人狼ゲームにおける 人工知能と人間の知能との差異の研究

後藤 彩乃　　（指導教員　西村 俊二）

平成31年1月25日

Differences between Human intelligence and artificial intelligence in werewolf games

AYANO GOTO　　(ACADEMIC ADVISOR　SHUNJI NISHIMURA)

概要： 人工知能技術は，囲碁，将棋，チェスなどの「完全情報ゲーム」においては研究が進められており，すでに人工知能が人間に勝利した例も存在する．このことから，人工知能は特定の分野においては人間を凌駕しているといえるだろう．そこで次の研究題材として，「不完全情報ゲーム」である人狼ゲームをプレイする人工知能である「人狼知能」が注目されている．毎年人狼知能エージェントの強さを競い合う大会が開催されるなど，人狼知能の研究は進められているが，人間らしい戦法を取らせるエージェントの研究はあまり行われていない．本研究では，人間ならば信頼を獲得するために行う戦術, 具体的には人狼が自身の信頼を上げるため仲間の人狼を捨てるような戦法を実装し，この戦術がどの程度有効であるかを勝率や周囲からの信頼度という基準を用いて調査し，考察を行った．また，人工知能における人狼ゲームでの信頼形成についても考察を行った．結果的に今回定めた信頼度の基準においては少しの信頼を得ることはできていた．しかし，勝率は大きく下がっていた．また，得られた信頼も一時的なものであり，現状のままでこの作戦は実装するだけの価値がないと思われる．しかし，将来的には信頼を得る手段として活用できる可能性はあると考えている. 人狼知能がさまざまな戦法を行うようになり, 戦法に対する共通認識が確立されたとき，現在よりも効果が上昇することが期待できるのではないかと考えられる.

キーワード: 人工知能，人狼ゲーム

1. 緒言

近年の人工知能技術の発展を象徴するものの一つに，ゲームプレイング AI がある. ゲームプレイング AI は，囲碁や将棋，チェスなどの，すべての情報がすべてのプレイヤーに公開されている「完全情報ゲーム」においては研究，開発が進められており，すでに人間に勝利した例も多く存在している．しかし，情報がプレイヤーにより隠蔽される「不完全情報ゲーム」においては, まだ研究開発が十分に行われていない状態にある．そこで注目されているのが，「不完全情報ゲーム」である人狼ゲームである[1].

人間がコミュニケーションを取り合う中では，「だます」，「見破る」，「説得する」などコミュニケーションは日常的に行われているが，これらを人工知能に実装しようという試みはあまりされていない状態にある．また，現在の人工知能は，他者や社会などの周囲の状況を認識して行動を起こすものではなく，そのための技術も確立されていない．現在，社会からのコミュニケーション AI に対する信頼は高くなく，人工知能技術が社会の中で活用されていくには，社会的なコミュニケーションの実現が必要とされる．人狼知能の実現には，「だます」，「見破る」，「説得する」などのコミュニケーションが行えるようになることが必要である. このような理由から, 人狼知能の研究は，新たな人工知能の研究題材として適しているといえる. 人狼知能エージェントの強さを競う大会が毎年開催される[2]など，人狼知能の研究は進められている．しかし，実際に人間の戦法を組み込んだ人間らしいエージェントを作成しようとする研究はあまりされていない. 一般の人狼ゲームの戦略的な行動指針(以後，「通説」と表す)の有効性は検証されており, 現在の人狼知能は通説に沿ったプログラムがされている[3] [4]. しかし，人間は通説に沿ったプレイを必ずするわけではない. 人間らしいエージェントとは，通説通りの行動をするだけではなく，さまざまな戦法も使うことができるエージェントであると思われる．そこで本論文では，人間が行う戦法を人狼知能に実装した場合の効果を，勝率，信頼度などの基準を用いて評価を行う．また，人狼知能における人狼ゲームでの他のエージェントに対する信頼の形成について調査する.

2. 人狼知能

2.1 人狼ゲーム

まず，人狼ゲームについての説明を行う．ゲームのプレイヤーには役職が割り振られ，役職の持つ能力を利用して勝利を目指す．役職は村人陣営の役職と人狼陣営の役職に分かれており，それぞれの陣営は異なる勝利条件を持っている．村人陣営の勝利条件は，人狼を全員追放することである．人狼陣営の勝利条件は，村人の人数を人狼の人数と同じ数まで減らすことである．また，村人陣営は自分が村人であるという情報しか持っておらず，誰が人狼であるかも知らない．そのため，村人陣営は与えられた情報やゲームの状況，情報をもとにした対話から，誰が人狼であるか推理を行う．それに対して人狼陣営は，ゲーム開始時に仲間の人狼を知ることが出来る．人狼陣営は仲間の人狼と協力しながら人狼であることがばれないように，ときには嘘をつきながら立ち回ることが基本的な戦術となる．人狼は相手を騙すゲームであるとも言われるが，自分の主張を論理的に説明し，周囲の人々を説得するゲームでもある．

