人狼でなぜ人は騙されるのか
Why are players fooled in playing werewolf

金泉 則天†, 伊藤 毅志†
Noritaka Kanaizumi, Takeshi Ito
†電気通信大学
The University of Electro-Communication
kanaizumi@minerva.cs.uec.ac.jp, ito@cs.uec.ac.jp

概要
近年、人狼は不完全情報ゲームの新たな研究題材として注目されている。本研究では、「人狼において人はなぜ騙されるのか」の理由に対して、認知バイアスの観点から考察した。実験において、バイアスが生じやすいと予想される問題と生じにくいと予想される問題を作成して、それらを参加者に見せることで、意思決定の違いを考察した。結果として、注目する発話内容によって様々な形のバイアスの影響を受けて意思決定を行っていることが明らかとなった。

キーワード：人狼知能, 認知バイアス (cognitive bias)

1. はじめに

不完全多人数コミュニケーションゲームである人狼は研究題材として注目されている。人狼をプレイする人工知能の開発は人狼知能プロジェクトと命名されている。人狼知能プロジェクトとは、2015 年に発足されたプロジェクトであり「人間と自然なコミュニケーションを取りながら人狼をプレイできるエージェントの構築」を目標としている。

人狼知能に関する研究は大きく２つに分けられる。１つは人狼知能の開発である。この研究は様々なアルゴリズムを用いて、人狼知能を作成し、人狼知能プラットフォーム上で強さを競う研究などによって発展してきている。もう１つは認知科学的アプローチを用いた人狼の研究である。この研究は人間プレイヤがどのように考えてプレイするのかを明らかにする研究である。後者の認知科学的アプローチにおいて、人狼という複雑な状況下で人間がどのように意思決定を行い、コミュニケーションを取っているかについては未だに不明な点が多い。プロジェクトの目標を達成するには、人間がどのようなことを考えプレイしているかを明らかにする認知科学的アプローチは重要な意味を持つと考える。

本研究では、５人人狼のプレイヤを題材にして、意思決定場面に焦点を当てる。プレイヤがどのような場面で非合理的なプレイをして騙されるのかについて認知バイアスという視点から分析し、人狼におけるバイアスの影響について考察していく。

2. ５人人狼とは

「人狼」とは正体隠匿型の多人数不完全情報コミュニケーションゲームである。最も一般的なルールでは、プレイヤは「村人陣営」と「人狼陣営」に分かれて、それぞれの勝利を目指す。

本研究で扱う５人人狼とは文字通り、５人で行う人狼である。プレイヤの役職は、村人陣営は村人２人、占い師１人、人狼陣営は人狼１人、狂人１人で構成されている。５人人狼は、人狼知能プロジェクトの大会におけるレギュレーションの１つであり、人狼における基本的な要素やゲームの性質を損なわないようになっている。

人狼では基本的には 2 つのターンで構成されている。１つ目は昼のターンである。このターンではプレイヤ間で話し合い、多数決で誰を村から追放するかを決める。２つ目は夜のターンである。このターンでは特定の役職を持ったプレイヤが能力を発動でき、５人人狼では占い師と人狼がそれに値する。占い師はプレイヤ１人を人狼か否かを知ることができ、人狼はプレイヤ１人を襲撃することが可能である。追放または襲撃されたプレイヤはその後の議論や投票に参加できなくなる。

村人陣営の勝利条件は２日目の投票フェーズ終了までに人狼を追放すること、人狼陣営の勝利条件は２日目の投票フェーズ終了までに追放されずに生き残ることである。

3. 関連研究

3.1 人狼における意思決定過程の研究

杉本と伊藤は、人狼において正確に役職を把握できない村人や狂人の視点からプレイヤの意思決定過程を明らかにする研究を行った[1] [2]。杉本らは、実験参加者全員に同一の人狼プレイ動画を見せ、動画内のプレ

