因果推論を用いた人狼知能プロトコルによる返答生成について

Response Generation by AI Wolf Protocol Using Causal Inference

福井敬徳[1*]　川部勇太[1]　野々山幾也[1]　岩田員典[2]　伊藤暢浩[1]
Takanori Fukui[1] Yuta Kawabe[1]　Ikuya Nonoyama[1]　Kazunori Iwata[2]　Nobuhiro Ito[1]

[1] 愛知工業大学
[1] Aichi Institute of Technology
[2] 愛知大学
[2] Aichi University

Abstract: In the recent years, the evolution of artificial intelligence (AI) has influenced the development of human-interactive communication games. The Werewolf game was originally a real-world, face-to-face, indoor game played between a minimum of four players. AI Wolf is an academic project of such Werewolf game. In the project, a player of AI Wolf has been developed for a few years. But, conversations among players is not enough to find something important from the conversations and speak new one based on them. In this paper, we propose a generating method of a response for an AI Wolf speech in the AI Wolf protocol. Furthermore, we evaluated our method through some experiments. As a result, we confirmed that our method could generate a correct response in some games.

1　はじめに

近年，人狼ゲームをプレイできる人工知能の研究が盛んにおこなわれている．人狼ゲームをプレイする人工知能を人狼知能プロジェクトでは「人狼知能」と呼んでいる．

人狼知能プロジェクトとは，人狼知能の構築を目指すプロジェクトである [1]．このプロジェクトでは，人狼ゲームを計算機上でプレイすることができる人狼知能プラットフォームが提供されている．また，人狼知能プラットフォームでは人狼ゲームをプレイする「人狼知能エージェント」（以降，エージェント）の開発環境も備えている．さらに，エージェントの強さを競うために人狼知能大会が毎年開催されている．

一般的に，人狼ゲームとは，複数人のプレイヤが対面しておこなうゲームとして知られている．プレイヤは村人陣営と人狼陣営に分かれ，村人と人狼は互いの排除を目指す．誰が人狼であるかは不明である．そのため，会話によって正体を探る必要がある．

会話は，会話の起点となる発話とその発話に対する返答に分類することができる．会話は 1 つの発言に対して複数の返答により構成されるため，適切な返答は会話において円滑な話し合いをするために必要である．しかし，現在の人狼知能大会では，会話の起点となる発話は多いが，発話に対する返答が適切にされていないという問題がある．この問題により，コミュニケーションゲームとしての側面が大きく損なわれていると考える．

本研究では，エージェント同士が会話をおこなうために，正しい返答を生成するモデルの検討と作成を，因果推論を用いておこなう．因果関係を求める要因を人狼知能大会のログから抽出する．また，因果関係があるかを統計的な指標を用いて確認する．

作成した正しい返答を生成するモデルをエージェントに組み込み，他のエージェントと対戦させることで，会話が成り立つような返答ができているかを確認し，モデルの評価をおこなう．また，返答の有無や適切な返答が，勝率に影響があるか確認する．

本研究では，第 3 回人狼知能大会に出場したエージェントと対戦をおこない，モデルが適切な会話をしているかを確認し，勝率に影響を与えるかも確認した．勝率は，返答をおこなわない場合と比較して，モデルを搭載したエージェントの方が高いことを確認した．

*連絡先：愛知工業大学大学院
愛知県豊田市八草町八千草 1247
E-mail: itfukui0922@icloud.com

2　人狼知能プロジェクト

2.1　人狼知能プロジェクトとは

人狼知能プロジェクトとは，人狼ゲームをプレイする人工知能の構築を目指すプロジェクトである．同プロジェクトは高度な知能の創出，および人と人工知能との高度なコミュニケーションを実現するために，人と自然なコミュニケーションをとりながら人狼ゲームを楽しむことができる人狼知能の構築を目指している [1]．ここで，人狼ゲームをプレイする人狼知能を人狼知能エージェント (以降，エージェント) と呼ぶ．

2.2　人狼ゲーム

人狼ゲームとは，アメリカのゲームメーカー Lonny Labs. が 2001 年に発売したパーティゲーム「汝は人狼なりや？」[2] およびその派生ゲームの総称である．

一般的に人狼ゲームとは，複数人のプレイヤが対面しておこなうゲームとして知られている．プレイヤは村人陣営と人狼陣営に分かれ，村人と人狼は互いの排除を目指す．村人陣営のプレイヤは誰が人狼であるかを知らない．そのため会話によって正体を探る．

