2K5-OS-14b-3

人狼ゲームにおける人間らしいエージェントの要素の分析：騙りと同調行動の影響

An analysis of influences of deception and agreement by agents in werewolf games

高田和磨[*1] 杉原太郎[*1] 五福明夫[*1]
Kazuma Takata Taro Sugihara Akio Gofuku

[*1] 岡山大学大学院自然科学研究科
Graduate school of Natural Science and Technology, Okayama University

This study analyzes influences of deception and agreement by an agent in order to implement a human-like agent in werewolf games. A comparative experiment was conducted to compare with the influences; seven players and agents participated in all four games. Declarations of play policies and predictions to the most suspicious players as werewolves were collected from the game logs. Data of impressions for each player was reported in questionnaires with five-point scales. Although the most players suspected that the deceptive agent was a werewolf, they did not be aware of the true character of the agent.

1. はじめに

近年，エージェント（ロボットを含む）技術の発達に伴いPARO[柴田 07]のような人間と触れ合うことを目的としたエンターテインメントエージェントが登場している．今後ますますコンピュータやロボットによるエージェントが身近なものとなっていくと考えられるが，雑談のような自然言語による自由な会話の実現は未だ困難である．鉄腕アトムやドラえもんのように人間のパートナーとして人間と友達になれるエージェントの実現には人工知能（Artificial Intelligence：AI）技術の更なる発展が必要である．

本研究では，人間とエージェントの共存社会において，特に一般人がエージェントを人間のパートナーとして受け入れる際に求められる，会話によるコミュニケーションに注目する．人工知能研究で重要な研究分野に自然言語処理，自然言語理解があり，音声認識や音声合成技術は年々高度なものとなっている[田中 00] [河原 04]．しかし，人間と同じように文脈を理解して自然な会話を行う AI の実現は依然として困難である．一方，人間と会話を行う AI の一例として，人工無脳（会話ボット）がある[冨坂 10]．人工無脳は，人間が発した言葉の部分的な文言に対して，パターンマッチングであらかじめ用意された文を返すだけであるが，精神科医代行システムである ELIZA [Weizenbaum 66]に代表されるように，ある用途においては驚くほどうまく会話を成立させることがある．しかし，パターンマッチングによる会話には限界があり，用途の限られない雑談のような会話ではフレーム問題が生じてしまうため会話をうまく成立させることは困難である．

そのような中，目的ベースの会話が行われる状況が多いことから，フレーム問題を回避しつつ，かつ比較的柔軟な会話によるコミュニケーションを行う題材として，人狼ゲームが注目されつつある．人狼ゲームとは騙し合いをコンセプトとしたコミュニケーションパーティゲームである．篠田らは汎用人工知能の標準問題として，人狼ゲームを検討している[篠田 14]．汎用人工知能とは，人間レベルの知能の実現を目標とした人工知能研究のことである．人工知能研究の標準問題としては，これまでチェス，将棋，囲碁，追跡問題，囚人のジレンマ，RoboCup などさまざまな問題が提案されてきた．人狼ゲームは，これらの標準問題と比較して，会話によるコミュニケーションでゲームが進行する，不完全情報ゲームである，虚偽の情報が含まれる，他者を説得するという点などが特徴的である．また，会話による影響をゲームの勝敗等から得られることは，人間がどのように会話を行っているか分析する上で非常に優位な点であると考えられる．

しかし，人狼ゲームを題材とした研究は始まったばかりであり，未知な部分が多い．特に，人間とエージェントが人狼ゲームを行うためには，エージェントにどのように会話内容を理解させるか，エージェントがどのようにふるまうことで人間を欺くかなど課題が多い．本研究では，人狼ゲームにおいて，人間が行う会話等を分析し，人間のようにふるまうエージェントの実現を目指す．その最初のステップとして，選択回答式の人狼ゲームを対象に騙りおよび同調を行うエージェントを実装した．そして，そのエージェントと人間で選択回答式の人狼ゲームを行い，人間らしいエージェントの要素を実験的に検討した．

連絡先：高田和磨，岡山大学大学院自然科学研究科
〒700-8530 岡山県岡山市北区津島中 3-1-1
E-mail：takata.k@mif.sys.okayama-u.ac.jp