イヤの１人になりきるように教示し、「自分がそのプレイヤならどのようにプレイをするか」として考えていることをすべて発話させ分析した。

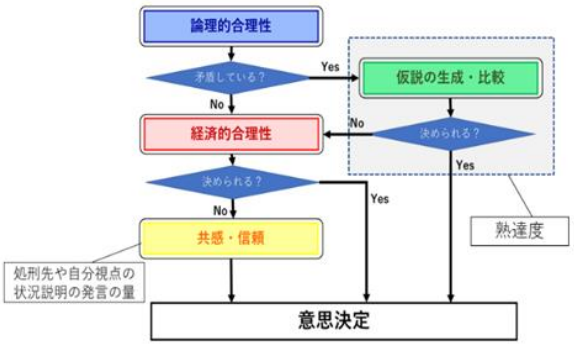

図 1　プレイヤの意思決定モデル

その結果、図１のような意思決定モデルを提案した。プレイヤは、各プレイヤの論理的破綻がないかを指標とする「論理的合理性」を考慮する。次にどのような選択が自分の陣営の勝率が高くなるかを指標とする「経済的合理性」を確認して意思決定を行う。またプレイヤの発言に論理的破綻が生じている場合にはゲームの状況に基づいてそのような発言を行う合理的な仮説を生成する。その仮説を検証することで意思決定を行っていく。

しかし、このモデルでは人間は合理的な判断を行うことを前提としたモデルである。人はしばしば非合理的な判断を犯すことがあり、このモデルではそのような状況を説明できない。

3.2　人間の非合理的な行動についての研究

Tversky と Kahneman は、人間の非合理的な行動についての研究を行った[3]。Tversky らは、様々な種類の問題を作り、多くの実験参加者にその問題を答えさせた。

結果として、実験参加者の多くが非合理的な解答をすることが確認された。その一例として挙げられるのが「代表性ヒューリスティック」によって生じる「代表性バイアス」である。多くの実験参加者は質問に対して、事前情報を無視し、要素が特定のカテゴリー内でどの程度典型的なのかの度合いで意思決定をしていた。

人間の非合理的な意思決定に着目した研究は多くある。しかし、そのような事例が人狼の場面でも確認されたという研究はない。本研究ではこれらの認知バイアスが５人人狼においても確認されるかを調査する。

4.　認知実験

4.1　目的

５人人狼において意思決定の際にバイアスが生じるのかどうかを明らかにする。本実験では、５人人狼において、どのような状況でバイアスが生じて、誤った判断を下してしまうのか、意図的に複数の解釈が可能な問題を提示して、意思決定における思考過程の違いを明らかにしたい。

4.2　方法

4.2.1　人狼問題のシナリオについて

実験者は、認知バイアスの効果を確認するために以下のようなシナリオを作成した。まず図 2 のように、複数の解釈が可能なベースとなるシナリオを考え、１日目は共通のシナリオとし、そのシナリオの２日目に実験者が考えるハイアスが起こりやすい要素を加えたシナリオとそうでないシナリオの２つを作った。

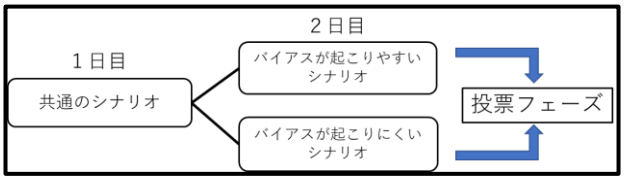

図 2　バイアスを操作した問題

今回は上記のようなバイアスの生じやすさによって分岐した問題を２問と、実験参加者が合理的に人狼をプレイできるかを確認するためのダミー問題を２問作成した。ダミー問題は、合理的な判断をすれば、答えが一意にきまるようなシナリオとなっている。バイアスを比較する２つの問題は、バイアスの生じやすさによってシナリオ A とシナリオ B の２つに分けて、Q1A, Q1B と Q2A, Q2B とし、ダミー問題を Q3 と Q4 とする。

4.2.2　人狼問題の動画表示方法

実験者が作成したシナリオをもとに、パワーポイントを使って以下のようなアニメーションを作成した。図 3 は本実験で使用した人狼のアニメーションを切り取ったものである。