ゲーム上では昼のフェーズと夜のフェーズがあり，2 つのフェーズを合わせて 1 日とする．村人陣営か人狼陣営のどちらかが勝利するまで何日間もおこなわれる．

昼のフェーズでは，会話によって人狼だと思われるプレイヤを投票により 1 名決定し，ゲームから排除する．この排除のことを追放と呼ぶ．夜のフェーズでは，人狼が村人から 1 名決定し，ゲームから排除する．この排除のことを襲撃と呼ぶ．また夜のフェーズでは，役職によって決められた能力を行使する．

村人陣営の勝利条件はゲーム上から人狼を追放することである．反対に，人狼の数と人間の数が同数となった場合，人狼陣営の勝利となる．どちらかの陣営の勝利条件を満たすことで，ゲームは終了する．

ゲームを開始する時にひとりのプレイヤに一つずつ，役職が決められる．以下にその役職と役職が持つ特殊な行動・能力について述べる．

村人陣営

村人 ： 特別な能力のない役職である．

占い師 ： 1 日のうちに 1 人だけ相手が狼であるかを知ることができる．

霊能者 ： 前の昼のフェーズに追放したプレイヤが人間であったか，人狼であったかを知ることができる．

狩人 ： 1 日のうちに 1 人だけ他のプレイヤを人狼から守ることができる．

狼陣営

裏切り者 ： 占い師や霊能者の結果からは人間だと判定されるが，この役職は人狼側の勝利条件で自分自身も勝利する．

人狼 ： 1 日のうちに 1 人だけプレイヤを襲撃することができる．また，人狼同士のみの会話をすることができる．

これらの役職の中で，人狼のみが，他のプレイヤの中で誰が人狼であるかを知っている．

2.3　人狼知能プラットフォーム

人狼知能プラットフォームとは，人狼ゲームを計算機上でプレイすることができる Java のパッケージである．また，人狼知能プラットフォームでは人狼ゲームをプレイする「人狼知能エージェント」（以降，エージェント）の開発環境も備えている．

人狼知能プラットフォームは，クライアント・サーバモデルとして構成されている．クライアントはエージェントの各行動に対する意思決定を行い，サーバは投票や襲撃といったシステム的な処理をおこなう．また，処理の内容をログデータとして出力する．

人狼知能プラットフォームを用いることで，エージェント同士のゲームプレイが可能になる．また，人狼知能プロジェクトが開催する人狼知能大会では，実際にエージェントの強さを競うために人狼知能プラットフォームが利用される．

人間同士がおこなう人狼ゲームにおいて，プレイヤの行動や発話の全てを計算機上で表現することは非常に困難である．そのため，人狼知能プラットフォームでは人狼知能プロトコルというプレイヤの行動や発話を抽象化したプロトコルが定義されている．

人狼知能プラットフォームでは，次の手順でゲームが進行する．

1. 参加するエージェントに役職が割り振られる．

2. エージェント同士が会話をおこなう．会話はターン制によって進行し，同ターンの発話は同時に発話されたものとなる．また，各エージェントは 1 ターンに 1 回，発話することができる．

3. 会話終了後，エージェントによる多数決によって追放するエージェントを 1 人決定し，追放されたエージェントをゲームから排除する．

4. 人狼の役職が振られたエージェントによる多数決によって襲撃するエージェントを 1 人決定する．襲撃されたエージェントをゲームから排除する．

5. 2~4 を繰り返し，人狼が全員追放された時点で村人の勝利，村人と人狼が同数となった時点で人狼の勝利となる．

2.4 人狼知能大会

人狼知能プロジェクトの目的のための 1 つの活動として，エージェントの強さを競う人狼知能大会が毎年開催されている．

また，人狼知能大会では，5 体のエージェントが参加するゲームと 15 体のエージェントが参加するゲームがある．それぞれ 5 人人狼，15 人人狼と呼ぶ．

2.5 人狼知能大会参加エージェントの共通問題

現在，開発されているエージェントは，会話において返答が重要視されてない問題がある．図 1 に示すように，会話は 1 つの話題に対して複数の返答によって構成されており，返答が大部分を占めている．