2.1.1 人狼ゲームの流れ

人狼ゲームは，昼のフェーズと夜のフェーズに分かれている．昼のフェーズではプレイヤー全員での対話が行われ，村人はここで公開された情報をもとに推理を行う．昼のフェーズの最後には，プレイヤー全員による追放対象を決定する投票が行われ，最も得票数の多いプレイヤーが追放される．夜のフェーズは，人狼の襲撃や，役職者の能力実行が行われる．この昼のフェーズと夜のフェーズをどちらかの勝利条件が満たされるまで繰り返していく．

2.1.2 人狼ゲームの役職

人狼ゲームには多くの役職が存在する．以下では本研究内で使用された役職のみを取り上げている．

1. 村人
   能力をもたない村人陣営の役職.
2. 占い師
   夜のフェーズにプレイヤーの 1 人を対象にして，対象のプレイヤーが人間であるか，人狼であるかを知ることができる．村人陣営の役職である．
3. 霊媒師
   夜のフェーズに前日追放されたプレイヤー1 人を対象にして，対象のプレイヤーが人間であったか，人狼であったかを知ることができる．村人陣営の役職である．
4. 狩人
   夜のフェーズにプレイヤーの 1 人を対象にして，そのプレイヤーを人狼の襲撃から守ることができる．村人陣営の役職である．人狼の襲撃先と狩人の護衛先が一致すれば，翌日の犠牲者はいない状況となる．
5. 人狼
   夜のフェーズに 1 人のプレイヤーを襲撃することができる．人狼陣営の役職である．人狼同士は，お互いが人狼陣営であることを把握し合っているので，村人に比べて多くの確定情報を持つことになる．
6. 狂人
   人狼陣営に協力する人間であり，人狼陣営の役職である．人間であるため，占い師からの占い結果は人間判定になり，霊媒師からの霊媒結果も人間判定になる．多くの場合，狂人は自分が占い師や霊媒師であると名乗り出るなどして，村を混乱させるために動く．誰が人狼であるかは把握していない．人狼陣営であるが人間なので，勝利条件のカウント時には人間として数えられる．その性質から，自分が追放されるように議論を誘導し，人狼の勝利に貢献するという戦法を取ることもある．

本研究では人狼知能大会に準拠した，村人 8 人，占い師 1 人，霊媒師 1 人, 狩人 1 人，人狼 3 人，狂人 1 人の 15 人のエージェントの組み合わせでゲームを行った．

2.2 人狼知能プロジェクト

人狼知能プロジェクト[5]とは人狼ゲームを人工知能にプレイさせようというプロジェクトである．人狼知能プロジェクトでは人狼知能のプロトコルとサーバを公開することで，多くの人がエージェントの作成に参加できるようにしている．また，作成したエージェントの強さを競う大会も開催している．人狼知能プロジェクトでは，人狼知能を汎用人工知能の新しい標準問題とすることを目指している[5]．本研究では，人狼知能プロジェクトの提供する人狼知能プラットフォームを使用し実験を行う[6].

2.2.1 人狼知能プロトコル

人間同士が人狼をプレイする場合は，会話には自然言語が用いられる．しかし，強い人狼知能の開発を目指しながら同時にゲーム内で自然言語を扱わせることは難しい．そこで表 1，表 2 のような，人狼知能独自のシンプルな会話プロトコルが用いられている．人間が人狼内で行う会話の半数以上は自分の思考を表現するものであり，それが表 1，表 2 の内容にあたる．残りの 20%は理由説明，15%が雑談，残りの 15%がその他の発言である[7].

表 1　プレイヤーの行動に関するプロトコル[7]

対象指定メソッド	メソッドの内容
vote	投票するプレイヤーを決める
attack	人狼が襲撃するプレイヤーを決める
guard	狩人が防衛するプレイヤーを決める
divine	占い師が占うプレイヤーを決める
会話メソッド	メソッドの内容
Talk	村全体への発話を行う
Whisper	人狼だけに対して発話を行う

表 2　会話メソッドで用いるプロトコル[7]

発話可能な内容	詳細
estimate	他プレイヤーの役職の推定
comingout	自分の役職を公言する
divined	占った結果を伝える
inquested	霊媒した結果を伝える
guarded	護衛したことを伝える
vote	投票したいプレイヤーを伝える
attack	人狼が襲撃したい人に投票する
agree	他プレイヤーの発言に同意する
disagree	他プレイヤーの発言に反対する
over	もう話すことはないとき使用
skip	様子見をしたいとき使用

3. 実験

本研究では，人狼知能 2017@CEDEC2017 決勝進出チームのエージェントである，kasuka，wasabi，AITKN を用いて実験を行う．人間が行っているが人狼知能が行っていないのではないかと考えられる戦法を挙げ，その戦法が実際に行われているかを調査する．行われていなければ，エージェントがその行動を行うようにプログラムを作成し，勝率や信頼度から戦術の評価を行う．