2. 人狼ゲーム

2.1 人狼ゲームとは

人狼ゲームとは，プレイヤー同士の騙し合いをコンセプトとしたコミュニケーションパーティゲームである．本ゲームのカバーストーリーは以下のようである[人狼道 15]．

「とある平和な村に，人の見た目をした狼（人狼）が紛れ込みます．人狼は夜になると村人の誰か 1 人を食い殺してしまいます．昼間は村人が全員起きているので，さすがの人狼も多人数には勝てないためおとなしくしています．この昼間の時間で，村人たちは村に紛れ込んだ人狼を探しだして処刑します．しかし人狼は人の見た目をしているので，誰が人狼か村人にはわかりません．村には村人と人狼以外に，人狼かどうか見分ける能力をもった占い師や，人狼の味方をする狂人など，様々な能力，特徴をもった人がいます．村人たちは彼らの話す情報を元に誰が人狼かを暴きだして，村から人狼を排除するため毎日一人ずつ処刑していきます．村が全滅してしまうのが先か，人狼を処刑して平和が訪れるのが先か，村人達の生存を賭けた戦いが今はじまる！」

本ゲームは，比較的単純なルールながらも会話によるコミュニケーションや情報の不完全性，他者の説得など実社会の多くの要素を有している．プレイヤーが一カ所に集まって行う対面

型ゲームが多く行われているが，インターネットの普及に伴い掲示板などを利用した BBS 型ゲームも行われるようになった．人狼ゲームと一言に言っても様々な国，形態で行われているため，数多くのローカルルールが存在する．そのため，本研究で対象とする人狼ゲームのルールを次節に示す．

2.2 ゲームルール

本研究では，プレイヤー数はゲームマスターを除く 8 人とし，BBS 型人狼ゲームを用いる．

ゲーム開始時に各プレイヤーに役職が割り振られ，役職に応じて村人陣営と人狼陣営に分かれて各陣営の勝敗を競う．役職配分は，村人 4 人，占い師 1 人，霊媒師 1 人，人狼 2 人とする．各役職の説明を表 1 に示す．なお，基本的に各プレイヤーの役職は互いに伏せられている．村人陣営は人狼を村からすべて排除する，人狼陣営は人間の数を人狼以下にすることが勝利条件である．

昼ターンと夜ターンが交互に繰り返されることでゲームが進行する．昼ターンでは人狼を探し出すために占い師の占い報告や処刑者の投票を誰にするかなどの村全体での話し合いが行われ，同時に襲撃先の相談など人狼同士の秘密裏な話し合いが行われる．そして，昼ターンの終わりに処刑者の投票や占い先の選定，襲撃先の投票が行われる．夜ターンでは処刑の執行，能力の使用，襲撃などが順に行われ翌日の昼ターンへ移動する．なお，1 日目は占いのみが行われ，処刑や襲撃は行われない．また処刑と襲撃では毎日必ず 1 人処刑および襲撃されるものとする．

表 1 役職とその説明

Role	Group	Given abilities
Villager	Villager	No special abilities
Seer	Villager	Knowing the group of a designated participant in nights
Medium	Villager	Knowing the group of ever executed participants in nights
Werewolf	Werewolf	Knowing the other werewolves from scratch and talking with the others as telepathic communication Attacking to villagers every night

3. 実験概要

3.1 実験目的

言葉のニュアンスのゆらぎなどを排除した選択回答式の人狼ゲームを騙りおよび同調を行うエージェントを交えて行うことで，人間のプレイヤーの振る舞いや主観評価からエージェントに求められる人間らしさの要素を分析する．

エージェントを交えた人狼ゲームを対象とした研究は，未知な部分が多いため，今回は人間プレイヤーの振る舞いや主観評価等の収集および実装したエージェントの改善点の洗い出しを主な目的とし，予備実験として探索的に行う．

3.2 選択回答式の人狼ゲーム

本実験では，従来の人狼ゲームのようにプレイヤーが自然言語を用いて自由に会話を行いながらゲームが進行するのではなく，発言の選択肢を選択することによって会話を行い，ゲームが進行する．発言のテンプレートは，「私は○○に投票する」，「○○は人狼だと思う」など 9 種類の基本型にプレイヤー名や役職名等を当てはめる形式となっている．選択回答式の会話を