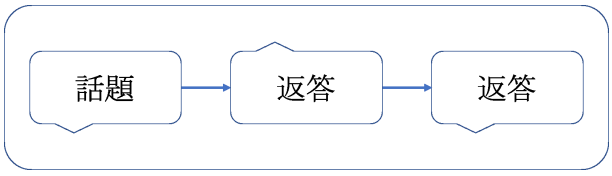

図 1: 会話の構成

ここで，話題とは会話の起点となる発話のことである．また，返答とは話題や他の返答に対しての発話である．

現状の人狼知能大会では，話題は多いが，返答が少ないため，会話がされているとは言えない．

この問題により，コミュニケーションゲームとしての側面が大きく損なわれている．

3 因果推論を用いた返答生成

3.1 本研究の目的

本研究では，人狼知能の実現に向け，2.5 節で示した会話において返答が重要視されていない問題に着目する．現在の人狼知能大会では，各エージェントが自分の意見を述べるだけであり，返答をおこなっているとは言えない．エージェントが自然な会話をおこなうためには，適切な返答をする必要があると考える．そこで，本研究では，相手の発話に対して自分の発話を決定する返答生成モデルを検討，作成する．

3.2 提案手法

本研究で提案する返答生成モデルは，図 1 で示した話題または返答に対して，共変量を観測し，返答となる発話を返すモデルである．ここで，共変量とは，因果推論において原因と結果の双方に影響を与える要因のことである [3]．以降，説明の便宜上，相手の発言である話題または発話を「話題」，返答生成モデルが出力する発話を「返答」として話を進める．

本研究で作成する返答生成モデルを黒木学（2017）[4] を参考に構造化方程式を用いて式 (1) のように定義する．

$$(X,Y,Z)=\begin{cases} x_i = g_x(\boldsymbol{z}_i, \epsilon_{x_i}) \\ y_i = g_y(x_i, \boldsymbol{z}_i, \epsilon_{y_i}) \\ \boldsymbol{z}_i = g_z(\epsilon_{\boldsymbol{z}_i}) \end{cases} \quad (i=1,2,...,n)$$

$$\boldsymbol{z}_i = \{z_1, z_2, ..., z_j\}\ (j=1,2,...,m)$$

ここで，X は話題となる発話，Y は返答となる発話，Z は共変量を示す．また，g は左辺と右辺が関数関係にあることを示す．

ϵ は X, Y, Z それぞれに影響を与える可能性のある未観測の要因である．また，n は返答生成モデルの作成手順で示したステップ 5 にて抽出された因果関係の個数である．さらに，m は返答生成モデルの作成手順で示したステップ 4 で選択された共変量の個数である．

本研究では，以下の手順で返答生成モデルの作成をおこなう．

1. 因果関係を求める発話の候補を人狼知能プロトコルをもとに作成する．

2. 作成した発話の候補をラベルとして，ログデータ上の発話に対してラベル付けをおこなう．

3. ログデータ上で出現頻度が高かった発話を因果関係を求める発話とする．

4. 本研究で扱う共変量を対戦中から得られる情報より選択する．

5. 全ての因果関係と共変量の組み合わせから有意であるものを抽出する．

6. 有意と判断された組み合わせを用いて返答生成モデルを作成する．

因果関係を求める発話と共変量は，人狼知能プロトコルと，対戦中に得られる情報をもとに定義する．

対象とするログデータは第 3 回人狼知能大会の決勝戦の 15 人人狼のログデータである．ただし，対象とする発話に関しては役職に関係なく全プレイヤーが取得できる通常発話とし，人狼同士のみの会話は対象としない．このログデータに対し，前手順で定義したラベルを貼り付ける．

組み合わせ可能な因果関係と共変量の組みから統計的な指標をもとに有意となる組み合わせを抽出する．また，抽出した組み合わせを返答生成モデルとする．

4 返答生成モデル

4.1 因果関係を求める発話

因果関係を求める発話を決めるため，ログデータ上の発話に対してラベル付けをおこなう．説明の便宜上，因果関係の原因を原因変数，結果を結果変数とし，話を進める．

対象とするログデータは第 3 回人狼知能大会の決勝戦，15 人人狼の 37,000 ログとする．本研究では，人狼知能プロトコルを参考に，ラベルを定義した．また，各ラベルをログデータ上の発話に対してラベル付けした結果の発話割合を算出する．定義したラベルを表 1 に示す．また，全発話の合計発話回数は，8,914,508 回であった．

表 1 より，発話割合が高い発話を原因変数と結果変数の対象ラベルとする．発話割合が高い発話を対象とした理由は，発話割合が高いほど，データ数が多いため，影響を計測しやすいと考えたためである．また，発話割合は低いが人狼ゲームにおいて重要とされる DivinedWolf も対象とする．

