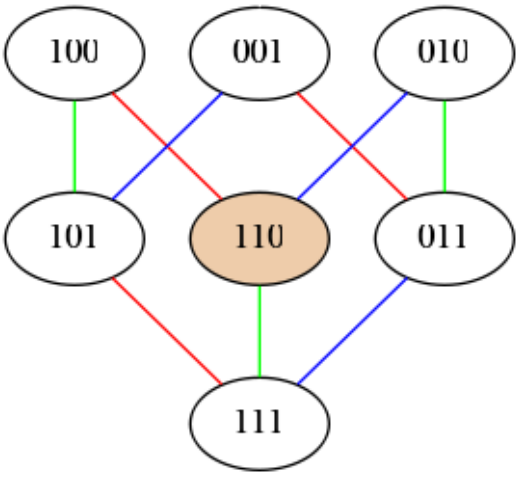

図 3.5: 公開告知論理：父親の告知を実行した後の知識状態

一度目の子供達の回答

では次に父親が「自分の額に泥が付いているかどうか分かるかい？」と子供達に対して尋ね，それに対して子供達が答える時を考える．図 3.4 において実世界であるノード 110 に注目する．すると，ノード 110 には 3 色のリンクがまだ接続されている事が分かる．即ち，3 人の子供達はまだ自分がどの可能世界に居るのか判別がついていない事が分かる．故に先程の父親の問いに対する子供達の答えは，全員「分からない」であると予想される．

子供 A が汚れているかどうか分からない，という状態を論理式で表すと，$\neg\Box_a d_a \wedge \neg\Box_a\neg d_a$ と表現出来る．この論理式を本来の意味のままで訳すると，

> 「エージェント A は，エージェント A が汚れていると知らない，かつ，エージェント A が汚れていないとも知らない」

となる．これを更に意訳するならば，

> 「エージェント A は自分が汚れていると知らないし，自分が汚れていないとも知らない」

となる．この告知の効果を考えると，「エージェント A が自分が汚れているとも汚れていないとも知らない」という事は，エージェント A のリンクが存在しない，という事に等しい．なぜならば，リンクとはそもそもある可能世界に居ると考えた時に，区別が付かない世界間に結線されるものであるからである．

図 3.5 を参照し，エージェント A のリンクが存在しないノードを探すと，ノード 100 が相当する．故に，「エージェント A が自分が汚れているか汚れていないか分からない」という告知によって，ノード 100 が削除される事になる．これはエージェント B とエージェント C も同様に考える事ができ，この論理式の a をそれぞれ b, c に置き換える事で同様に表現する事が出来る．以下に告知された論理式を示す．

$$\neg\Box_a d_a \wedge \neg\Box_a\neg d_a \text{ and } \neg\Box_b d_b \wedge \neg\Box_b\neg d_b \text{ and } \neg\Box_c d_c \wedge \neg\Box_c\neg d_c$$

そして同様に 2 名のエージェントが自分が汚れているか汚れていないか分からない世界，即ち赤いリンクと緑のリンクが繋がっていない世界を削除する．同様に図 3.5 を参照すると，ノード 010 には赤のリンクが，ノード 001 には緑のリンクが繋がっていない事が分かる．故にこの 3 つのノードとそれに接続されるリンクを削除する．この告知を実行した結果，得られるエージェントの知識状態を図 3.6 に示す．

図 3.6 を参照すると，実世界であるノード 110 に接続されている青と赤のリンクが無くなっている．つまり，エージェント A とエージェント B は先程の告知が実行された事で，自分がどの世界に居るのか把握する事が出来た事が分かる．これは，Muddy Children Puzzle を紹介した際の以下の文言の内容が公開告知論理の上で実行された事になる．

> 「もし自分が泥をつけていなかったとしたら，泥をつけている子供は『自分が汚れている』という事が分かるはずだ．しかし先程の問答では分からないと答えた．故に自分も汚れている．」

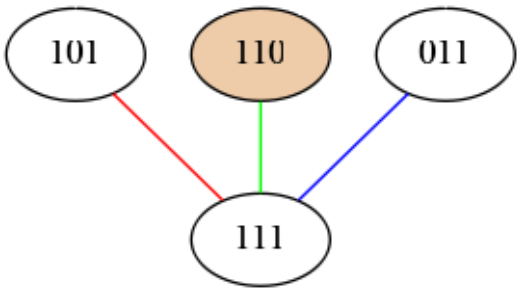

図 3.6: 公開告知論理：1 度目の子供達の回答の告知を実行した後の知識状態

この時点で，エージェント A とエージェント B は自分に泥が付いている事が分かったが，まだエージェント C は泥が付いているかどうか分からない．なぜならば，実世界であるノード 110 には，ノード 111 と繋がる緑のリンクが残っており，まだエージェント C 視点ではどちらの世界にいるのか判別出来ないからである．

二度目の子供達の回答

図 3.6 より，二度目の父親の問いに対する答えは，エージェント A と B は「自分が汚れているか汚れていないか分かった」となり，論理式で表すと $\Box_a d_a \vee \Box_a \neg d_a$ となる．これは

> 「エージェント A はエージェント A が汚れていると知っている，もしくは，
> エージェント A が汚れていないと知っている」

と解釈する事が出来る．そして，エージェント C の答えは先程と同様に $\neg\Box_c d_c \wedge \neg\Box_c \neg d_c$ であるので，これらの 3 つを全て満たす世界のみを残す．以下に実行された告知の論理式を示す．

$$\Box_a d_a \vee \Box_a \neg d_a \text{ and } \Box_b d_b \vee \Box_b \neg d_b \text{ and } \neg\Box_c d_c \wedge \neg\Box_c \neg d_c$$

即ち図上で分かりやすい解釈に表現を直すとすると，「赤と青のリンクが接続されていなく，緑のリンクが接続されている世界のみを残す」と言い換える事が出来る．これを満たす世界を図 3.6 上で探すと，実世界であるノード 110 しか存在しない．なので，2 度目の子供達の回答を告知と見なして実行すると，図 3.7 の知識状態が得られる．

図 3.7: 公開告知論理：2 度目の子供達の回答の告知を実行した後の知識状態

図 3.7 には実世界であるノード 110 しか残っておらず，全エージェントが実世界がノード 110 である事が把握出来た事を示せた．Muddy Children Puzzle はこのように公開告知論理を用いて解く事が出来る．

3.3 エージェント告知論理

エージェント告知論理とは，先述の公開告知論理を基に，エージェントが他のエージェントに対して嘘やブラフを告知する事を定義する為に Hans [12] によって考案された様相論理系で，エージェントが他のエージェントに対して嘘の告知を行う事や，自分自身で真偽のついていない事についてブラフの告知を行う，といった告知を定義している．

3.3.1 構文論

エージェント告知論理の構文論を以下に定義する．

$$\mathcal{L}(!_a, \mathsf{i}_a, \mathsf{i}!_a) \ni \varphi ::= p|\neg\varphi|(\varphi \wedge \psi)|B_a\varphi|\,[!_a\varphi]\,\psi|\,[\mathsf{i}_a\varphi]\,\psi|\,[\mathsf{i}!_a\varphi]\,\psi$$

ここで，$[!_a\varphi]\,\psi$は「エージェント a がφ という真実の告知を行った後,ψが成り立つ」と解釈出来る．また，$[\mathsf{i}_a\varphi]\,\psi$は「エージェント a がφ という虚偽の告知を行った後,ψが成り立つ」と解釈出来る．また，$[\mathsf{i}!_a\varphi]\,\psi$は「エージェント a がφ というブラフの告知を行った後,ψが成り立つ」と解釈される．

3.3.2 クリプキ意味論

エージェント告知論理における認識モデル $M = (S, R, V)$ は以下のように定義される．

1. $M, s \models p \iff s \in V(p)$
2. $M, s \models \neg\varphi \iff M, s \not\models \varphi$
3. $M, s \models \varphi \wedge \psi \iff M, s \models \varphi$ かつ $M, s \models \psi$
4. $M, s \models B_a\varphi \iff$ 全ての t において，$M, s \models \varphi$
5. $M, s \models [!_a\varphi]\,\psi \iff M, s \models B_a\varphi$ ならば $M_a^{\varphi}, s \models \psi$
6. $M, s \models [\mathsf{i}_a\varphi]\,\psi \iff M, s \models B_a\neg\varphi$ ならば $M_a^{\varphi}, s \models \psi$
7. $M, s \models [\mathsf{i}!_a\varphi]\,\psi \iff M, s \models \neg(B_a\varphi \vee B_a\neg\varphi)$ ならば $M_a^{\varphi}, s \models \psi$

なお，$M_a^{\varphi} = (S, R', V)$ は M の R を以下のように変更したモデルである．

- $R'_a = R_a$
- $a \neq b$ のとき, $R'_b = R_b \cap (S \times [\![B_a\varphi]\!]_M) = \{(s,t) \in S \times S | sR_bt \& M, t \models B_a\varphi\}$

ここで，公開告知論理の場合とは異なり，M_a^φ においては可能世界のセットを消去する事なく，到達可能性のみを消去する形をとっている．これは，虚偽やブラフの告知によって実世界が削除される事を避ける為にこのような定義になっている．そして，告知の発言者のエージェントと受け手のエージェントとで到達可能性のセットが異なるのは，虚偽やブラフの告知を行ったエージェント自身が，その虚偽やブラフの告知を信じる事を避ける為である．

クリプキモデル

今，2 名のエージェントが以下のクリプキモデルで表される知識を持っていると仮定する．

- 可能世界の集合 $S = \{s_0, s_1\}$
- 到達可能性関係 $R_a = \{(s_0, s_0), (s_1, s_1)\}$
  $R_b = \{(s_0, s_0), (s_0, s_1), (s_1, s_0), (s_1, s_1)\}$
- 付値 $V(p) = \{s_0\}$

これは，エージェント b が p である可能世界と $\neg p$ である可能世界との間にリンクが存在する事から, p であるか $\neg p$ であるかが判断出来ておらず，エージェント a は判断がついている事を表している．この知識状態の時，エージェント a が $\mathrm{i}_a\neg p$ という虚偽の告知を実行した時，エージェント告知論理によってクリプキモデルが変化する例を図 3.8 に示す．なお，画像中の赤色で表される矢印がエージェント a の，青色で表される矢印がエージェント b の到達可能性を示している．