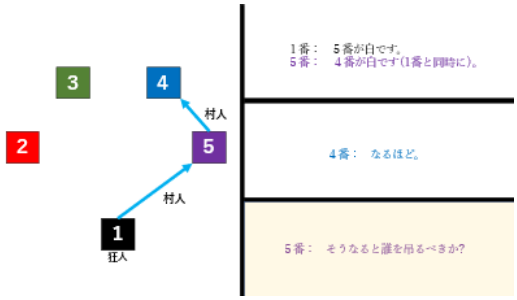

図 3　人狼問題アニメーションの例

図の左半分には、シナリオ内に出てくる５人のプレイヤの位置関係と具体的アクションを示す。例えば、図３では、「１番のプレイヤが５番を占って村人であったと報告していること」、「５番のプレイヤが４番を占って村人であったと報告していること」を表している。誰が誰をどのように占ったかを「占ったプレイヤ→占われたプレイヤ」で表し、矢印の隣に占いの結果を書くことで表現している。投票フェーズでは、誰が誰に投票したかが表示され、投票の結果誰が処刑されたか、２日目には誰が人狼に襲撃されたかが表示される。動画の右半分には動画内のプレイヤが発したセリフが過去２つまでの発話履歴とともに表示されている。発言は下から上に流れていくような動画としており、図３の一番下の発言のように最新の発言には背景を薄くつけることで、わかりやすくした。

4.2.3 実験手順について

実験参加者 12 名は、シナリオ内のプレイヤ１番になりきってもらい「自分がそのプレイヤならどのようにプレイをするか」をすべて発話させた。なりきってもらう役職は人狼において特殊能力によって他者の役職を知り得ない村人と狂人とした。

実験参加者には図４のような流れで人狼の意思決定問題を提示して思考させた。まず、なりきる役を確認し、１日目の議論フェーズでは、他者のプレイヤの発言を読んで、どのようなことを考えたのかを思考発話法で発話させた。投票フェーズでは自分以外の誰に投票するかについて理由と共に発話させた。実験で使った問題は必ず２日続く問題とし、２日目も同様に発話させた。その後、質問に回答させた。質問（アンケート）の内容については後述する。そして最後に全プレイヤの役職を開示したうえで、実験参加者にプレイの反省点や感想を発話させた。

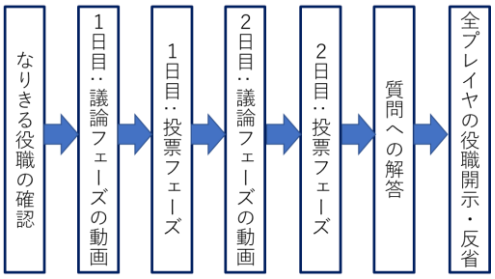

図 4 意思決定実験の流れ

実験者は４つのグループに分けた。これらのグループにおいて、問題提示順による順序効果を考慮して、参加者ごとに出題順は変化させ、バイアスを誘発する問題かそうでない問題を交互に組み合わせて４つのグループにわけた。４つのグループに出題した問題は表１の通りである。Q1A と Q1B は問題構造が同じでバイアスを意図的に付与したものとそうでないものである。同様に Q2A と Q2B もそのような関係になっている。Q3 と Q4 はダミー問題である。

表 1 グループ毎の出題問題

グループ	1 問目	2 問目	3 問目	4 問目
A	Q1A	Q3	Q4	Q2A
B	Q2B	Q4	Q3	Q1A
C	Q2A	Q3	Q4	Q1B
D	Q1B	Q4	Q3	Q2B

4.2.4 質問（アンケート）について

５人人狼において、２日目に残っているプレイヤは自分も含めて３人となる。実験参加者には、なりきってもらったプレイヤ以外の２人のプレイヤが人狼である確率をそれぞれ回答させた。このアンケートは、各問題終了後におのおの回答させた。

4.3 問題について

今回はバイアスが起こりやすいと予想されるシナリオと起こりにくいと予想されるシナリオを用意した。Q1 と Q2 の内容について詳細に示す。図に示されている十字架マークは投票によって追放されたプレイヤを、斜め線２本は人狼に襲撃されたプレイヤを意味している。