本研究で原因変数と結果変数となるラベルと対象ログ中の各ラベルの発話割合を表 2 に示す．

表 1: 第 3 回人狼知能大会での発話割合

ラベル名	説明
ComingoutVillager	村人であると役職公開
ComingoutSeer	占い師であると役職公開
ComingoutMedium	霊能者であると役職公開
ComingoutBodyguard	狩人であると役職公開
ComingoutPossessed	裏切り者であると役職公開
ComingoutWolf	人狼であると役職公開
VoteChange	前回と異なる投票先へ投票
VoteSame	前回と同じ投票先へ投票
EstimateVillager	村人と予想する
EstimateSeer	占い師と予想する
EstimateMedium	霊能者と予想する
EstimateBodyguard	狩人と予想する
EstimatePossessed	裏切り者と予想する
EstimateWolf	人狼と予想する
Divination	占う対象の表明
DivinedHuman	占い結果が人間
DivinedWolf	占い結果が人狼
IdentifiedHuman	霊能結果が人間
IdentifiedWolf	霊能結果が人狼
Guard	護衛対象の表明
Guarded	護衛
Attack	襲撃対象の表明
Agree	同意
Disagree	反対
RequestVote	投票の要請
RequestDivination	占い先の要請
RequestOther	その他の要請

表 2: 使用する原因変数と結果変数

データ名	発話割合 (%)
DivinedWolf	0.934
VoteChange	43.386
VoteSame	16.357
EstimateWolf	13.497
EstimateVillager	9.344

4.2 共変量の定義

共変量は対戦中に全てのエージェントが知ることができる情報から選択する．ただし，本研究では，原因変数となる発話以前の情報を扱わないため，エージェントの発話履歴など，時系列を含む情報は除外する．

本研究で定義した共変量とその共変量が取りうる値の範囲を表 3 に示す．原因変数となる発話を発見した際に，その原因変数に関して共変量となるデータを抽出する．

また，共変量はゲーム中にすべてのエージェントが知ることができる情報から選択している．これは，返答生成モデルが役職に関係なく適用可能とするためである．

表 3: 使用する共変量

データ名	説明	値
Day	発話の日付	0 ～ 13
Turn	発話のターン	0 ～ 19
AliveNum	発話の時点で追放されていないエージェントの数	1 ～ 15
CoCount	発話の前日までに，占い師または霊能者であると役職公開をしたエージェントの数	0 ～ 15
DiscordVoted	発話の発話者エージェントが，前日までに発話した最終投票先と実際の投票先が不一致だった回数の累計	0 ～ 12
MaxVoted	発話の投票先エージェントが，発話時点で最も投票が集まっているエージェントかどうか	0 or 1
TalkerVoted	発話の発話者エージェントの前日の得票率	0 ～ 1
TargetVoted	発話の投票先エージェントの前日の得票率	0 ～ 1
TargetWin	発話の発話先エージェントの勝率	0 ～ 1

4.3 因果関係と要変量の組合せの抽出

第 3 回人狼知能大会のプロトコル部門決勝戦の 15 人人狼より，ランダムに選んだ 1,000 ログを対象に抽出する．

原因変数や結果変数，共変量の抽出の流れを図 2 に示す．

1. 原因変数となる発話を探索する．
2. 原因変数となる発話から共変量を測定する．
3. 次ターンの発話の中で，結果変数となる発話を探索する．
4. 結果変数を発見した際，原因変数や結果変数，共変量を抽出する．
5. 1~4 を繰り返す．

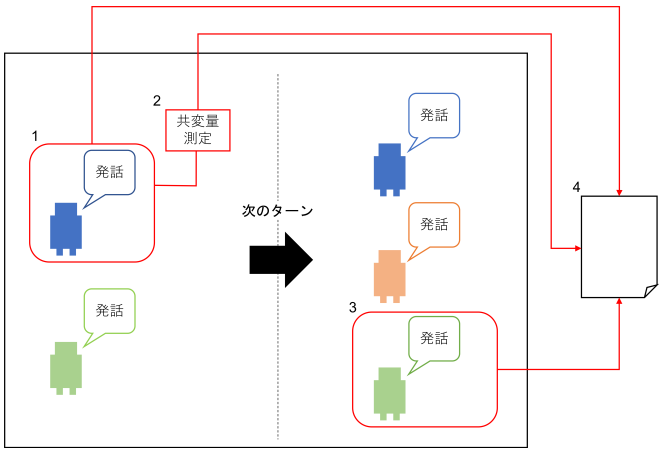

図 2: 抽出の流れ

5 実験と考察

5.1 実験の目的

本実験の目的は，原因変数と結果変数，共変量の組み合わせから優位となる組みを探し出すことである．また，これをエージェントに組み込んで正しい返答をしているか確認する．さらに，対戦させた勝率を測ることによって返答生成モデルが勝率に影響を与えるかを確認する．