ここで，告知を行ったエージェント a は p の可能世界へのリンクは切れないが，告知を受けたエージェント b は p の可能世界へのリンクが削除されてしまい，嘘の告知を受け入れてしまう．このような告知で，エージェント a の嘘によってエージェント b が騙された，という事が表現されている．

告知による信念変化の公理

以下にエージェント告知論理における信念変化の公理を示す．ここで，マルチエージェントに対応させた **HK** に加えて，次の公理から成る．

1. (A1) $\varphi \to (\psi \to \varphi)$

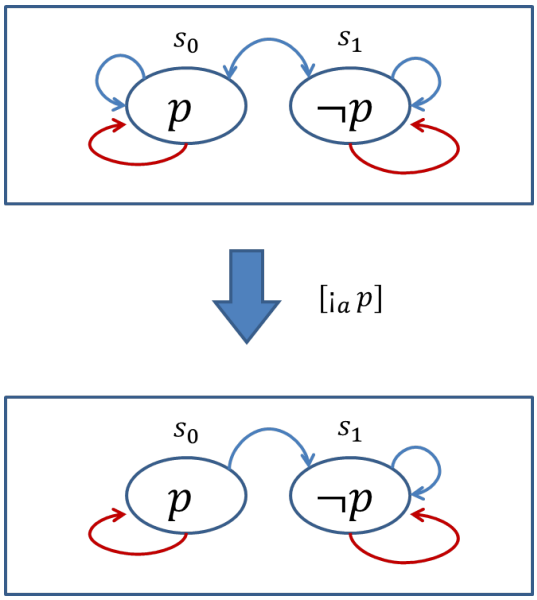

図 3.8: エージェント告知によるクリプキモデルの変化

2. (A2) $(\varphi \to (\psi \to \varsigma)) \to ((\varphi \to \varsigma) \to (\varphi \to \psi))$
3. (A3) $(\neg\varphi \to\!\!\to \psi) \to (\psi \to \varphi)$
4. (K) $\Box_i(\varphi \to \psi) \to (\Box_i\varphi \to \Box_i\psi)$
5. (MP) φ かつ$\varphi \to \psi$から,ψが推論出来る
6. (Nec) φ から，$\Box_i\varphi$ が推論出来る $(i \in I)$
7. $[!_a\varphi]\, B_b\psi \leftrightarrow (B_a\varphi \to B_b\,[!_a\varphi]\,\psi)$
8. $[!_a\varphi]\, B_a\psi \leftrightarrow (B_a\varphi \to B_a\,[!_a\varphi]\,\psi)$
9. $[!_a\varphi]\, B_b\psi \leftrightarrow (B_a\neg\varphi \to B_b\,[!_a\varphi]\,\psi)$
10. $[!_a\varphi]\, B_a\psi \leftrightarrow (B_a\neg\varphi \to B_a\,[!_a\varphi]\,\psi)$
11. $[!!_a\varphi]\, B_b\psi \leftrightarrow (\neg(B_a\varphi \vee B_a\neg\varphi) \to B_b\,[!_a\varphi]\,\psi)$
12. $[!!_a\varphi]\, B_a\psi \leftrightarrow (\neg(B_a\varphi \vee B_a\neg\varphi) \to B_a\,[!!_a\varphi]\,\psi)$

ここで定義されるように，φ について真実の告知を行う際は，発言者は必ず φ である事を知っているし，逆に φ についての虚偽の告知を行う際は $\neg\varphi$ であると知っている必要がある．また，φ であるとも $\neg\varphi$ であるとも知らない場合に，φ であると告知する事をブラフ告知と定義している．

第4章　知識状態観測システム

4.1　設計

本研究の最終的な目標は「汝は人狼なりや？」のプレイログを入力として用い，各プレイヤーをエージェントとして置き換えて考える事によって，発言によってその知識状態の遷移を観測する事であるが，直接人狼のプレイログを用いるには複雑過ぎる為，まずは論理パズルとして有名かつ人狼より知識状態を簡潔化して考える事が出来る Muddy Children Puzzle をエージェント告知論理を用いて解く機構を考えた．子供達をエージェントとして考え，それぞれの回答を公開告知と見なす事で，エージェント告知論理を用いて子供達の知識状態の遷移が観測出来ると考えた．

通常，Muddy Children Puzzle においては，子供達は父親の問いに対して常に正直に答えるが，本研究の目標でもある人狼では，非協力的な人狼役エージェントは嘘をつき，嘘をつかれた人間役エージェントはそれを見抜く，あるいはその嘘を見抜けずに受け入れてしまう，といった知識状態の遷移が発生する為，子供達が意図的に嘘をついた場合はどうなるかを整理する仕組みを用意する．また同様に，Muddy Children Puzzle では父親の問いに対して一斉に子供達が回答を行うが，通常の対話や，人狼が行われるチャットや掲示板を考えると，一斉に回答を行うという事は非現実的であり，通常ならば発言に前後関係が生じ，それにより知識状態の遷移の仕方が異なる事が一般的であると思われる為，子供達が一人づつ回答を行う場合はどうなるかを考える．

4.1.1　矛盾を信じる事無く嘘をつく事が出来るエージェント

嘘をつくエージェントを考える時，エージェント告知論理を用いれば，実現は容易である．ただし，エージェント告知論理においては，エージェント自身の知識と他のエージェントの告知とに矛盾が発生した場合は，到達可能世界のセットが失われ，そのエージェントは「何も信じていない」という状況が発生しうる．しかし，通常のゲームにおいて,「何も信じていない」という状況は現実的では無い為，自分の知識と告知とを照らし合わせ，確固たる自分の知識と告知内容とが矛盾した場合は告知を棄却する，という形でエージェント告知論理の改変を行う．

改変したエージェント告知論理の構文論を以下に定義する．

$$\mathcal{L}(!_a, \mathsf{i}_a) \ni \varphi ::= p|\neg\varphi|(\varphi \wedge \psi)|B_a\varphi|\,[\varphi!_a]\,\psi|\,[\varphi\mathsf{i}_a]\,\psi$$

ここで，通常のエージェント告知論理と区別する為に，真実告知を $[\varphi!]\,\psi$，虚偽告知を $[\varphi\mathsf{i}_a]\,\psi$ と表記する．続いて，意味論を再定義する．

1. $M, s \models p \Longleftrightarrow s \in V(p)$
2. $M, s \models \neg\varphi \Longleftrightarrow M, s \not\models \varphi$
3. $M, s \models \varphi \wedge \psi \Longleftrightarrow M, s \models \varphi$ かつ $M, s \models \psi$
4. $M, s \models B_a\varphi \Longleftrightarrow$ 全ての t において，$M, s \models \varphi$
5. $M, s \models [\varphi!_a]\,\psi \Longleftrightarrow M, s \models \varphi$ ならば $M^{\varphi!_a}, s \models \psi$
6. $M, s \models [\varphi\mathsf{i}_a]\,\psi \Longleftrightarrow M, s \models B_a\neg\varphi$ ならば $M^{\varphi\mathsf{i}_a}, s \models \psi$

ここで，$M^{\varphi!_a} = (S, R', V)$ は以下のように定める．

$$R'_b = \begin{cases} R_b & (S, (R_c)_{c\in A\setminus\{b\}}, R_b \cap (S \times [\![\varphi]\!]_M), V), s \models B_b\bot\text{の場合} \\ R_b \cap (S \times [\![\varphi]\!]_M) & \text{それ以外の場合} \end{cases}$$

また，$M^{\varphi\mathsf{i}} = (S, R', V)$ は以下のように定める．

- $R'_a = R_a$
- $a \neq b$ の時，

$$R'_b = \begin{cases} R_b & (S, (R_c)_{c\in G\setminus\{b\}}, R_b \cap (S \times [\![\varphi]\!]_M), V), s \models B_b\bot\text{の場合} \\ R_b \cap (S \times [\![\varphi]\!]_M) & \text{それ以外の場合} \end{cases}$$

なお，G は全エージェントの集合とする．

再定義を行った点は他エージェントに対して告知を行った時である．従来のエージェント告知論理では発言者が告知を実行すると，聞き手はそれを単純に受け入れるのみであったが，本研究においては，聞き手の知識を考慮に入れて知識状態の更新を行う．例えば，φ が真であるとエージェント A と B が知っているとする．その時，エージェント A が $\neg\varphi$ と告知を行っても，エージェント B はその告知を受け入れない．なぜならば，エージェント B は自分自身で φ という知識を保持しているので，それと矛盾する告知を受け入れる事はしないからである．

また，ある嘘の告知を受け入れてしまい，騙されたエージェントが，本人にとっては真である知識を告知する際にも，同様の対応が必要と考えた為，虚偽の告知を行う際だけでなく，真実の告知を行う際の意味論についても再定義を行った．これによって，嘘に対して騙されたエージェントがその嘘から得た知識を告知を行った場合に，最初に騙されな

かったエージェントがその発言者が嘘によって騙されて行った告知も，知識と矛盾していればそれも受け取らなくなる．

Muddy Children Puzzle の例を挙げるとするならば，子供 A が汚れているか否かは，子供 B と C は常に知っている．子供 A が汚れている時，(本来はこのような発言は行わないが) 子供 C が「子供 A は汚れていない」などと発言した場合などは，子供 A はそれが定かではない場合は，その告知を受け入れるが，子供 B は自分の持っている情報と矛盾が発生する為，自分の知識状態を更新しない．これによって，発言者が嘘をついているか否かを聞き手が自身の知識ベースで判別する事が出来る．

また，エージェント告知論理において，告知の発言者が真偽の付かない事に対して真である，あるいは偽であると告知する事をブラフ告知 (Bluffing Annoucement) と定義している．しかし今回は，エージェントらの告知内容は「自分の額に泥が付いているかどうか分かる」,「自分の額に泥が付いているかどうか分からない」の二つのみであり，仮に「分かる」状態ならば，嘘をつこうとすると「分からない」になり,「分からない」状態ならば嘘をつこうとすると「分かる」になる為，このルールではブラフ告知を定義する事が出来ない為，今回の実装からは除いた．