人間がプレイした人狼ゲームにおいては，人狼が占い師であると名乗り，信頼を勝ち取って勝利する例も存在している．このように，実際は異なる役職であるが，自分はその役職であると発言することを「騙る」という．先行研究[8]では人間の対戦ログについて基準を設けて，占い師や占い師を騙る人狼，狂人の信頼度を測っており，人狼や狂人が占い師を騙った場合には本物の占い師に比べて信頼されにくい傾向があることが分かっている．先行研究で用いられた信頼度の基準は 3 つあり，占い師候補の被投票数が 0 であるか，初日の占い結果が人間の場合に判定先のプレイヤーの村人陣営からの投票が 0 であるか，初日の占い結果が人狼である場合に村人陣営からの投票が半分以上あるかである．この条件を 3 つすべて満たした場合は信頼とし，満たさないものがある場合は疑われているとした[8]．本研究では人狼知能同士のエージェントの対戦ログを用い，また，先行研究とは異なった基準での信頼度測定を行う．

3.1 占い師を騙る人狼による身内切り

人間が行っているが人狼知能が行っていないのではないかと考えられる戦法の 1 つに，占い師を騙る人狼による身内切りが挙がる．これは，占い師を騙る人狼が仲間の人狼に人狼判定を出し，それによる周囲からの信頼の上昇を狙う戦法である．身内切りを受けた人狼が追放され，翌日霊媒師に占い師を騙った人狼が正しい占い結果を出していたことを村人に伝えてもらえば，周囲からの信頼度が上昇するのではないかと考えている．この占い師を騙る人狼による身内切り(以後，「身内切り」と表す)という戦法を行っているかについて，kasuka，wasabi，AITKN の役職を人狼に設定して対戦ログを作成し，調査を行った．役職が人狼になったとき占い師を騙っていた kasuka について 3000 件の対戦ログについて調べたところ，4 件については身内切りを行っていることが確認されたが，実行回数が少ないため kasuka に身内切りを実装し，その効果と信頼度の推移について調査を行った．今回は 1 日目に仲間の人狼に人狼判定を出すようにした．

3.1.1 ゲームとエージェントの設定

村人 8 人，占い師 1 人，霊媒師 1 人，狩人 1 人，人狼 3 人，狂人 1 人の 15 人で行う．エージェントの構成は，kasuka(kasukaA，kasukaB，kasukaC も含む)が 5 人，wasabi が 5 人，AITKN が 5 人である．以下の表 3，表 4 のようなグループ A，グループ B のエージェント構成で対戦ログを 3000 件作成した．グループ A の人狼の kasukaA が占い師を騙るが身内切りを行わないエージェントであり，グループ B の人狼の kasukaB が身内切りを行い占い師と信頼勝負を行うエージェントである．グループ B の kasukaC は，信頼勝負の作戦に協力するようにプログラムを変更したエージェントである．

表 3　グループ A のエージェント構成

ログ A	人数	役職指定
kasukaA	1	人狼
kasuka	4	ランダム
wasabi	5	ランダム
AITKN	5	ランダム

表 4　グループ B のエージェント構成

ログ B	人数	役職指定
kasukaB	1	人狼
kasukaC	2	人狼
kasuka	2	ランダム
Wasabi	5	ランダム
AITKN	5	ランダム

kasuka，kasukaA，kasukaB，kasukaC の 4 つのエージェントの情報について，騙る役職，占い先選択，襲撃優先度の 3 つの項目で表 5 にまとめた．狼になったとき何を騙るかが，表 5 の「騙る役職」列である．占い師を騙ったときの占い先の選択基準が「占い先選択」列である．役職が人狼であったとき，どの役職を優先して襲撃するかが「襲撃優先度」列である．kasuka は，人狼になったときにランダムで占い師を騙るか村人を騙るかを決め，占い師になった場合は占い対象として人狼陣営らしくない人を選択する．また，襲撃先は狩人，占い師，霊媒師，村人，人間の優先度で決定する．これを kasukaA，kasukaB，kasukaC についても表 5 のようにまとめた．kasukaA は，元のエージェントである kasuka を必ず占い師を騙るようにプログラムを変更したものである．kasukaB は，kasukaA が初日に身内切りを行うようにし，襲撃の優先度を変更したエージェントである．kasukaC は kasukaB と同様の襲撃の優先度に変更したエージェントである．kasukaB と kasukaC は霊媒師（3 日以降），狩人，村人，人間の順に優先度を変更した．霊媒師の襲撃の優先度を 3 日目以降に変動させた理由は，占い師を騙った人狼が人狼に人狼判定を出していたという事実を霊媒師によって村に発表してもらった後に襲撃させるためである．霊媒師の襲撃の優先度を 3 日目以降に最も高くした理由について説明する．今回使用したエージェントらは狂人も人狼も霊媒師を騙ることがなかったため，ゲーム内で霊媒師は本物であると確定する．そこで本物の霊媒師と違う結果を出すと偽物であることが明らかになってしまい，信頼勝負の土台が整わないことになると考えためである．占い師を襲撃しないようにプログラムを変更した理由は，占い師が襲撃を受けた場合は人狼でないことが確定してしまい，信頼勝負にならないと思わるためである．襲撃を受けた場合人狼でないことが確定するのは，人狼は人狼に対して襲撃を行えないためである．