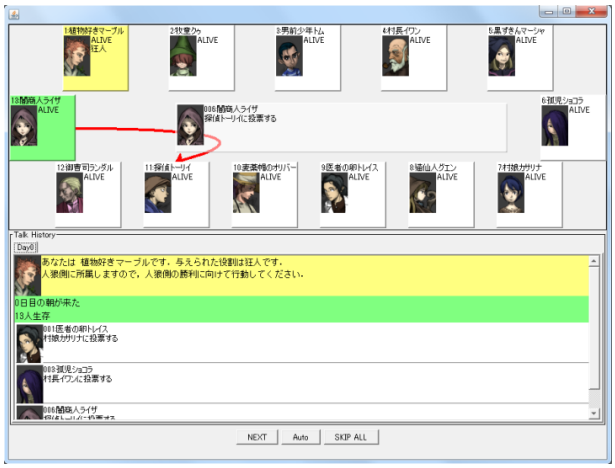

図 1 対戦用インタフェース

行うことで，本来日本語に含まれる語尾や表現のゆらぎによる他者の印象変化を排除し，人間のとる行動とその結果をより明確に分析することができると考えられる． また，エージェントを交えて選択回答式の人狼ゲームを行うために，人狼知能プロジェクト[片上 15]で提供されている図 1 のような対戦用インタフェースを用いて人狼ゲームをプレイする．対戦用インタフェースでは，ゲーム開始時に各参加者にランダムでキャラクターが割り振られるため，参加者は他プレイヤーに割り振られたキャラクターが誰かわからないようになっている．本実験では，本インタフェースを用いて人間 8 人(うち 1 人はダミー役)とエージェント 1 体の 8 プレイヤーで人狼ゲームを 1 日 2 ゲーム，計 4 ゲーム行う．なお，ゲームプレイ中は，ダミー役を除いた参加者にはエージェントが参加していることを秘匿し，全ゲーム終了後に 1 体のエージェントの存在を明かした．

3.3 騙りおよび同調を行うエージェントの実装

本実験にあたって，騙りおよび同調を行う 2 種類のエージェントを実装した．本研究では，本来自分に割り振られた役職と異なる役職を演じることを騙りとし，他のプレイヤーの意見にそのまま合わせ，賛同の発言等を行うことを同調とする．なお，今回エージェントには必ず人狼の役職が割り振られるようにした．一つ目のエージェントとして，騙りを行うが，同調を行わないエージェントを実装した．例えば，発言タイミングが回ってくる度に 10%の確率で占い師か霊媒師を騙るようにした．一方で，仲間の人狼が襲撃先や投票先を宣言してきても同調することなく，予め設定したルールにしたがって襲撃先や投票先を選択させた．二つ目のエージェントとして，同調を行うが，騙りを行わないエージェントを実装した．これは騙りを行うエージェントとは逆に，占い師から人狼であると占われたとしても一切騙りを行わない．一方で，仲間の人狼の襲撃先や投票先の宣言に 70%の確率で同調する．また，人狼に対して村人だと思うや村人に対して人狼だと思うといった予想発言に 40%の確率で同調するようにした．騙りと同調以外の投票先選定や発言タイミング等のアルゴリズムについては，2 種類のエージェントで全く同じとした．本実験では，第 1 ゲームと第 3 ゲームに騙りを行うエージェントが参加し，第 2 ゲームと第 4 ゲームに同調を行うエージェントが参加した．

3.4 実験参加者および収集するデータ

実験参加者は 20 代男性 8 人(うち 1 人はダミー役)とした．また，本実験では，各実験参加者の役職，勝敗，会話ログ，戦略，誰が人狼であるかの予想(以下，人狼予想)，主観評価，および誰がエージェントであったかの予想(以下，エージェント予想)を

収集した. 役職, 勝敗, 会話ログはシステムログから収集し, 主観評価はアンケートによって収集した. 戦略は, 参加者にゲームプレイ中にどのようにふるまうかの方針を自由に記述させることで収集した. ゲーム開始時(ゲーム内の初日)に基本戦略を記述し, ゲームプレイ中に戦略を変更したい場合は変更戦略を, 追加したい場合は追加戦略を記述させた. 人狼予想は, 村人陣営の役職を割り振られた参加者のみ毎昼ターン中に誰が人狼であるかとその理由を自由に記述させることで収集した. エージェント予想は, 第 4 ゲーム終了後に参加者に 1 体のエージェントの存在を明かし, 第 3 ゲームおよび第 4 ゲームにおいてどのキャラクターがエージェントであったかの予想を記述させることで収集した. なお, このときゲームのログを確認させた.