4.1.2 発言順を考慮した告知

通常の Muddy Children Puzzle では一斉に行っていた回答を，子供一人ずつ順番に発言（告知）を行うよう Muddy Children Puzzle を定義し直す事で，より現実的な対話における知識状態の遷移を実現する．ここで注意せねばならない点は，父親の問いに対して，子供達はそれぞれのタイミングで分かっているか否かを回答する，という点である．本来の例では，最初の父親の問いに回答する際，子供達が持ちうる情報は「子供達の内，一人以上が汚れている」しか存在しないが，例えば子供達が A，B，C の順で回答していくと仮定した時，子供 B が回答する時は，父親の発言と，それに対する子供 A の回答を持ち合わせた上での回答になるという事である．発言の間隔が極めて短い場合はこれによらない場合も考えられるが，現実的に考えた際には，誰かの発言を聞いた瞬間にそれぞれのエージェントの知識状態が更新されるはずと考え，このように Muddy Children Puzzle を再定義した．

4.2 実装

4.2.1 実装環境

実装環境は以下の通りである．

- OS：Windows7 Professional 64bit
- 実行環境：XAMPP 1.8.1

- 使用言語：PHP 5.4.7, JavaScript

論理式を取り扱うという点では，当研究室でも研究が豊富である Prolog を使う事が望ましいと考えられたが，筆者のプログラミング経験より，PHP を用いて Web アプリケーションという形で実装を行った方が実装コストの軽減に繋がると考え，PHP で実装を行った．また，知識状態のクリプキモデルを描画する為に PEAR ライブラリの Image_GraphViz を導入している．

4.2.2 実装するエージェント

本研究で提案する機構上に実装するエージェントは，Muddy Children Puzzle に登場する額に泥をつけているかも知れない 3 名の子供達である．子供達をそれぞれ A，B，C と定義し，子供達はそれぞれ知識と行動原理を持つ．

エージェントが持ちうる知識

エージェントは自分自身が泥を付けているか否かは判別出来ないが，他エージェントが額に泥を付けているか否かは判別出来る．また，それは覆る事のない絶対的な情報であると仮定する．例えば，子供 A は自分の額に泥が付いているか否かを知ることは出来ないが，子供 B と子供 C が額に泥を付けているかどうかはエージェント自身が観察する事から知る事が出来る．また,「子供 A の額に泥が付いている」という情報を d_a と定義する．

エージェントの行動原理

本研究で実装を行ったシステムにおいては，エージェントは常に嘘をつく事ができると仮定している．エージェントが告知出来る情報は「自分が額に泥を付けているかどうか分かる」か「自分が額に泥を付けているかどうか分からない」の 2 点のみである．

ここで，実際に告知される内容を考えると,「子供 A が額に泥をつけているかどうか分かる」は,「子供 A の額に泥が付いていると知っている」もしくは「子供 A の額に泥がついていないと知っている」と書き換える事が出来る為，論理式で表すと，$\Box_a d_a \vee \Box_a \neg d_a$ となる．

同様に,「子供 A が額に泥を付けているかどうか分からない」とは,「子供 A の額に泥がついていると知らない」かつ「子供 A の額に泥がついていないと知らない」と書き換えられる事から，論理式では $\neg\Box_a d_a \wedge \neg\Box_a \neg d_a$ と表す事が出来る．そして，エージェントは自分が持ちうる知識と告知内容が矛盾しない場合のみ，他者の告知を受け入れる事で，自身の知識状態を更新する．

実装したシステム

図 4.1 に実装したシステムのスクリーンショットを示す．画面左側のクリプキモデルの図が告知を行う前の画像，右側のクリプキモデルの図が告知を行った後の画像である．現在は告知を何も行っていない状態なので，単純にノードだけが並んでいる画像と全エージェントの知識状態のリンクを張った状態の画像が並んでいる．

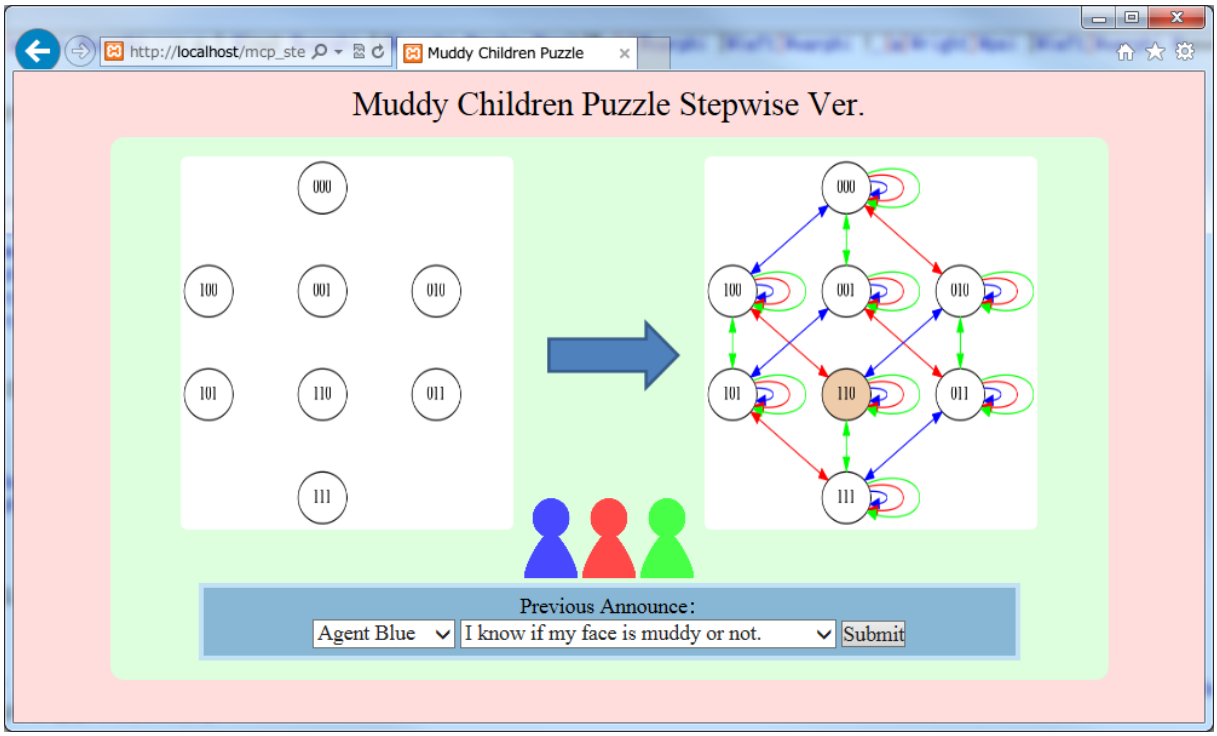

図 4.1: 実装したシステムのスクリーンショット

画像下部にはセレクトボックスが二つあり，そこで「どのエージェントが」「(自分の顔が汚れているかどうか) 分かる／分からない」という告知を行うかを選択し，送信ボタンを押すと，画面右側のクリプキモデルに対して告知を行って左右のクリプキモデルの図を更新する．告知を実施するエージェントを選択する操作を図 4.2，実施する告知の内容を選択する操作を図 4.3 に示す．

そして，実際に告知を実施した際の挙動を図 4.4 に示す．この例であれば，初期状態から父親が「君達の内，一人以上が顔に泥を付けているよ」という内容の $d_a \vee d_b \vee d_c$ という告知を実施した事になる．なお，父親の告知内容だけは「(自分の顔が汚れているかどうか) 分かる／分からない」という内容ではなく，先述の論理式をシステム上で固定して告知する指定になっている．