表 5　エージェント情報

	騙る役職	占い先選択	襲撃優先度
kasuka	占か村人	人狼陣営らしくない人を選択	狩人>占>霊>村人>人間
kasukaA	必ず占	人狼陣営らしくない人を選択	狩人>占>霊>村人>人間
kasukaB	必ず占	初日に仲間の狼へ黒判定，以降は ksuka と同じ	霊 (3 日目～)>狩人>村人>人間
kasukaC	必ず村人		霊 (3 日目～)>狩人>村人>人間

3.1.2　評価基準

本研究では戦術の効果について，「勝率」，「信頼度」，「人狼の生存人数」，「占い師を騙った人狼と占い師の追放回数とその比率」という基準を用いて評価する. 信頼度の基準としては, 狩人からの護衛率と，村人からの投票率を用いて評価を行う．

1. 勝率

   人狼陣営が全ゲーム中どれだけ勝利したかを表す．戦術がどれだけ実用的であるかがよく表れると考えられる．

2. 狩人からの護衛率

   狩人が占い師を護衛したとき，本実験で用いた配役では現時点で一番本物だと思っている占い師を護衛していると考えられる．護衛回数で信頼を測るという手段も考えられるが，エージェントによって後半まで生き残るログの個数に大きな違いがあるときに適切な結果が得られないと判断し，護衛率という基準を考えた．

   この護衛率とは，狩人が生きているかつ自分が生きているという条件の中で，自分を護衛した割合と定義した．これを日付ごとに集計し，狩人からの信頼度の日ごとの推移とした．

3. 村人からの投票率

   村人からの投票数は，村人が誰を疑っているかをよく表すと考えている．そのため（投票してきた村人の人数/生存している村人の人数）の平均を日付ごとに集計し，村人からの信頼度の日ごとの推移とした．

4. 人狼の生存人数

   今回は占いを騙る人狼による身内切りという作戦を実装するため，2 日目の時点で人狼の人数は大きく減少すると考えられる．しかし，それ以降の人狼の減り方についての予測はできない．信頼が取れていたのならば，緩やかな減り方になると考えられる．

5. 占い師を騙る人狼と占い師の追放回数とその比率

   グループ B では人狼陣営が占い師を襲撃しなかったもののみを対戦ログとして集計する．そのため占い師の死亡はプレイヤー全体による追放会議のみで発生する．また，人狼は人狼を襲撃できないため人狼の死亡も追放会議のみでの発生となる．そこで人狼と占い師の追放回数と比率を調査することで，プレイヤー全体の疑いが占い師を騙っている人狼と本物の占い師のどちらに向いているのかが明らかになるはずである．

これら 5 つの基準を用いて，実験 1～実験 5 で戦法の評価を行った．

3.1.3 実験結果

1. 実験 1

実験 1 では評価基準 1 をもとに評価を行う．まず，グループ A とグループ B の対戦ログについて，人狼陣営の勝率の比較を行った．以後，グループ A の対戦ログをログ A，グループ B の対戦ログをログ B とする．表 6 は，勝率についてまとめた表である．ログ A とログ B の対戦ログを 3000 個作成したが，ログ B については 3000 個のログの中から占い師を一度も襲撃しなかったかつ霊媒師への襲撃が 3 日目以降に行われていたもの 2204 個を対象として検証している．これは，占い師と単純な信頼勝負を行った結果についての勝率や信頼度を測るためである．

表 6 グループ A とグループ B の人狼陣営の勝率

	ゲーム数(回)	人狼勝(回)	人狼負(回)	人狼勝率(%)
グループ A	3000	1748	1252	58.3
グループ B	2204	396	1808	18.0

実験 1 の表 6 の結果から，身内切りを行った場合は，勝率が約 40 ポイントと大幅に下がることが分かった．人狼 BBS という人間がプレイする掲示板型人狼ゲームの場合，村人陣営の勝率は約 60%程度になる[6]．グループ A の村人陣営の勝率は 41.7%，グループ B は 82.0%である．グループ A の村人陣営の勝率は人狼 BBS のものより 20 ポイントほど低く，元々人狼陣営が強い傾向があるといえるだろう．しかしその点を考慮しても，この結果のみで言えばグループ B で行った身内切りの戦法の有効性は認められない．

2. 実験 2

実験 2 では評価基準 2 をもとに評価行う．狩人からの護衛率の推移について，ログ A の kasukaA とログ B の kasukaB について求め，図 1 のような折れ線グラフにまとめた．