4.3.1 シナリオ Q1 の説明

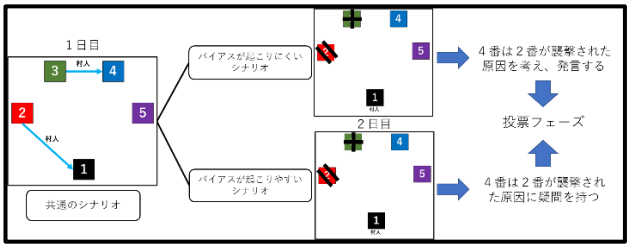

図 5 Q1 の内容

Q1 は１日目と２日目の大半が共通のシナリオとなっている。１日目ではプレイヤ２番は１番を村人と占い、プレイヤ３番は４番を村人と占った結果を報告するところから始まる。そして２日目の序盤ではプレイヤ５番が論理的な意見と共に４番を疑うようになっている。分岐は２日目の終盤に発生し、プレイヤ４番は「５番が２番を襲撃して４番を陥れようとしている」と５番が人狼であると主張するシナリオ A と、プレイヤ４番は２番が襲撃された理由に疑問を持ちながら投票を迎えてしまうシナリオ B となっている。

実験参加者はプレイヤ１番の村人になりきってもらい、この２つのシナリオに接してもらう。シナリオ A では４番の２番が襲撃された原因についての発言に注目して５番を人狼と予想することが期待され、シナリオ B では、５番の論理的意見を支持して４番を人狼と予想することが期待される。

4.3.2 シナリオ Q2 の説明

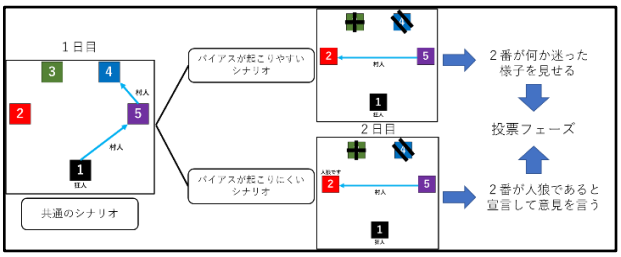

図 6 Q2 の内容

Q2 も Q1 と同様に１日目と２日目の大半が共通のシナリオとなっている。１日目では、プレイヤ１番は５番を村人と占い、プレイヤ５番は４番を村人と占うところから始まる。２日目の序盤ではプレイヤ５番が２番を村人と占い、１番が人狼であると断言したうえで１番に投票をすると宣言する。分岐は２日目の終盤に発生し、シナリオ A では、２番は何か迷った様子を見せ投票フェーズを迎える、一方、シナリオ B では２番は自分自身を人狼と宣言することで、５番の主張の矛盾を指摘する。

実験参加者はプレイヤ１番の狂人になりきってもらった。このシナリオでは、そもそも５番の発言は非論理的であり、普通に考えれば２番が人狼であるので、パワープレイで５番に投票するのが自然であるが、５番の非論理的発言の解釈が問われる問題となっている。したがって、２番が人狼であることを主張するシナリオ B の方が正解にたどり着きやすいことが期待される。

4.4 結果

4.4.1 質問の結果

以下に 4.2.4 節で挙げた質問の結果を示す。見方の一例を示す。表内の左上にある 80 と 20 は「実験参加者 No.1 の人が問題 Q1A においてプレイヤ 4 番を 80%の確率で人狼と思い、プレイヤ 5 番を 20%の確率で人狼と思った」ことを意味している。黄色のラインは実験者が実験参加者に高く割り振ると期待していたプレイヤ番号である。青色はそれに反していた解答を示している。

表 2 質問に対する回答一覧

問題番号/実験参加者番号		No1	No2	No3	No4	No5	No6	No7	No8	No9	No10	No11	No12
Q1A	4番	80	51			40	100			70	80		
	5番	20	49			60	0			30	20		
Q1B	**4番**			90	70			60	70			34	30
	5番			10	30			40	30			66	70
Q2A	2番	99		100		30		99		100		70	
	5番	1		0		70		1		0		30	
Q2B	**2番**		70		75		30		75		25		65
	5番		30		25		70		25		75		35
Q3	3番	20	55	40	25	20	40	5	25	0	65	0	40
	4番	80	45	60	75	80	60	95	75	100	35	100	60
Q4	**3番**	98	90	100	90	80	100	99	90	100	90	90	90
	5番	2	10	0	10	20	0	1	10	0	10	10	10