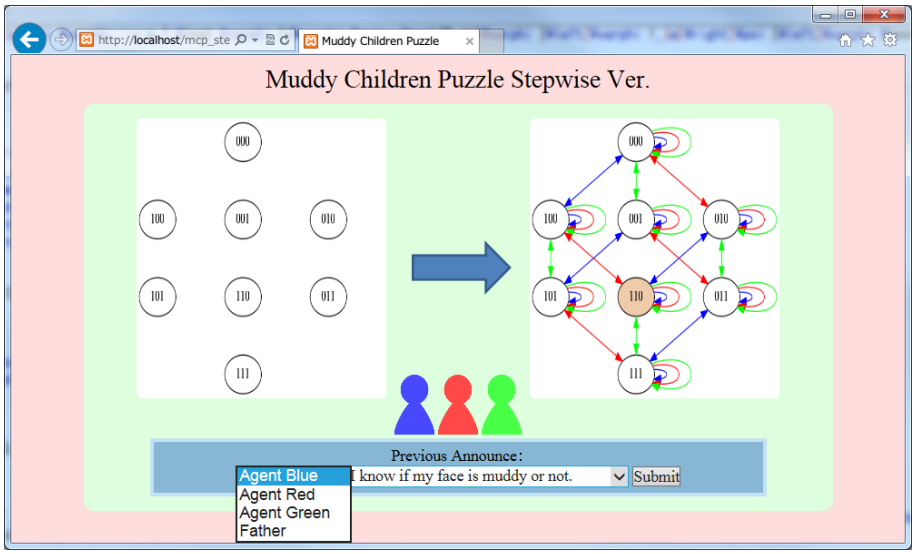

図 4.2: 告知を実施するエージェントを選択

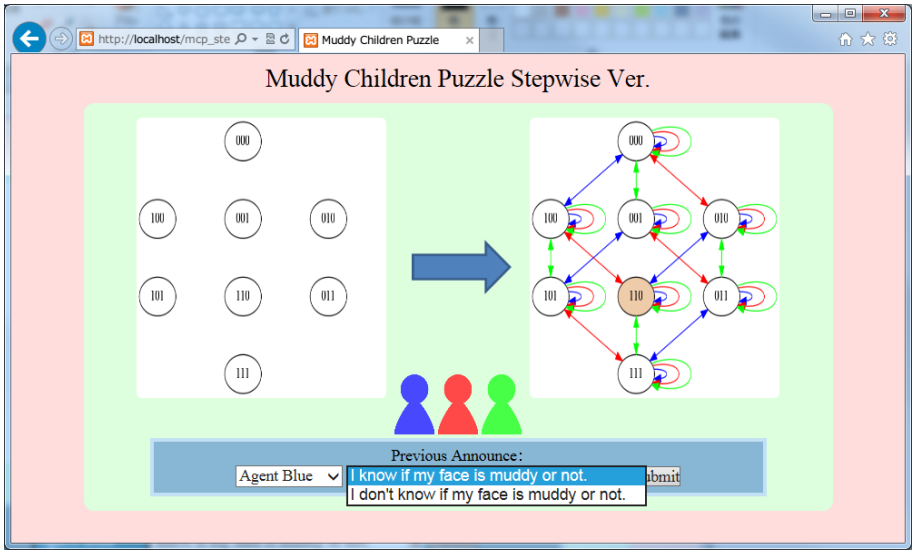

図 4.3: 実施する告知の内容を選択

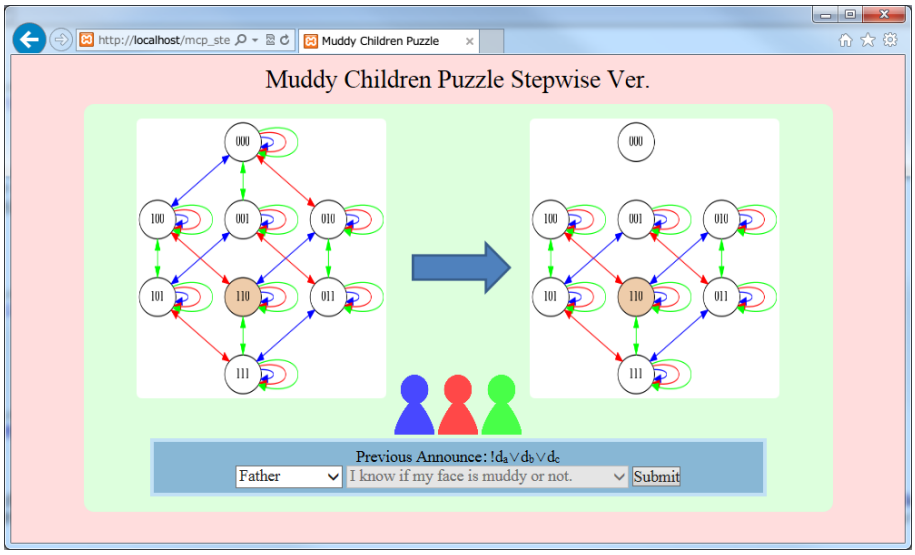

図 4.4: 初期状態に対して父親の告知を実施した状態

アーキテクチャ

図 4.5 に実装したシステムのアーキテクチャを示す．システムの処理される手順としては以下の通り．

1. ユーザがシステムへブラウザを介しアクセス
2. 初回アクセス時に実世界とノードからリンクの集合を生成
3. セレクトボックスを使用して告知するエージェントと内容を選択
4. 告知内容が実世界で真か偽か判定
5. 告知を各エージェント毎に実行
6. もし告知により実世界から出るリンクが 1 本も無くなったら告知を棄却してロールバック
7. エージェントの数だけ 6-7 を実行
8. 8 が終了したら，その時点でのクリプキモデルを.dot ファイルに書き出し
9. 書きだした.dot ファイルを PEAR ライブラリを用いて png 形式に書き出し
10. png ファイルを JavaScript を用いてバックグラウンドで取得，ブラウザ上に描画
11. 3 へ戻る

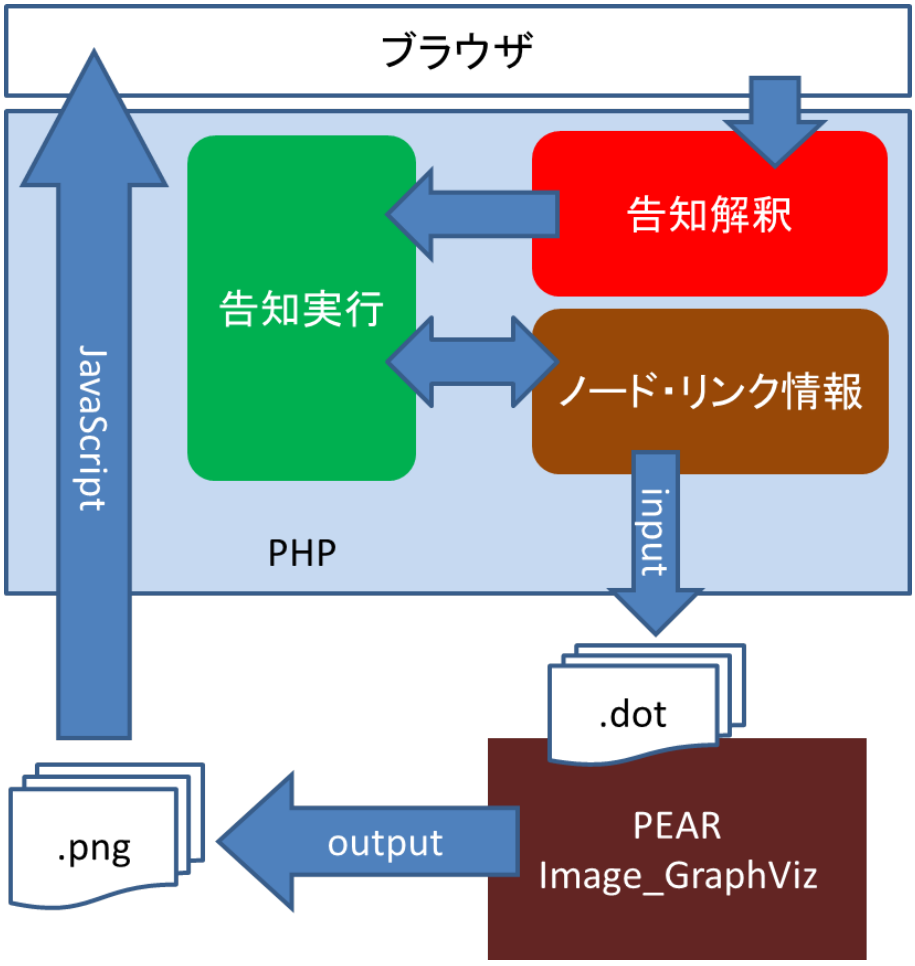

図 4.5: システムアーキテクチャ

第5章　知識状態観測システムの検証

提案した機構で実施した検証は，以下の条件の下で行っている．

1. 登場するエージェントは3名で，それぞれリンクの色はA（青）B（赤）C（緑）に対応する
2. ノード内の数字は，左から順にA，B，Cが汚れている(=1)か，汚れていない(=0)かを表す
3. エージェントAが汚れている事をd_a，汚れていない事を$\neg d_a$と表す．下付き文字はそれぞれエージェントA，B，Cに対応する
4. エージェントA，B，Cがあるノードからあるノードに到達可能性がある事をそれぞれ青赤緑の矢印で表す
5. 嘘やブラフをつくのはエージェントAのみで，BとCは常に真実のみを告知する
6. 最初に父親が告知を行った後，エージェントの発言順はA，B，Cの順で行われる

5.1　真実告知

まずは全てのエージェントが常に真実の告知しか行わないケースを考える．

初期状態

今設定してある実世界は$d_a \wedge d_b \wedge \neg d_c$，つまりノード110である．3名が一斉に告知を行うケースに関しては公開告知論理の節において解説を行い，全エージェントが実世界がノード110である事が理解出来ていた．まずはそれについて，同様の実世界を設定した上で，告知を段階的に行うと結果がどう変化するかを確認する．

図5.1に初期状態を示す．状態は全員汚れていない(=000)から全員汚れている(=111)までの計8通りある．今，実世界は$d_a \wedge d_b \wedge \neg d_c$に設定した為，ノードで言えば110に相当する．

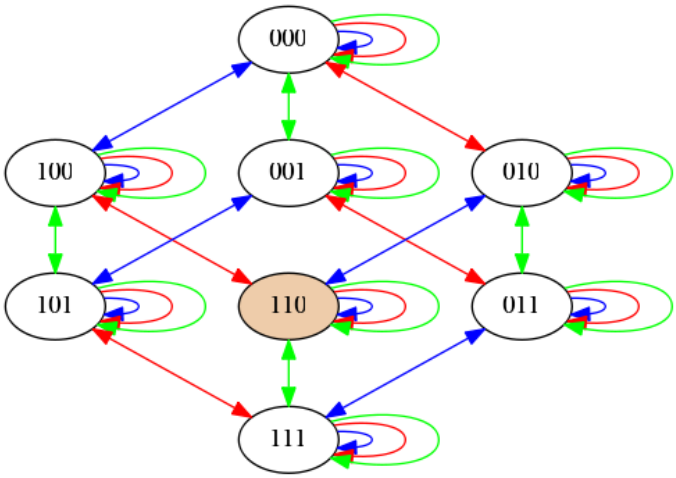

図 5.1: 初期状態

父親の告知

図 5.1 の状態から父親が告知を行った状態を図 5.2 に示す．父親が行った告知の内容は「君達の内，一人以上が汚れている」というものであるので，形式化すると $d_a \vee d_b \vee d_c$ となる．この状態の内，父親の告知内容を満たさないのはノード 000 である為，各エージェントが持つノード 000 へ繋がるリンクが削除される．

Muddy Chidlren Puzzle における父親は，エージェント達を外部から観測している部外者のような存在であり，父親の発言は常に絶対的で真実である，という前提が存在する．即ち，エージェントがエージェントに対して行うエージェント告知とは異なり，一般的な公開告知に相当する為，図に現れる動作としては告知内容に反する世界からの到達可能性と，告知内容に反する世界への到達可能性との両方を削除される事になる．