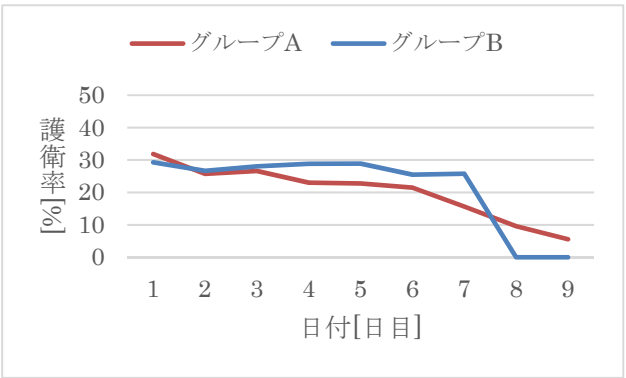

図 1 グループ A の人狼とグループ B の人狼の狩人からの護衛率

1 日目の時点では，グループ A の kasukaA とグループ B の kasukaB の護衛率はほぼ差が無いが，僅かに kasukaA の方が高い．2 日目になると，どちらの護衛率も低下するが，低下量は kasukaB の方が少ない．2 日目からは kasukaA での護衛率の低下傾向がはっきり見られ，次第に狩人からの信頼度は低下していくといえるだろう．それに対してログ B の kasukaB は，やや下がり気味ではあるがほぼ一定を保ち続けている．8 日目に kasukaB の信頼度は 0 に急落しているが，これは 8 日目に到達したかつ狩人が生きているゲームが少なく，正確なデータが取れなかったためだと思われる．ゲームが終盤に近付くと抽出に該当するデータ数が少なくなっていき，後半になるにつれてデータの正確性も低下していくといえる．

また，図 2 は図 1 のグループ B の人狼の護衛率のグラフに占い師と狂人の護衛率を追加し比較を行ったものである．今回使用したエージェントでは狂人が必ず占い師を騙るようになっているので，占い師候補 3 人の信頼度を狩人からの護衛率という基準を用いて比較した．

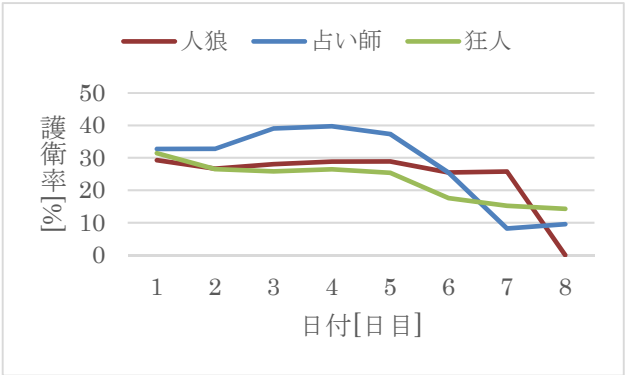

図 2 グループ B の 3 役職の狩人からの護衛率

1 日目の時点で，人狼が護衛される確率は占い師より約 5 ポイント低く，3，4，5 日では 10 ポイント程度の差が生まれる．占い師は 7 日目には護衛率の大きな低下が見られるが，これも人狼と占い師の信頼が逆転したわけではなく，データが少ないために正確な結果が得られなかったためと思われる．そこを除くと，護衛率については占い師の方が高い状態が維持され続けているという結果が得られた．

実験 2 より，図 2 より占い師を騙る人狼は占い師と比較して信頼を得ることが難しいと思われる．しかし，図 1 から，身内切りを 1 日目に行うことで，信頼度の下降が少し軽減されていた．

3. 実験 3

実験 3 では，評価基準 3 をもとに評価を行う．村人からの投票率について推移をグラフにまとめた．図 3，図 4 はそれぞれグループ A とグループ B における占

い師候補 3 人の村人からの投票率をまとめたものである．

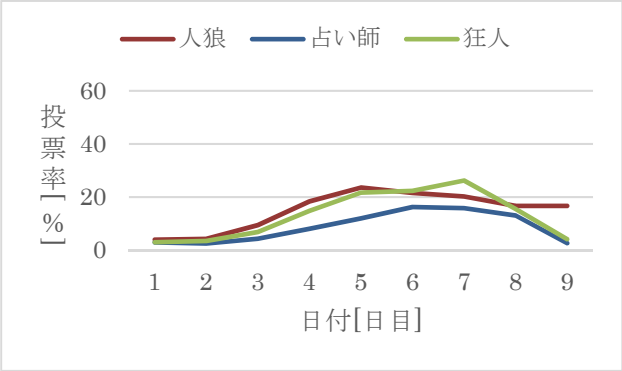

図 3　グループ A の 3 役職の村人からの投票率

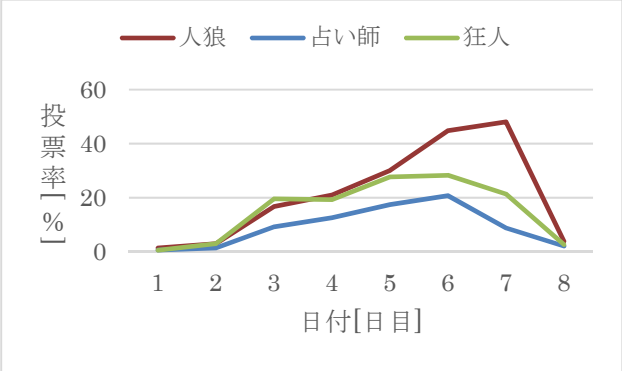

図 4　グループ B の 3 役職の村人からの投票率

図 3 では，人狼は占い師に比べて村人から疑われやすい傾向があることが読み取れる．初日の時点では人狼，占い師，狂人の村人からの被投票率はほぼ変わらない．図 4 のログ B についても，初日はほぼ人狼も占い師も狂人も村人からの被投票率に差はないが，後半になるにつれて投票を受ける確率が大きく上昇していると分かった．図 3 と図 4 を比較すると，身内切りを行った図 4 の方が後半になるにつれての人狼の被投票率の上昇が大きかった．