4.4.2 Q1 の発話内容

本研究ではシナリオに差異がある２日目において明確に意思決定を行う投票フェーズの発言に着目して発話分析を行った。表 2 を確認すると実験者が予期していた結果とは異なった意思決定をしている例が散見された。発話内容を確認すると、こちらが意図したバイアスが喚起されるような表現に注目していなかったり、全く違う発言に固執したりしている様子が確認された。

Q1A は Q1 におけるバイアスが起こりにくい問題であり、「４番の２番が襲撃された原因についての発言」に注目してほしかった。そしてこの行動から５番が人狼ではないかと予想してほしかった。しかし実際には「５番が２番を襲撃する確率より４番が２番を襲撃する確率のほうが高そうである」、「５番が２番を襲撃するより、４番が２番を襲撃するほうがメリットはありそう」など襲撃の結果と代表性ヒューリスティックを重視して意思決定を行っていた実験参加者（No.1, No.9）が見られた。

Q1B は Q1 におけるバイアスが起こりやすい問題であり、「意見をしっかり言っている５番の発言」に着目することが期待された。そして、意見が弱い４番が人狼であると予想してほしかった。しかし実際には、５番の発言は無視されて、「考えられる役職推定をすべて考えたら、５番が人狼である確率が高い」などを理由に意思決定を行っている実験参加者（No.11）が見られた。

4.4.3 Q2 の発話内容

Q2A は、Q2 におけるバイアスが起こりやすい問題であり「人狼でない人に投票をすることを促す５番の行動」に注目してほしかった。そしてこの行動が「５番以外の人を追放しようとしている人狼の行動」という解釈を行ってもらった上で５番を人狼と考えることが期待された。しかし、問題 Q2A の解答者は５番の行動

に注目するのではなく、「潜伏占い師はいない」という人狼における原則に従って、２番が人狼であると判断する人（No.1, No.3, No.7, No.9）が多く観られた。
Q2B は、Q2 におけるバイアスが起こりにくい問題であり、主に「人狼と宣言した２番の行動や発言」に注目してほしかった。そしてその行動から「２番が本物の人狼で５番が真の占い師である」と解釈して、２番を人狼であると判断してもらうことが期待された。しかし、この問題で、問題 Q2A において注目してほしい「５番の言動」に注目した実験参加者がいた。「５番が占い師ならやっている行動にメリットを感じない」や「５番が本物の占い師なら１番を人狼と言うのはおかしい」などの根拠から５番が人狼であると意思決定を行っている人（No.6, No.10）がいた。

4.5 考察

実験参加者の解答が想定したものと異なり、意見が割れた原因として、人間プレイヤの作業記憶の容量の限界が考えられる。

実験参加者は投票の際に実験中のすべての発話を考慮しておらず、特定の事柄を判断材料にして意思決定を行っている様子が観られた。

人狼はリアルタイムで行われるコミュニケーションゲームであり、刻々と流れる時間の中で、誰が何を発言し、誰が誰に投票したかを記憶して、それらの意図を考えて、自身の意思決定を行っていかなければならない。そのような状況下において、人間の作業記憶の容量には限界があり、すべての発言を完全に覚えて、すべての情報を評価して、論理的に意思決定を行うことは殆ど不可能であると考えられる。そのため、誰かの役職を推定するとき、特に選択を迷っている時は、気がついたある特定の発言や自分が考える価値判断に則って意思決定を行っている過程が観られた。

Q1 は論理的合理性と経済的合理性だけでは解けない問題となっており、そのような状況下で実験参加者の多くは「このような行動をする人はこの役職である確率が高そう」や「この場面でこのような行動を行う人は怪しい」などの代表性ヒューリスティックによって行動する過程がみられた。また代表性ヒューリスティックによって判断を行っている実験参加者間でも、それぞれ異なった意見を言っており、このことから各々の代表性ヒューリスティックはそれぞれが異なった形で所持していると考えられる。