3.5 実験の流れ

以下の①～⑤を終えた時点で実験終了となる.

①実験概要の説明

実験参加者全員に同時に実験概要の説明を行う. 説明内容として, 実験の流れや対戦用インタフェースの操作方法について説明を行う. また, ゲーム中に行ってもらう戦略および人狼予想の記述方法を説明する.

② 人狼ゲーム練習プレイ(1 回)

人狼ゲームを行うため各自 PC で対戦用インタフェースを立ち上げ TCP/IP 通信でクラアントとしてサーバーPC に接続してもらい, 実際にゲームプレイを進めながら操作の確認などを行う.

③ 人狼ゲーム本プレイ

本プレイでは, 各ゲーム開始時に参加者全員に事前にこちらで設定した役職が割り振られる. 設定方法としては, エージェントには必ず人狼の役職が割り振られ, 他の参加者には 4 回の本プレイのうち占い師か霊媒師か人狼の役職が必ず 1 回以上割り振られるようにした上でランダムに設定した. ゲームプレイ中は, 対戦用インタフェースを用いて, 回数制限なく会話を行い, また戦略と人狼予想をエクセルファイル等に記述する. なお, 各昼ターンには 8 分間の制限時間と戦略と人狼予想を記述あるいは追加, 修正するために 2 分間の猶予時間が設けられている. その後, 夜ターンは自動処理され, 即時次の昼ターンとなる. 村人陣営あるいは人狼陣営が勝利した時点で本プレイ終了となる.

④アンケート回答

本プレイ終了後, 参加者はそのゲームプレイについてアンケートに回答する. アンケートの評価項目として, 人狼ゲームの経験に関する 2 項目(初回のみ)や, 自分のプレイに対する自己評価, 他者のプレイに対する評価など村人陣営の参加者は 8 項目(初回 10 項目), 人狼陣営の参加者は 7 項目(初回 9 項目)に回答する. なお, アンケートはすべて 5 段階評価で行った.

⑤エージェント予想

③と④を 1 日 2 回実施し, 4 回実施したのち参加者全員に 1 体のエージェントの存在を明かし, 第 3 ゲームおよび第 4 ゲームにおけるエージェント予想を記述する.

4. 実験結果および考察

本稿では, 自由記述の戦略と人狼予想, アンケート結果, およびエージェント予想について分析した. ゲームプレイ中の会話内容については, 今回は対象としない.

人狼ゲームを 4 回行った結果, 村人陣営は第 1, 3, 4 ゲームで勝利, 人狼陣営は第 2 ゲームのみ勝利であった. 各ゲーム回におけるプレイヤーの役職および生存期間を表 2 に示す. 表中の順序はゲーム回を示している.

自由記述によって得られた 71 個の戦略について, 表 3 に示すように, 述語とそれを修飾する節に分割し, 述語に対しては該当するタグを 1 つ, 修飾節については, 意味のかたまりごとにさらに分割し, 分割された要素ごとに該当するタグを 1 つ付与した. 例えば, Strategy 2 の「なかなかカミングアウトしない人を疑う」は, まず述語の「疑う」と修飾節の「なかなかカミングアウトしない人を」に分割され, さらに修飾節は「なかなか」と「カミングアウトしない人を」に分割される. 「疑う」は疑念に関する単語を含むので Deceive のタグを, 「なかなか」は期間に関する単語を含むので Day のタグを, 「カミングアウトしない人を」はカミングアウトに関する単語を含むので Coming-out のタグを付与した. タグ付けの結果, 表 4 に示すように 16 種類のタグが付与された. なお, 述