実験 3 では，ログ A の占い師を騙った人狼と比較して，初日に人狼を言い当てたログ B の占い師を騙った人狼の方が投票を受けやすくなっていた．初日に人狼を見つけた占い師に対してここまで多くの疑いが向けられる理由は不明である．しかし，図 3 と図 4 の占い師と狂人のグラフを比較すると，図 4 のグループ B の方が投票を受ける確率が上昇していた．人狼の村人からの被投票率が上がらず占い師と狂人の被投票率が上がっていたならば，自分以外の信頼を落としたと思われ，効果があったといえる．今回は人狼の被投票率の上昇が非常に大きい．

4.　実験 4

実験 4 では評価基準 4 をもとに評価を行う．図 5 は，ログ A とログ B の人狼の平均生存人数のグラフである．グループ B の占い師を騙った人狼は 1 日目に身内切りを行うので，図 5 のように 2 日目時点で生存平均人数が 2.1 人程度に落ちる．そこから直線的にグラフは下降し，ログ B では 5 日目の時点での人狼平均生存人数が約 1 人となっている．ログ A の人狼が平均生存人数 1 人となるのはログ B の場合の約 2 日後の 7 日目であった．

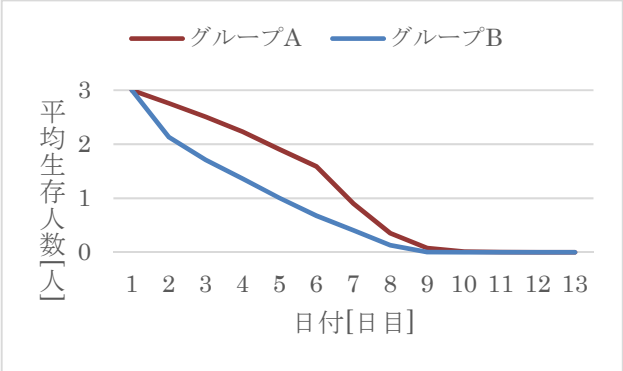

図 5　人狼生存人数の平均

図 5 より，ログ A の 1 日目から 6 日目までの傾きとログ B の 2 日目から 6 日目までの傾きに違いがあることが分かった．図 6 は，ログ A のグラフの 1 日目から 6 日目の線の上から直線を引き，その直線をログ B の 2 日目以降の線に重ねたものである．ログ A の人狼平均生存人数の減り方の傾きよりも，ログ B が傾きの方が大きい．ログ B では 1 日目から 2 日目にかけて人狼の生存人数がほぼ 1 人減る．そこから傾きが緩まることが無く下降し続ければ，敗北に繋がると思われる．平均生存人数という基準で評価を行ったが，信頼が取れていたとはいえない結果になった．

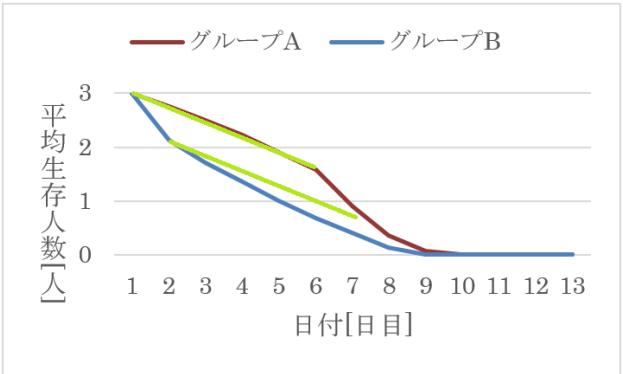

図 6　人狼生存人数の平均 2

5.　実験 5

実験 5 では，評価基準 5 をもとに評価を行う．ログ

B において，日ごとに占い師を騙った人狼と占い師が追放された回数を求めた．表 7 は追放回数を表としてまとめたものであり，図 7 は人狼の追放と占い師の追放が行われる比率の日ごとの推移をグラフにしたものである．

表 7 ログ B における日ごとの追放回数

	人狼	占い師
1 日目	17	18
2 日目	24	31
3 日目	356	215
4 日目	402	241
5 日目	503	236
6 日目	432	187
7 日目	210	38
8 日目	6	1
合計	1950	967