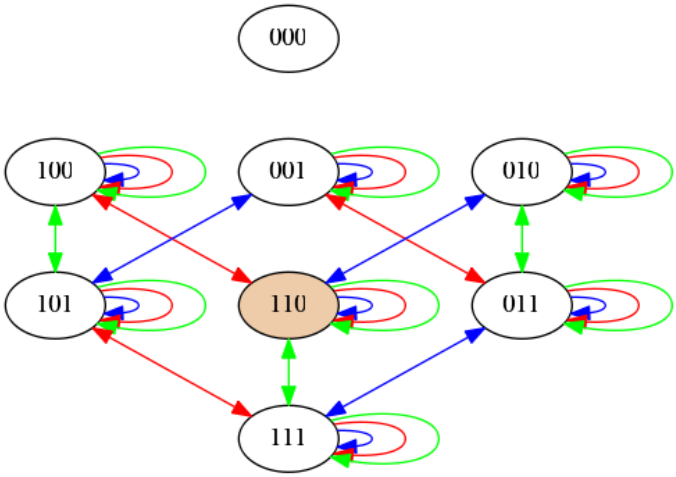

図 5.2: 父親が最初の告知を行った状態

エージェント A の告知

続いて，エージェント A が発言する順番である．図 5.2 を見ると，エージェント A にとって，実世界であるノード 110 には，反射律から存在するノード 110 へのリンクと，ノード 010 へのリンクが存在する．Muddy Children Puzzle においては，自分自身が汚れているかどうかは分からない為，今エージェント A が持ち合わせている情報は「$d_a \vee d_b \vee d_c$」と「d_b」と「$\neg d_c$」だけである．これだけの情報ではエージェント A は自分がノード 110 にいるのか，それともノード 010 にいるのかは判別出来ない．

そこでエージェント A の告知内容は「自分が汚れているかどうか分からない」となる．これを形式化すると $\neg\Box_a d_a \wedge \neg\Box_a \neg d_a$ である．これは,「エージェント A が汚れていると知らない，かつ，エージェント A が汚れていないと知らない」と解釈する事が出来る．これをエージェント告知を行う為に告知演算子を追加して表現すると，$!_a \neg\Box\, d_a \wedge \neg\Box\neg d_a$ となる．これは,「エージェント A は，エージェント A が汚れていると知らない，かつ，エージェント A が汚れていないと知らない」と解釈する事が出来る．

エージェント A の正直な告知 $!_a(\neg\Box_a d_a \wedge \neg\Box_a \neg d_a)$ が実行されると，エージェント達は $\neg\Box_a d_a \wedge \neg\Box_a \neg d_a$ を満たしていない世界へ向かうリンクを切断する．即ち，エージェント A が一つの世界しか見えていない世界へのリンクを切断する事になる．図 5.2 を参照すると，ノード 100 だけがエージェント A が一つの世界しか見えていない事になる．もし，エージェント達がノード 100 に居たとするならば，エージェント A はノード 100 においては反射律から来るノード 100 へ向かうリンクしか存在しない．即ち，ノード 100 が実世界であるとするならば，エージェント A は自分の額に泥が付いているか否か判別出来るはずである．しかし，実際の告知では「自分が汚れているかどうか分からない」だった為，ノード 100 へ向かうリンクを全てのエージェントが切断した．

この告知が実行された結果を図 5.3 に示す．

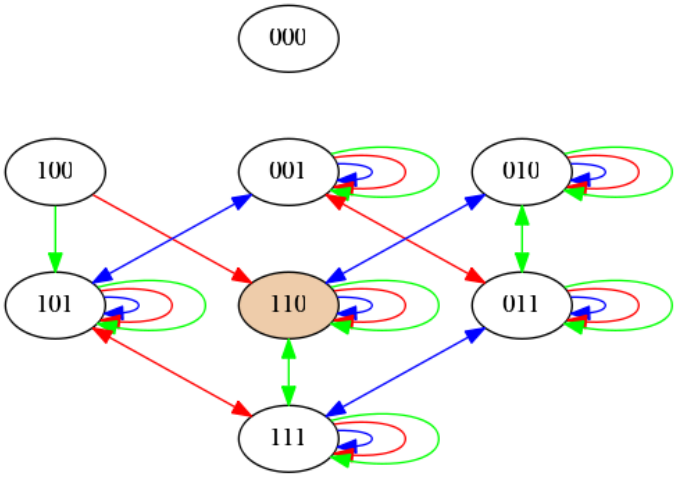

図 5.3: エージェント A が正直に告知を行った後の状態

エージェント B の告知

次はエージェント B が発言する順番である．図 5.3 を参照すると，エージェント B は実世界であるノード 110 においては，反射律から来るノード 110 へのリンクしか存在しない．これは，先程のエージェント A の告知によって，ノード 110 からノード 100 へ向かうリンクが切断された為である．エージェント B は元々「d_a」と「$\neg d_c$」である事は観察から既に知っている為，ノード 100 とノード 110 のどちらかに居るはずだという事は分かっていた．それが先程のエージェント A の告知によってノード 100 である可能性が削除された為，エージェント B は実世界がノード 110 である事を知る事が出来た．

なお，この時点ではエージェント A とエージェント C は，実世界であるノード 110 において，リンクの行き先が複数存在する為，まだ実世界がどこであるかは判別出来ていない．故にエージェント B の告知内容は正直に「自分が汚れているかどうか分かる」と告知する．これを形式化すると $\Box_b d_b \vee \Box_b \neg d_b$ となり，意味としては「エージェント B が汚れている事を知っている，もしくは，エージェント B が汚れていない事を知っている」となる．これに告知演算子を追加すると $!_b(\Box_b d_b \vee \Box_b \neg d_b)$ となる．

エージェント B の正直な告知 $!_b(\Box_b d_b \vee \Box_b \neg d_b)$ が実行されると，エージェント達は $\Box_b d_b \vee \Box_b \neg d_b$ を満たしていない世界へのリンクを切断する．即ち，エージェント B が複数の世界が見える世界へのリンクを切断する事になる．図 5.3 を参照すると，$\Box_b d_b \vee \Box_b \neg d_b$ を満たすノードは，ノード 100 とノード 010，そしてノード 110 の三つである．後者の二つの世界のエージェント B のリンク，即ち赤の矢印を参照すると，反射律から来るそれぞれのノードから自身のノードへ向かうリンクしか存在しない．また，ノード 100 はノード 110 へ向かうリンクしか存在しない．エージェント B がこの三つの世界に居ると仮定するならば，エージェント B が $!_b \Box_b d_b \vee \Box_b \neg d_b$ という告知が出来る．なので，エージェント達はこの三つのノードへ向かうリンク以外を削除する．だが，ノード 100 に向かうリンクはもとより誰も持っていないので，実際にはノード 010 とノード 110 へ向かうリンク以外を削除する事になる．この告知が実行された結果，リンクの切断が行われた結果を図 5.4 に示す．

図 5.4 を参照すると，ノード 010 とノード 110 へ向かうリンク以外が全て切断されている事が確認出来る．ここで実世界におけるエージェント C を確認すると，エージェント C は実世界であるノード 110 において，反射律から来るリンク以外は存在しない．即ち，エージェント C も実世界がノード 110 であるという事が理解出来たのである．これでエージェント C とエージェント B が実世界がどこであるか判別出来た事になる．

エージェント C の告知以降

これ以上は告知を繰り返しても，これ以上に知識状態が収束する事は無く，エージェント A はいつまでたっても実世界がどこであるか判別がつかない．なぜならば，次はエージェント C が発言する順番であるが，エージェント C が発言する内容は「自分が汚れているかどうか分かる」であり，形式化すると $\Box_c d_c \vee \Box_c \neg d_c$ である．この告知によって，

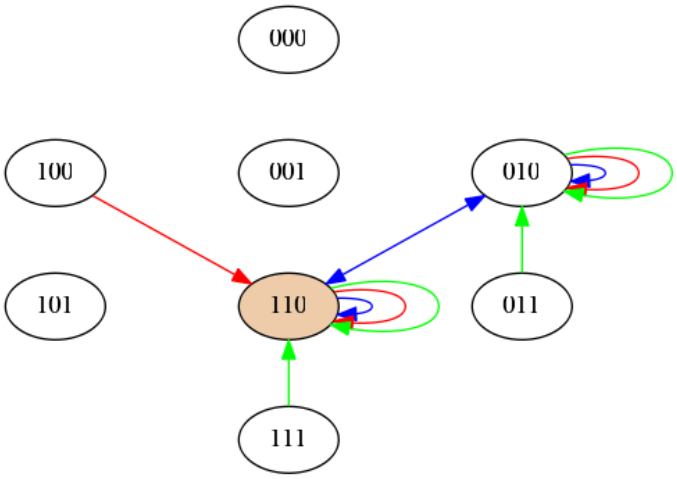

図 5.4: エージェント B が正直に告知を行った後の状態

エージェント C の到達可能性を示す緑の矢印が 1 本しか出ていない世界へのリンクを残す，すなわち，緑の矢印が複数本出ている世界へ向かうリンクを切断するのだが，全てのノードにおいて緑の矢印が出ているのは 0 本ないしは 1 本である．故に知識状態に変化は生じない．

同様に次のエージェント A は未だに実世界であるノード 110 からはノード 110 へ向かうリンクとノード 010 へ向かうリンクと 2 本存在しており，実世界がどちらであるか判別が付かない．故に告知内容は $\neg\Box_a d_a \wedge \neg\Box_a \neg d_a$ である．これは全エージェントにとって既知の情報である為，知識状態の更新は行われない．次のエージェント B の告知も同様である．

故に，全エージェントが常に正直な告知を実行すると，少なくとも実世界を $d_a \wedge d_b \wedge \neg d_c$ に設定し，エージェントの発言順を A，B，C と設定したこの例においては，エージェント A はいつまでも実世界が定まる事がなく，エージェント B とエージェント C はすぐに実世界がどこであるかを理解する事が出来た．

段階的な告知の効果

通常の公開告知論理において 3 名のエージェントが一斉に告知を行った場合は，全エージェントが正直に告知を行うと，全エージェントの知識状態が一意に実世界に定まる事が確認出来ていたが，全エージェントが正直に，しかし段階的に告知を行う事によって，このケースであればエージェント A が何度告知を続けても実世界がどこであるのか理解する事が出来なかった．これは，エージェント A にとって，エージェント B の「汚れているかどうか分かった」という告知が，父親の告知の時点で理解出来ていたのか，エージェント A の告知も聞いた上で理解出来たのかが判別出来ないから発生した現象であると考えられる．