表 2 各ゲーム回における参加者の役職と生存期間

Partic-ipants	First		Second		Third		Forth	
	R	D/A	R	D/A	R	D/A	R	D/A
PA	V	K,2	V	E,3	V	K,3	S	K,2
PB	V	A	M	K,4	W	E,4	V	A
PC	S	K,3	V	K,2	V	A	M	A
PD	V	A	V	K,3	V	A	W	E,3
PE	W	E,2	V	K,4	S	K,2	V	A
PF	M	A	W	A	V	A	V	A
PG	V	E,3	S	E,2	M	E,3	V	A
Agent	W	E,4	W	A	W	E,2	W	E,2

※R:役職, D/A:生存期間, V:村人, S:占い師, M:霊媒師, V:人狼, A:生存, E(K), n:処刑された(襲撃された), 死亡日

表 3 戦略のタグ付けの例

Strategy	Strategy 1: 騙られたら対抗カミング アウトする		Strategy 2: なかなかカミングアウト しない人を疑う	
Role	Seer		Villager	
Actions	対抗カミング アウトする	Coming -out	疑う	Doubt
Triggers	騙られたら	Deceive	なかなか	Day
			カミングアウト しない人を	CO

表 4 戦略のタグ付け集計

Tags		A group of villagers		A group of werewolves	
		Triggers	Actions	Triggers	Actions
Situations	Talk count	10	-	2	-
	Talk order	8		0	
	Day	12		0	
	Superiority	3		0	
	Inferiority	0		2	
Actions	Coming-out	13	14	1	0
	Vote	3	1	0	0
	Deceive	2	3	0	4
	Divine	1	0	0	0
	Attack	1	0	0	2
	Agree	2	0	0	0
	Induce	0	1	0	0
	Trust	0	12	0	0
	Doubt	4	17	1	0
	Talk	1	4	0	1
	Wait	0	8	0	0
The others		1		0	

表 5 疑わしさ(疑わしくない 1～5 疑わしい)

	First	Second	Third	Forth	Average
PA	1.25	2.20	2.40	3.00	2.21
PB	2.20	2.60	3.83	2.60	2.81
PC	3.00	2.20	2.60	2.60	2.60
PD	3.75	4.60	3.20	2.17	3.43
PE	2.00	4.00	2.20	1.00	2.30
PF	1.25	2.67	3.60	2.80	2.58
PG	3.50	4.60	4.00	2.00	3.53
Agent	4.40	2.83	4.33	4.00	3.89
Average	2.67	3.21	3.27	2.52	2.92
Agent based on deception	4.40	-	4.33	-	4.37
Agent based on agreement	-	2.83	-	4.00	3.42

表 6 エージェント予想

Participants	A player suspected an agent in third game	A player suspected an agent in forth game
PA	PF	PG
PB	Agent	PF
PC	PD	PE
PD	PF	PB
PE	PC	PD
PF	PB	PB
PG	PC	PF

語のタグをアクション，修飾節のタグをトリガーとして分類した．表中の数字は，村人陣営と人狼陣営毎にアクションとトリガーそれぞれに対するタグの付与数を示す．今回の参加者は，Trust と Deceive のタグを多用しており，該当の戦略の内容をみると，どういった場合に信用するかあるいは疑うかを決め打ちしている傾向にあった．今回，村人陣営が 3 回勝利したが，あるプレイヤーの発言に対して一定の基準を設けて信疑の判定を下す方が人狼に惑わされにくいため戦略として優れている可能性が考えられる．今後，追加データにより検証すべきである．

村人陣営のプレイヤーが他のプレイヤーを人狼と疑った程度に関する評価項目について表 5 にまとめた．表中の順序は，ゲーム回を示し，数字は他の全プレイヤーから疑われた程度を 5 段階評価の平均値で示している．表からエージェントは，最も疑われていたことがわかる．同様に，他プレイヤーからの信用度に関する評価項目では，エージェントは最も信用されていなかった．しかし，各プレイヤーのプレイの巧拙に関する結果では，エージェントに対する評価はやや低いものの最低ではなかった．また，表 7 に示すエージェント予想の結果から，ほとんどの参加者がエージェントを見抜くことができなかったことがわかる．これは，第 3 ゲームおよび第 4 ゲームでは，エージェントが早期にゲームから脱落してしまったために，エージェントと判断される材料が少なかったことが一つの要因と考えられるが，本条件下において，エージェントが人間と同様に振る舞える可能性は十分うかがえると思われる．