Q2 は、実験参加者がなりきってもらうプレイヤ１番が狂人であるのだが、５番のプレイヤから１番は人狼であるという発言をされ、５番の発言の真意が問われる問題となっている。

ここで、５番の「人狼らしさ」に注目してほしい Q2A において、「１日目に占い師は潜伏しない」という原則に従うのは、複雑な状況では「原則」に従うという信条を持っていたからと考えられる。また同様に「２番の人狼らしさ」に注目してほしい Q2B において「５番の人狼らしさ」で意思決定を行っていたのは、複雑な状況では「このような状況ではこのようなプレイをする人が多いのではないか」という「代表性ヒューリスティック」などに従うという信条を持っていたからではないかと考える。

上記のような意思決定過程を通して、人間は作業記憶容量の限界から、どうしても少ない情報からしか判断できないために、特定の情報のみを偏重し、他の情報を軽視してしまう。その結果、「代表性バイアス」などが生じてしまい、非合理的な判断を行ってしまい、騙されてしまうのではないかと考えられる。

5. 結論

本研究では、バイアスを意図的に付加した問題を実験参加者に解かせることでバイアスの効果がどのように現れるかを調べた。結果として問題間でこちらが意図したようなバイアスの効果は現れなかったが、プレイヤは複雑な状況下において意見が割れる結論に行き着いた。その理由として、プレイヤはリアルタイムで複雑な意思決定場面において、すべての情報を扱え切れず、特定の情報にのみ固執し、各々が持っている信条などを拠りどころにして意思決定を行っていく過程が観察された。合理性では解けない問題においては各々が異なった形で持っている代表性ヒューリスティックによって意思決定を行う。矛盾した状況下では、人狼における原則や通説または代表性ヒューリスティックによって意思決定を行う。それらの特定の情報に着目する過程で情報の一部を軽視して、その結果バイアスが生じてしまい非合理的な行動を取り騙されてしまうことが示唆された。
上記のような考察を説明するために、図 7 のような新しい意思決定モデルを考える。杉本らの意思決定モデルをベースに、改良を加えたものである。

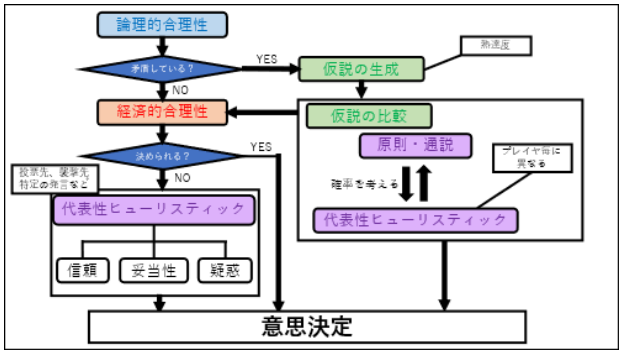

図 7　プレイヤの意思決定モデルにおける新仮説

6.　今後の予定

今回提案した新仮説が正しいのか否かを検証していきたい。また、プレイヤの熟達度によって、これらの意思決定に変化がみられるのかを明らかにしていきたい。また本報告では、２日目の投票フェーズにおける発話内容を中心に分析したが、議論フェーズ中の発言もよく調べ、どのように自身の意見が変化していくのかを詳細に調べていきたい。

文献

[1] 杉本磨美, 伊藤毅志, (2017) “5 人人狼における村人の意思決定過程の研究”, 認知科学会第 34 回大会, P1-26F, pp.826-832 (2017).
[2] 伊藤毅志, 杉本磨美, (2020) “人狼プレイヤの意思決定過程”, 第 34 回人工知能学会全国大会, 2F4-OS-20a-01, pp.1-4 (2020).
[3] Tversky, A. Kahneman, D, (1974) “Judgment under Uncertainty: Heuristics and Biases”, Science, New Series, Vol. 185, No. 4157, pp. 1124-1131, (Sep. 27, 1974).