5.2　虚偽告知（1）

先ほどの例ではエージェント A はいくら告知を繰り返しても知識状態が一意に定まらずに，自分の額に泥が付いているか否かを知る事が出来なかった．そこで，同様の実世界を設定した上で，エージェント A が虚偽の告知を実行する事で，得られる知識状態がいかに変化するかを確認する．

初期状態と父親の告知

図 5.5 に初期状態を示す．先程の全エージェントが正直な場合から実世界を変更していない為，図 5.1 と同様の初期状態になり，変更点はない．また，図 5.6 は図 5.5 で示される知識状態に対して父親が $d_a \vee d_b \vee d_c$ の告知を行った後の知識状態であるが，こちらも同様に先程の正直なエージェントの場合と同様で，図 5.2 と同じ知識状態になる．

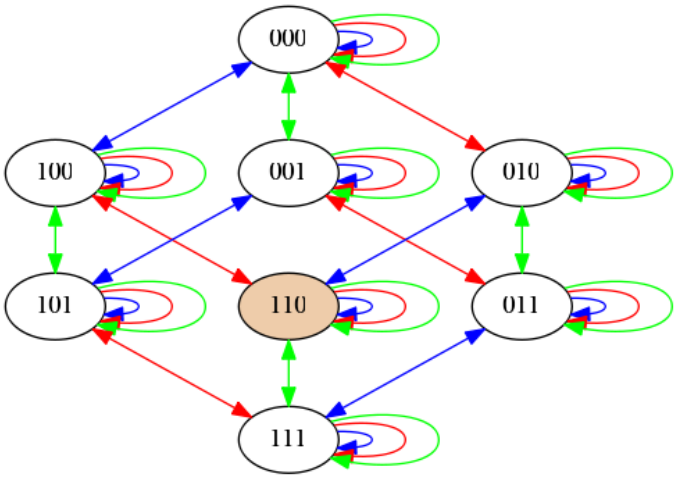

図 5.5: 初期状態

エージェント A の虚偽告知

ここでも先程の正直なエージェントだけで構成される知識状態と同様，エージェント A はまだ自分が汚れているかどうか分からない．しかし，先程の実験結果からエージェント A はいつまでも自分の額に泥が付いているか否かを把握する事は出来ない．

そこで「自分の額が汚れているかどうか分かった」と嘘をつく．これを形式化すると $\Box_a d_a \vee \Box_a \neg d_a$ となり，この式は「エージェント A が汚れていると知っている，もしくはエージェント A が汚れていないと知っている」と解釈される．これをエージェント告知論理を用いて嘘をつくと形式化すると，$\mathrm{i}_a(\Box_a d_a \vee \Box_a \neg d_a)$ となり，「エージェント A は，エージェント A が汚れていると知っている，もしくはエージェント A が汚れていないと知っている，と虚偽の告知を行った」と解釈される．

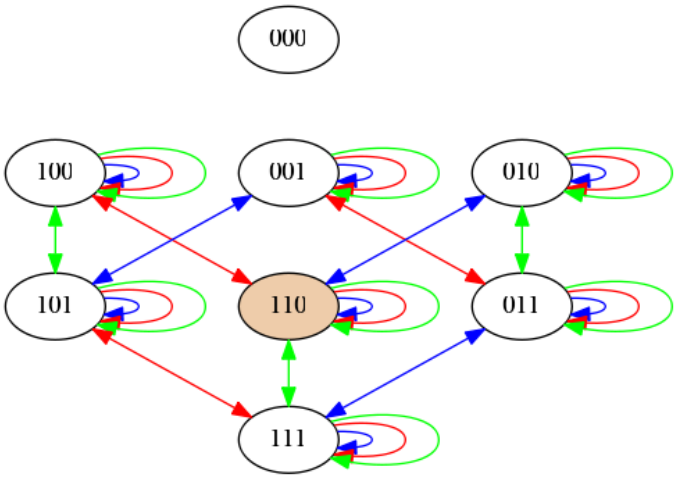

図 5.6: 父親が告知を行った後の状態

この告知をまずは意味論を変更しない状態で実行した結果を図 5.7 に示す．

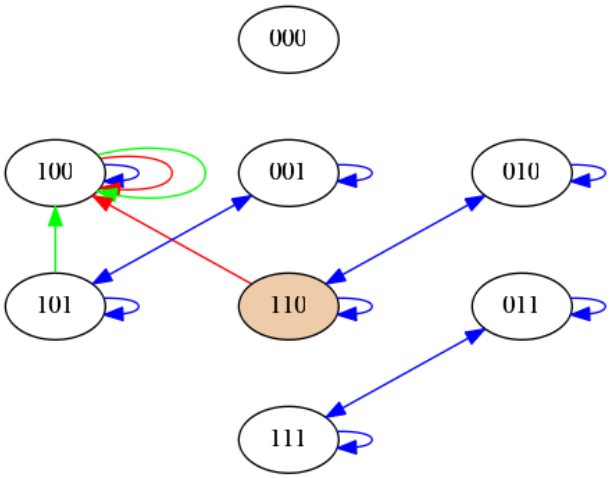

図 5.7: エージェント A が虚偽の告知を行った後の状態（1）

図 5.6 を参照すると，$\Box_a d_a \vee \Box_a \neg d_a$ を満たす世界はノード 100 しかない．それ以外のノードは，エージェント A のリンク，即ちエージェント A の到達可能性を示す青の矢印が 2 本ないしは 0 本存在する．なので，ノード 100 へ向かうリンク以外を全て削除する．すると図 5.7 の知識状態になる．すると，発言者であるエージェント A のリンクには変更がない．それは発言者が発言の内容を嘘と自覚している為である．自分で発言した嘘を信じる，というのは極めて現実的ではない為，虚偽告知は発言者の知識状態に一切の影響を与えない．

そして，発言を受け取ったエージェント B とエージェント C はノード 100 へ向かうリンク以外が切断されたが，注目すべき点はエージェント C のリンクである．エージェント C は実世界であるノード 110 において，d_a と d_b を知識として持ちあわせていたが，エー

ジェント A の虚偽である告知を正直に受け取った場合，実世界であるノード 110 からの全てのリンクが消滅してしまう．つまり，自分が持っている知識とエージェント A の告知とが矛盾してしまっている．このままの状態では，エージェント C は「何も信じていない」状態になってしまい，一種の思考停止状態のような状況に陥ってしまう．

元々のエージェント告知論理においては，この状態に陥る事で「エージェント A が嘘をついた」という事が判明するだけで終了してしまうのだが，本研究においては，告知と知識との間に矛盾が発生した場合は，告知内容を棄却する処理を行った．そうする事で，嘘の告知を受けたエージェントは，その告知の発言者の嘘を見抜いた上で，自分の知識状態に矛盾を起こさせない事が可能になる．ここで，エージェント B は実世界であるノード 110 からノード 100 へ到達可能性のリンクが残っている為，エージェント B の中では知識に矛盾が発生しない為，この嘘の告知を受け入れてしまう．

なぜならば，エージェント B は元々の知識として d_a と $\neg d_c$ は持っており，実世界からはノード 110 とノード 100 へのリンクが存在していた．そこで，エージェント A の嘘の告知により，ノード 110 からノード 110 へ向かうリンクが切断され，実世界ノード 110 からはノード 100 への到達可能性しか存在しない為，ノード 110 に居ながらもノード 100 へ居ると錯覚してしまう．即ち，エージェント A は嘘の告知を行う事で，エージェント B をだます事に成功したが，エージェント C にはその嘘を見抜かれてしまった事になる．図 5.7 の状態からエージェント C が嘘を見抜いた事でエージェント C のリンクだけをロールバックした知識状態を図 5.8 に示す．

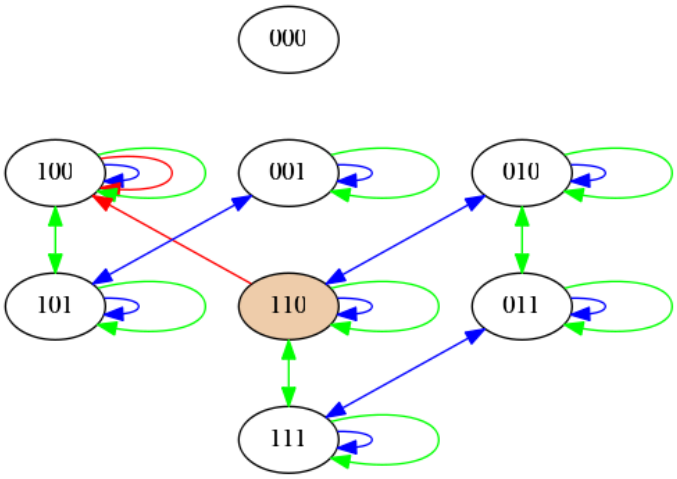

図 5.8: エージェント A が虚偽の告知を行った後の状態（2）

エージェント A の虚偽である告知の後の各エージェントの知識状態は図 5.8 の通りになる．エージェント A は自分の嘘を信じる事は無い為，知識状態のリンクは変更しないし，エージェント C も自分が持っている知識と告知とが矛盾する為，告知を嘘と断定し棄却する．しかし，エージェント B にとっては元々の自分が持っている知識と矛盾しない告知だったので，エージェント A の嘘を信じてしまう．

エージェント B の告知

続いてはエージェント B の発言する順番になるが，エージェント B にとって実世界であるノード 110 からは到達可能性が一つしか存在しない，即ち自分の額に泥が付いているかどうか判別出来ている事になる．しかし，前述の通りそれはエージェント A の嘘によるものである．それでも，エージェント B にとっては判別出来ている事になる．なので，エージェント B の告知内容は「自分の額に泥が付いているかどうか分かる」になり，形式化すると $\Box_b d_b \vee \Box_b \neg d_b$ となる．発言者であるエージェント B にとっては真実なので，告知演算子を追加すると，$!_b(\Box_b d_b \vee \Box_b \neg d_b)$ となる．これまでと同様に，解釈としては「エージェント B は，エージェント B が汚れていると知っている，もしくはエージェント B が汚れていないと知っている」となる．