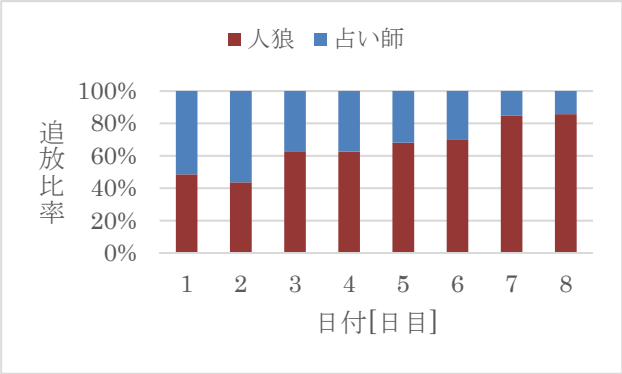

図 7　ログ B における人狼と占い師の追放比率

1 日目においては，人狼と占い師の追放回数は 17 回と 18 回でほぼ同じである．2 日目には人狼の追放率が 50%を下回り，人狼の追放率は 1 日目から 2 日目にかけて減少した．ここに身内切りの効果が表れたと思われる．しかし 3 日目になると人狼の追放率が 60%を超え，以後も上昇が続いていく．2204 個のログの中の 1950 個で人狼の追放が行われ，967 個のログで占い師の追放が行われた．その割合は約 88%と約 44%であり，約 2 倍の差がつくという結果となった．ログ B には人狼が占い師を襲撃したログは含めていないため，人狼が占い師を襲撃したために占い師の追放回数が少ないということは無い．

実験 5 では，日数が経過するにつれて本物の占い師と比べて占い師を騙った人狼の信頼が下がっていくことが分かった．身内切りを行ったとしても，その効果は 2 日目に追放の割合が減少していたことに現れていたが，非常に一時的な信頼を得るのみにとどまった．

3.1.5　実験のまとめ

初日に信頼の差は見られないが，ゲームが後半になるにつれて占い師を騙った人狼の信頼は次第に落ちていく傾向が見られた．また，人狼知能における信頼は一度正しい「あるエージェントが人狼である」という結果を出すだけで取れるものではないかった．身内切りをした直後はその効果も見られたが，以後，その信頼はあまり持続しなかった．以下の表 8 に，実験 1～実験 5 の効果についてまとめた．

表 8　戦法の効果

	効果
実験 1(勝率)	×
実験 2(護衛率)	◎
実験 3(投票率)	×
実験 4(生存率)	×
実験 5(追放回数)	○

実験 1，実験 3，実験 4 の 3 つの基準において負の効果となった．これを記号×で表している．良い効果が見られたものは実験 1 のみであり，これを記号◎で表す．少しの良い効果が見られたものは実験 5 のみであり，これを記号○で表した．表 8 の結果を総合すると，今回実装した戦法について，期待した効果は得られなかったといえる．

4. 考察

4.1 勝率の低下について

実験 1 で勝率が低下した原因について考察する．人狼の人数が減ることは人狼陣営の「村人を人狼と同じ人数まで減らす」という勝利条件から，人論陣営の敗北につながると予測できる．今回調査に使った 15 人中 3 人狼の組み合わせは，追放が可能な回数は基本的には 7 回である．さらに，狩人の護衛成功により追放可能回数が増加することもある．そこで身内切りの戦法を取った場合は，7 回のうちの初回で人狼を追放できるので，あとの 6 回で 2 回人狼を追放すればよい．6 回の追放回数があるのならば，占い師候補 3 人全員と占い師候補のそれぞれの目線での確定人狼 3 人を全員追放することができる余裕がある．それができれば，村人陣営の勝利となる．村人陣営が確定勝利する流れにならないように人狼と狂人は結果を調整すると考えられるだろうが，村人に有利な展開であることは確かである．よって人狼陣営の勝率を下げることにつながったと推測できる．しかし，人間も毎回のようにこの

身内切りという戦術を使うわけではない．もしも人間が毎回この戦術を取るのであれば，勝率が下がるだろうと予測できる．今回は戦術の細かい実行条件などは指定せず，無条件に初日から戦術を実行するとしている点も勝率の低下の原因であると考えられる．

4.2 信頼獲得について

表 8 より，今回の実験では占い師を騙った人狼が信頼を得られたとは言えない結果になったと読み取れる．信頼をあまり得られなかった原因について考察する．人間がプレイした人狼においては，初心者の村人はあまり深読みをしないため，身内切りをしてもそれほど大きい効果が得られない可能性があると考えられる．しかし経験を積むことで「狼ならばこんなリスクのある行動をとらない」と読むことができるようになり，身内切りが効果を発揮することも考えられる．つまり，初心者村人の経験が増えるにつれて効果が上がってくることがあると考えられ，ある時点では身内切りを行った方がより高い勝率となる場合も考えられる．そこからさらに経験が増すことで，再び身内切りの効果が下がることも考えられる．人狼知能はこの戦法においてはいわば初心者であるため，効果が得られなかったのではないかと予測した．

現在，人狼知能は勝ちを目指してプログラムされている．しかし，人間は純粋に価値を目指すのではなく，時には「あっと驚かせたい」という思いから，様々な戦略を試すことがある．そうした理由から人狼知能が経験した戦略の数は多くはないと考えられる．今回実装した占い師を騙る人狼による身内切りという戦法は，「人狼ならばこういう戦法はとらないのではないか」という共通認識が存在した上で大きい効果が得られるものであると考えられる．これから人狼知能が様々な戦法を行うようになり，戦法に対する考えを持つようになることで，今回実装した「相手の考えや認識の裏をかく」ような戦法も使うことができるようになり，通説のみにとらわれない，より人間らしいエージェントができると考えられる．