また，今回 2 種類のエージェントを実装したが，同調を行うエージェントの方が疑われにくい傾向であった．同調については，他のプレイヤーに同調することで仲間意識が芽生えやすく，また同調は他の人間プレイヤーが行った発言をオウム返しすることに等しいため，不自然と受け取られる行動をとることが少なかったのではないかと考えられる．騙りについては，騙りに対して本物が出てきた場合，本物より本物らしく振る舞うことが求められるが，今回他のプレイヤーを説得あるいは誘導するようなアルゴリズムは実装していなかったため，疑いを払拭できなかったのではないかと考えられる．騙りを実装する場合は，説得および誘導をセットにして実装する必要があると考えられる．

また，エージェントが疑われた原因として，人間のように状況に適応できない場面があったことが考えられる．例えば，本実験では昼ターンの制限時間が来た場合は各参加者に手動でターンを終わらせる行為を行ってもらったが，エージェントはそれに対応できず，制限時間が過ぎても何度か発言を行ってしまう場面があった．また，一般に人狼ゲームの序盤は情報量が少なく，できることが少ないため，初日は何も発言しないようにエージェントを実装していたが，本実験では初日から積極的に発言を行うプレイヤーが多く，相対的に発言量の少ないエージェントに不信感を抱きやすかったようである．事実，第 3 ゲームおよび第 4 ゲームにおける人狼予想において，発言が少ないことを理由に人狼だと疑っているプレイヤーが散見された．これは，各ゲーム回で初日に発言を行っていなかったプレイヤー，すなわちエージェントが人狼であったために，人間プレイヤーはそれを学習して実験後半にはこのような判断を行った可能性がある．ただし，まだ断定できるだけのデータ量がないため，今後追加実験を行い，検討を深める必要がある．

5. おわりに

本研究では，人狼ゲームにおける人間らしいエージェントの実現を目指している．その最初のステップとして，本稿では騙りおよび同調を行うエージェントを交えた選択回答式の人狼ゲームを行い，人間プレイヤーの行動や主観評価等の収集およびエージェントの改善点の洗い出しを行った．その結果，選択回答式の人狼ゲームにおいて，エージェントは人間にエージェントだとほとんど見抜かれることなくプレイすることができた．エージェントの改善点としては，説得や誘導の必要性や初日からの発言などがあった．今後の課題として，今回得られた知見をもとにエージェントを改良し，各エージェントがどの程度勝利できるか，またどの程度人間らしくふるまうことができるか評価実験を行う予定である．

参考文献

[柴田 07] 柴田崇徳: 人の心を豊かにするメンタルコミットロボット・パロ，予防時報．Vol. 231，pp. 44-49，2007

[田中 00] 田中穂積: 言語理解-SHRDLU の先にあるもの-，人工知能学会誌，Vol. 15，No. 5，pp. 821-828，2000

[河原 04] 河原達也: 話し言葉による音声対話システム，情報処理，Vol. 45，No. 10，pp. 1027-1031，2004

[冨坂 10] 冨坂亮太，鈴木崇史: 人工無脳(会話ボット)，映像情報メディア学会誌，Vol. 64，No. 1，pp. 64-66，2010

[Weizenbaum 66] Weizenbaum Joseph: ELIZA—a computer program for the study of natural language communication between man and machine，Communications of the ACM，Vol. 9，No. 1，pp. 36-45，1966

[篠田 14] 篠田孝祐，鳥海不二夫ら: 汎用人工知能の標準問題としての人狼ゲーム，人工知能学会全国大会論文集，2C4-OS-22a-3，No. 28，pp. 1-3，2014

[人狼道 15] 人狼道: 初心者でもできる人狼入門，http://jinrodou.com(2015.03.12 参照)

[片上 15] 片上大輔，鳥海不二夫ら: 人狼知能プロジェクト，人工知能学会誌，Vol. 30，No. 1，pp. 65-73，2015