この告知を実行した結果を 5.9 に示す．

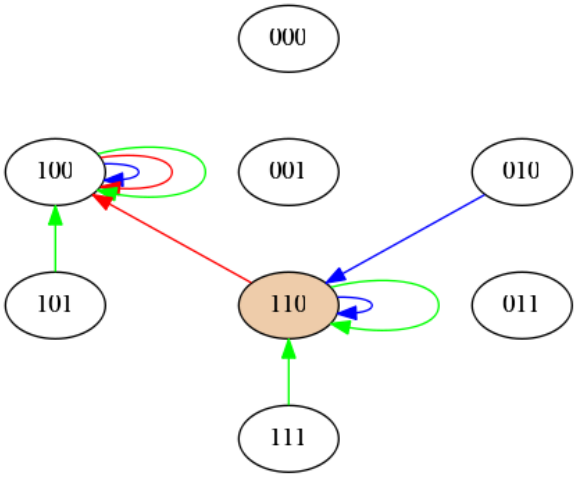

図 5.9: エージェント B が正直に告知を行った後の状態

エージェント B はエージェント A による嘘の告知をそのまま真実であるとして受け入れた状態で「額に泥が付いているかどうか分かる」と（少なくともエージェント B 本人にとっては）真実の告知を行った．すると，告知を受け取ったエージェントは，エージェント B の到達可能性が一つしか存在しない世界へのリンクのみを残して他のリンクを切断する．エージェント B の到達可能性が一つしか存在しない世界とは，この状況ではノード 100 とノード 110 の 2 つになる．それぞれ，ノード 100 はノード 100 へ，ノード 110 はノード 100 へのリンクしか存在しない．故に，告知を受け取ったエージェント A とエージェント C はこの 2 つのノードへ向かうリンク以外を全て切断する事になる．

すると，実世界であるノード 110 を参照するとエージェント A とエージェント C は反射律のノード 110 からノード 110 へのリンクだけが存在し，エージェント B はノード 110 からノード 100 へのリンクしか存在しない．これで全エージェントが自分が汚れているかどうか分かった事になる．

だが，エージェント B だけは自分が汚れていないと信じているが実際には汚れており，エージェント A によって「騙された」状態になっている．エージェント C も同様の嘘の告知を受け取ったものの，自分の知識と矛盾する，という点から嘘の告知を見抜きつつ，自分が汚れているか否かを正しく把握する事が出来ている．

虚偽の告知の効果

エージェント A は前項で常に正直な告知を行った場合，何度告知を実行しても知識状態が一意に定まる事はなかった．しかし，今回のケースにおいては，敢えて嘘をつく事によって，少なくとも 1 名のエージェントを騙した上で自分が汚れているか否かを正確に理解する事が出来た．言い換えれば，嘘をつくことで本来知り得なかった情報を知る事が出来たと言える．

今回のケースではエージェント C には嘘がバレてしまっていたが，人狼のケースに置き換えて考えると，例えばエージェント A とエージェント C が人狼で，エージェント B が人間だと考えた場合に，人狼役のプレイヤーが共謀して人間役のプレイヤーを騙すような事もこのエージェント告知論理の改変形を用いて表現する事が出来ると考える．

5.3 虚偽告知（2）

先ほどの例ではエージェント A が「自分に泥が付いているかどうか分からない」状況において,「自分に泥が付いているかどうか分かった」と虚偽の告知を実行する事によって，本来真実のみを告知した時には得られなかった「自分の額に泥が付いているか否か」を知る事ができ，エージェント B を騙す事も出来た．しかし，エージェント C にはその嘘を見抜かれてしまっていた．そこで，今度は実世界を変更する事で,「自分に泥が付いているかどうか分かっている」状況において,「自分に泥が付いているかどうか分からない」とエージェント A が虚偽の告知を実行する事で，どのような効果が得られるかを確認する．

初期状態と父親の告知

図 5.5 に初期状態を示す．今回の検証では実世界をノード 110 からノード 100 へ変更して実験を行う．初期状態は図 5.10 となる．このように実世界を設定する事で，エージェント A は最初の父親の告知によって,「自分に泥が付いているかどうか分かる」状況になる．

続いて，父親の告知を実行する事で得られる知識状態を図 5.11 に示す．これは図 5.10 で示される知識状態に対して父親がこれまでと同様に，$d_a \vee d_b \vee d_c$ の告知を行った後の知識状態であるが，こちらでは期待した通り，既に実世界であるノード 100 にはエージェント A のリンクは 1 本しかなく，エージェント B とエージェント C のリンクは 2 本存在する為，エージェント A のみが実世界がノード 100 である事が理解出来ており，エージェント B とエージェント C はまだどこが実世界なのか理解出来ていない状態になっている．

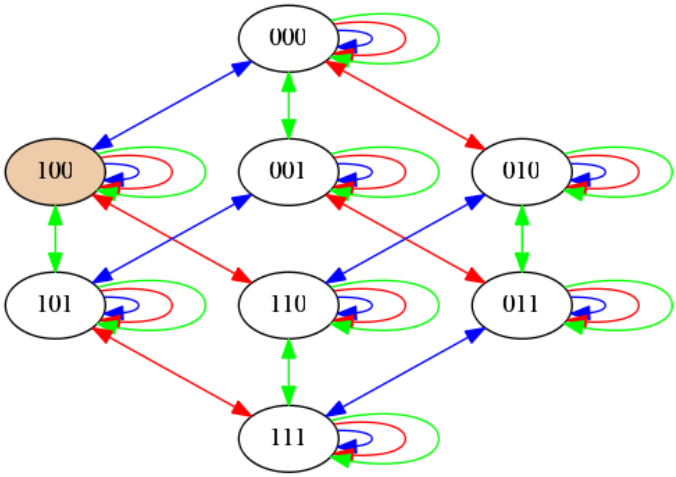

図 5.10: 初期状態

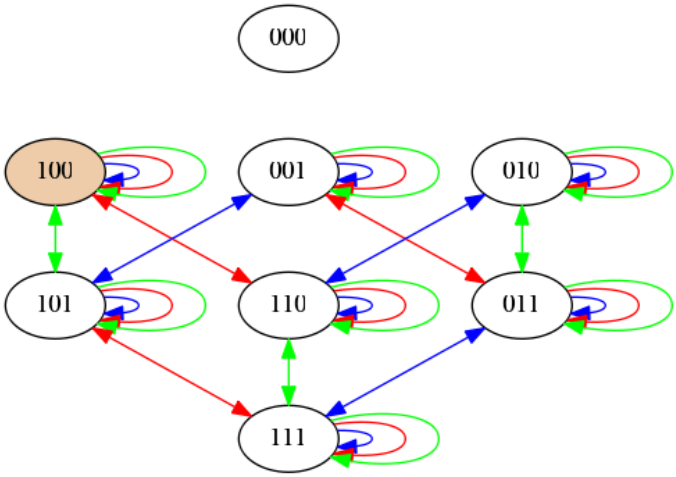

図 5.11: 父親が告知を行った後の状態

エージェント A の虚偽告知

先程の場合はエージェント A はまだ自分が汚れているかどうか分からなかったが，今回の場合では既に自分が汚れているかどうかわかっている．そこで,「自分の額が汚れているかどうか分からない」という内容の嘘をつく．これを形式化すると $\neg\Box_a d_a \wedge \neg\Box_a \neg d_a$ となり，この式は「エージェント A が汚れていると知らないし，エージェント A が汚れていないとも知らない」と解釈される．これをエージェント告知論理を用いて虚偽告知として形式化すると，$\mathrm{i}_a(\neg\Box_a d_a \wedge \neg\Box_a \neg d_a)$ となり,「エージェント A は，エージェント A が汚れていると知らない，かつエージェント A が汚れていないと知らない，と虚偽の告知を行った」と解釈される．

これまでの実験と同様に，この論理式は,「エージェント A が自分が汚れているかどうか分かる」可能世界へ向かうリンクを削除する為，図の上で分かりやすい解釈をすれば，ノード 100 へ向かうリンクを削除する事になる．そして，エージェント B とエージェント C は現時点での知識状態においては，エージェント A が嘘をついていると判別出来ないので，共にエージェント A の嘘を受け入れてしまう．

この告知を実行した結果を図 5.12 に示す．

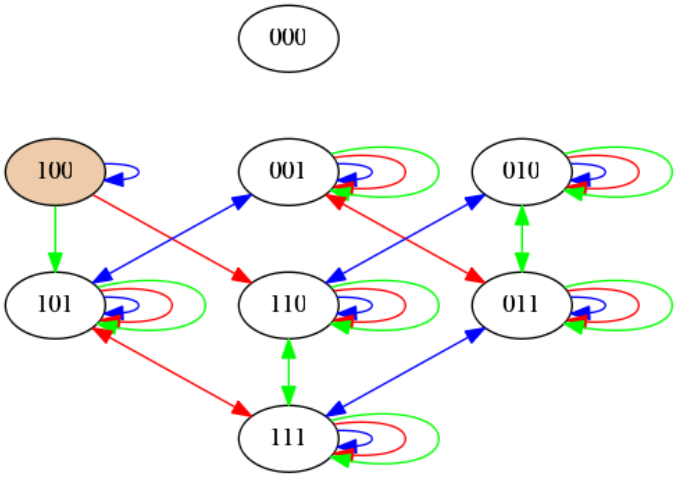

図 5.12: エージェント A が虚偽の告知を行った後の状態

この告知を実行すると，エージェント B とエージェント C は実世界からそれぞれ別のノードに到達可能性を残している状況になる．すると，実世界において，エージェント B はノード 110 への，エージェント C はノード 101 へのリンクしか存在しておらず，先程の実験結果とは異なり，エージェント 2 人とも騙せる事が出来たと考えられる．そして，先程までの実験と同様，エージェント A は自分が嘘の告知を行ったと自覚しているので，エージェント A のリンクには変更を加えない．

エージェント B の告知

続いてエージェント B が発言を行う順番である．図 5.12 において，エージェント B は実世界であるノード 100 からノード 110 を信じており，かつノード 110 には他のノードへ向かうリンクは存在しない．なので，エージェント B はエージェント A の告知によって騙されているのにも関わらず,「自分が汚れているかどうか分かる」と信じてしまい，告知も正直に「自分が汚れているかどうか分かる」と発言してしまう．