この「確立されていない」と考えられる共通認識の 1 つの例に，「人間という占い結果が事実だった」と「人狼という占い結果が事実だった」の価値の重さが挙げられるのではないかと予測した．今回，仲間の人狼に人狼判定を出し追放させ，霊媒師によって本当に人狼であったことが明らかになったときには少しの信頼度向上が見られた．しかし，その効果は想定よりも小さいものだった．人間の思考では，多くいる人間を人間であると結果を出してそれが事実だったときと，少ない人狼を人狼であると結果を出してそれが事実だったときでは，後者の方がより本物の占い師のように見えると考えられる．したがって，「あるプレイヤーに対して人狼判定を出して霊媒師と結果が繋がったという事実に対する周囲のプレイヤーからの評価」が「霊媒師と結果が同じだった」という加点のみで，それに加えて「人狼に黒判定を出した」という重みがかかっていない可能性があるのではないかと考察した．このことから，「人間という占い結果が事実だった」と「人狼という占い結果が事実だった」という結果の後者の方の価値が大きいことが共通認識として確立されれば，より今回実装した戦術の効果が増すのではないかと考えられる．

4.3 日数経過に伴う信頼低下

今回の実験で信頼度が日数の経過につれて低下した原因について考察する．ゲームが終盤に近付くと，人狼は村人の中に仮想の狼を作り出さなければならない．しかし，仮想狼にされた村人からは，占い師を騙っている狼が自分を狼に仕立て上げてきた偽占い師と分かってしまう．そのため村人からの投票率という基準で信頼度を測るならば，信頼度は日が経つにつれて低下すると考えられる．占い師を騙る人狼が作った仮想狼の村人は，当然敵対関係となる．自分を疑っている村人はできるならば村に居てほしくないが，狼扱いしている村人を襲撃することでその村人が人狼でないことが全てのプレイヤーの視点で明らかになり，占い師を騙る人狼が本物の占い師でないことがプレイヤー全員に知られてしまうことがあり得る．また，狩人に対して人狼判定を出してしまった場合には，狩人が狩人であると公言することで，占い師を騙る人狼が本物の占い師でないことが全てのプレイヤーに明らかになってしまうことがある．そのため，今回設定した基準においては，日が経つごとに信頼度が低下していったと考えられる．本実験で見られた日数経過に伴う信頼低下は，人間が行う人狼ゲームにおいても起こると推測できる．

5. 結言

本論文では，人間が行うが人狼知能が行わない戦法を人狼知能に実装した．また，その効果について勝率と自分で定義した信頼度の基準のもとで評価を行った．結果，勝率は大幅に下がったが，少しの信頼は獲得できていた．「だます」，「説得する」という行為は，共通認識を利用することでより効果が増すことがある．人狼知能エージェントは様々な戦法の経験をしていないため，戦法に対する認識が確立していない．したがって現在の人狼知能に対しては人間と同じ戦法をしたとしても，同様の効果は得られなかったのだと考えられる．人狼知能がさまざまな戦法を行うようになり，戦法に対する共通認識が確立されたとき，現在よりも効果が上昇することが期待できるのではないかと考えられる．

謝辞

本研究に際して，様々なご指導を頂きました西村俊二講師に深謝いたします．また，この研究の機会をくださった情報工学科の先生方，そして多くの知識を下さいました西村研究室の先輩，同期の皆様に厚く御礼申し上げます.

参考文献

[1] 鳥海不二夫・狩野芳伸・大槻恭士・園田亜斗夢・中田洋平・箕輪峻 : 人狼知能で学ぶ AI プログラミング，株式会社マイナビ出版，2017

[2] 人狼知能プロジェクト，人狼知能大会，http://aiwolf.org/aiwolf_contest，[アクセス日： 21 1 2019].

[3] 伊藤幹太，Reijer Grimbergen : 人間同士の人狼ゲームで用いられる戦術を反映させた人狼知能の研究, ゲームプログラミングワークショップ 2017 論文集（2017）

[4] 神田直樹，伊藤毅志 : 人狼サーバによる自動対戦を用いた通説の検証~人狼は占い師を騙るべきか~，ゲームプログラミングワークショップ 2015 論文集(2015)，pp.20-24

[5] 片上大輔，鳥海不二夫，大澤博隆，稲葉通将，篠田孝祐，松原仁 : 人狼知能プロジェクト(<特集>エンターテイメントにおける AI)，30 巻 1 号 pp65 - 73，2015

[6] 人狼知能プロジェクト，人狼知能プラットフォーム，http://aiwolf.org/server，[アクセス日：11 10 2018].

[7] 鳥海不二夫・片上大輔・大澤博隆・稲葉通将・篠田孝祐・狩野芳伸 : 人狼知能 ーだます・見破る・説得する人工知能ー，森北出版株式会社，2016

[8] 園田 亜斗夢，鳥海 不二夫 : 人狼ゲームにおける信頼の分析，2017 年度人工知能学会全国大会（第 31 回）（2017）