しかし，エージェント C から見たとき，エージェント C が信じているノード 101 においては，エージェント B はまだ信念が固定されていないはずなので，エージェント C はエージェント B の「(A に騙された上での）正直な告知」を「虚偽の告知」と断定してしまい，告知を棄却してしまう．告知を棄却するので，このエージェント B の告知によるエージェント C のリンク削除は行われない事になる．

この告知の結果を図 5.13 に示す．

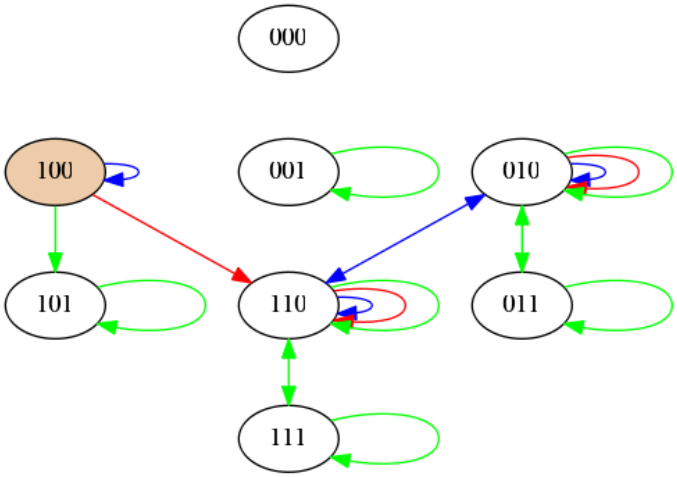

図 5.13: エージェント B が真実の告知を行った後の状態

この告知によって，エージェント A とエージェント B は,「エージェント B が汚れているかどうかエージェント B が分かる」世界からのリンク以外をカットする事になる．図 5.12 において，エージェント B が汚れているかどうかエージェント B が分かる世界とは，ノード 100 とノード 010 とノード 110 の三つである．そこで，エージェント A とエージェント B はこの三つの世界から出るリンク以外を削除する．だが，先述の通り，エージェント C にとってエージェント B は虚偽の告知を行っているように見えるので，告知を棄却し，リンクを削除する事はしない．

エージェント C の告知

続いてエージェント C が発言を行う順番である．図 5.13 において，エージェント C は実世界であるノード 100 からノード 101 を信じており，かつノード 101 には他のノードへ

向かうリンクは存在しない．故にエージェント C は（エージェント B と同様に）エージェント A の告知によって騙されているのにも関わらず,「自分が汚れているかどうか分かる」と信じてしまい，告知もエージェント B 同様に，正直に「自分が汚れているかどうか分かる」と発言してしまう．

しかし，こちらもまた同様であるが，エージェント B から見たとき，エージェント B が信じているノード 110 において，エージェント C はまだ信念が固定されていないはずなので，エージェント B はエージェント C の「（A に騙された上での）正直な告知」を「虚偽の告知」と断定してしまい，告知を棄却してしまう．こちらでも同様に告知を棄却するので，このエージェント C の告知によるエージェント B のリンク削除は行われない．

この告知の結果を図 5.14 に示す．

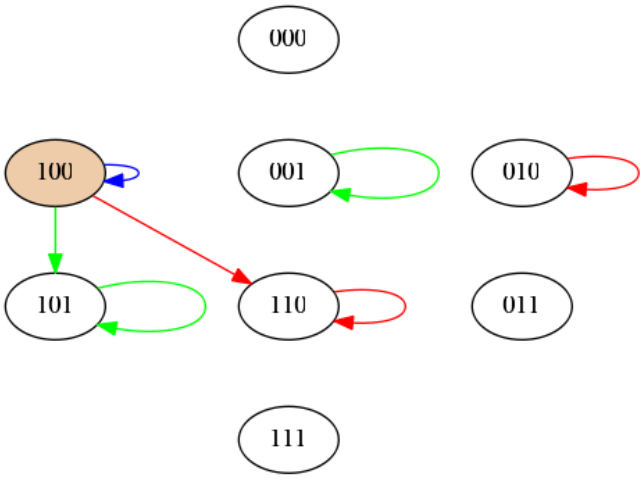

図 5.14: エージェント C が正直に告知を行った後の状態

この告知により，エージェント A とエージェント C は「エージェント C が汚れているかどうかエージェント C が分かる」世界からのリンク以外を削除する．図 5.13 においてエージェント C のリンクが 1 本しか存在しない世界とは，ノード 100 とノード 001 とノード 101 の三つである．そこでエージェント A とエージェント C はこの三つのノードから向かうリンク以外を削除する．

虚偽の告知の効果

エージェント A は前項で「自分が汚れているかどうか分からない」状態で「自分が汚れているかどうか分かる」と嘘をついた．しかし，その時はエージェント B を騙す事は出来たが，エージェント C にはその嘘を見抜かれてしまった．

今回はエージェント A は「自分が汚れているかどうか分かる」状態で「自分が汚れているかどうか分からない」と嘘をついた．そして，今回の場合ではエージェント B とエージェント C をそれぞれ別の世界を信じさせる事で，自分の告知を疑わせる事無く，しか

もエージェントBにはエージェントCが，エージェントCにはエージェントBが嘘つきであると疑わせる事が出来た．

今回の検証ケースを人狼で置き換えて考えるなら，例えば，エージェントAが人狼役で，エージェントBとエージェントCを人間役と仮定してみる．すると，エージェントAが嘘の告知を他のエージェントに信じさせる事ができれば，他のエージェント同士を疑わせた上で，自分は信頼を勝ち取る事も出来る，といった事も表現出来るのではないかと考えられる．

第6章　おわりに

6.1　まとめ

本論文では，マルチエージェントシステムの一例として古典的論理パズルである Muddy Children Puzzle を，エージェント告知論理の変形を用いて解き直す事で，エージェントの知識状態を，公開告知が行われた際に，各エージェントが持つ知識状態と，それに準ずる，あるいは反する知識の候補を得て，それぞれのエージェントがその公開告知を受け入れるか，あるいは棄却するかという事が確認できる機構を提案した．Muddy Children Puzzle においては，告知を段階的に行う，エージェントが嘘をつく事を想定する，という 2 点において問題を再定義した．また，エージェント告知論理は，告知を受けたエージェントが，それが自分自身が持つ知識と矛盾する告知だった場合には，告知を棄却するよう意味論を再定義した．

実験としては，Muddy Children Puzzle を基にして，3 名のエージェントが自分自身が泥をつけているかつけていないかを分かっているか否か，という事を段階的に 1 名ずつ告知を行い，他のエージェントはその告知を受け入れるか否かという事を確認した．この実験において，システム上でエージェントは任意のタイミングで嘘をつく事ができ，受け手のエージェントはその告知が自分自身の持つ知識と矛盾する場合はその告知を受け入れずに棄却するシステムを実装した．

その結果として，エージェントが嘘をつく事によって本来持ち得なかったはずの知識を得る事が出来る，エージェントが嘘をつかれた時，自分自身の知識と矛盾しない場合には，嘘を受け入れてしまって実世界とは異なる世界に居ると錯覚してしまう，自分自身の知識と矛盾する場合には，告知を棄却する事で嘘を見抜く事が出来る事を示した．

6.2　今後の課題

今後の課題として以下のような問題が残った．

1. 本研究の最終的な目標は「人狼」の解析であったが，その前段階である Muddy Children Puzzle の再定義とその解決を行う段階までしか進める事が出来なかった．

2. 信憑性のある公開告知によって，受け手のエージェントが持つ知識を覆す，という事が表現出来なかった．
   現状では，一度嘘を受け入れたエージェントが，後々の告知によって，「以前の告知

は嘘であった」という事が判明した場合に，以前の知識状態までロールバックして，その時点の知識状態から嘘を見抜いて反映する，という事は出来ない．

3. 告知が嘘であると判別出来た時，本研究の中では単純にその告知を棄却するだけであったが，その告知が嘘であると判別出来た受け手のエージェントは，告知された論理式の否定を取る事で知識状態の更新をする，というアプローチも出来るのではないか．

4. 機構の設計を行う際に，エージェント告知論理を改変したが，その際に公理を定義出来ていなかった．

参考文献

[1] Are You A Werewolf?, The Official Are You A Werewolf? Homepage, http://www.wunderland.com/LooneyLabs/Werewolf/

[2] van Ditmarsch, Hans,van der Hoek, Wiebe, Kooi Barteld. *Dynamic Epistemic Logic*, Synthese Library, Vol. 337.

[3] Hans van Ditmarsch, Jan van Eijck, Floor Sietsma, Yanjing Wang. On the Logic of Lying, *Games, Actions and Social Software. Lecture Notes in Computer Science* Volume 7010, 2012, pp. 41-72.

[4] 山田友幸. ”社会的コミュニケーションの論理的ダイナミクス”, International Symposium on Methodologies for Intelligent Systems. Vol. 201. 1989.

[5] Plaza, J. A., 'Logics of public communications'. Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems. pp. 201-216.1989

[6] Chapter 5 Knowledge and Imformation Flow, Logic in Action Open Course Project, http://www.logicinaction.org/

[7] Baltag, Alexandru, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicions, *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge* Morgan Kaufmann Publishers Inc., 1998.

[8] 沼岡千里, 大沢英一, 長尾確. ”マルチエージェントシステム”, 共立出版 (1998)

[9] 伊藤孝行, 新谷虎松, ”マルチエージェントシステムのための実装技術とその応用”, 人工知能学会論文誌 12.1 (1997)

[10] 杉山公造, 永田晃也, 下嶋篤, 梅本勝博, 橋本敬. ナレッジサイエンス:知を再編する 81 のキーワード, 近代科学社 (2003)

[11] 石田亨, 片桐恭弘, 桑原和宏, ”分散人工知能”, コロナ社 (1996)

[12] van Ditmarsch, Hans. The Ditmarsch Tale of Wonders-The Dynamics of Lying, *Reasoning About Other Minds: Logical and Cognitive Perspectives* (2011): 65.