5.2 因果推論による分析結果

原因変数と結果変数．共変量の組合せの中から c 統計量が高いものを表 4 に示す．c 統計量とは，組合せの適合度を数値化したもので，0 から 1 の値の範囲をとる．本研究では，大林準（2016）[5] の論文を参考に，0.7 以上のものを有意とする．

また，因果効果を IPW 推定量を求めることで判断する．IPW 推定量とは，因果効果の強さを数値化したものであり，-1 から 1 の範囲をとる．原因変数と共変量から結果変数が生起する場合を 1，しない場合を-1 とし，0 は因果効果がないことを示す．

分析した結果より，c 統計量が有意であった組み合わせと IPW 推定量を表 4 に示す．

表 4: 返答生成モデルに用いる因果関係

	原因変数	結果変数	共変量	c 統計量	IPW 推定量
1	DivinedWolf	EstimateWolf	Turn, MaxVoted, TargetWin, AliveNum, DiscordVoted, CoCount, TalkerVoted, TargetVoted	0.81	0.34
2	EstimateVillager	EstimateWolf	Turn, MaxVoted, AliveNum, DiscordVoted, TargetVoted	0.82	-0.32
3	EstimateWolf	EstimateWolf	Turn, MaxVoted, TargetWin, AliveNum, DiscordVoted, CoCount, TalkerVoted, TargetVoted	0.80	0.20
4	DivinedWolf	EstimateVillager	Turn, MaxVoted, TargetWin, AliveNum, DiscordVoted, CoCount, TalkerVoted, TargetVoted	0.81	-0.23
5	EstimateVillager	EstimateVillager	Turn, MaxVoted, AliveNum, DiscordVoted, TargetVoted	0.82	0.45
6	EstimateWolf	EstimateVillager	Turn, MaxVoted, TargetWin, AliveNum, DiscordVoted, CoCount, TalkerVoted, TargetVoted	0.80	-0.12

IPW 推定量が正の値の場合は，結果変数を返答として用い，負の値の場合は，結果変数を返答しないとする．表 4 に示した 6 つの因果関係を返答生成モデルとする．

5.3 基準値

作成した返答生成モデルをもとに返答をおこなうか，おこなわないかを決める基準値を定義する．本研究では，基準値を式 1 のように定義した．ここで，IPW は因果効果，b_0 は c 統計量を算出したときのロジスティック回帰の切片，b_i は c 統計量を算出したときの各共変量に対するロジスティック回帰の各予測値，Z_i は各共変量の値，N は共変量の個数である．

$$\text{基準値} = |\text{IPW}| \times (b_0 + \sum_{i=1}^{N} b_i Z_i) \tag{1}$$

原因変数に対応する発話に対して，複数の結果変数が対応する場合は，基準値の高い方を選択する．

返答生成モデルが返答を生成する流れを以下に示す．

1. 相手の発話を認識する．
2. 相手の発話が原因となる発話であるか判断する．
3. 原因となる発話である場合に，モデルを元に返答となる発話を選択する．
4. 原因となる発話と 3 で選択した発話に対して，因果効果と共変量を元に実際に発話するか基準値と比較する．
5. 基準値を超えた発話を返答として生成し，発話する．

5.4 実験エージェントの作成

返答生成モデルを搭載したエージェントが勝率に対して影響を与えるのかを確認するために 3 種類のエージェントを作成した．

1 つ目は返答生成モデルを搭載したエージェント（以降，返答生成モデルエージェント），2 つ目はランダムに発言を返すだけのエージェント（以降，ランダムエージェント），3 つ目は発言を一切おこなわないエージェント（以降，返答なしエージェント）である．

5.5 対戦環境

本実験では，第 3 回人狼知能大会の参加エージェントの 14 体と実験エージェントによる 15 人人狼をおこなう．実験エージェントの役職は村人に固定し，100 ゲーム 1 セットとして 100 セット対戦をおこなった．

5.6 対戦結果

実験エージェントの対戦の結果を表 5 に示す．

表 5: 対戦結果

実験エージェント	勝率
ランダムエージェント	61.4
返答生成モデルエージェント	59.4
返答なしエージェント	53.7

5.7 返答生成モデルの考察

本実験で作成した返答生成モデルは，表 4 より，相手の発話に同調する返答は生成するが，相手の発話に反駁

する返答は生成しないモデルといえる．例えば，同調する例として，原因変数が EstimateVillager と，結果変数が EstimateVillager のとき IPW 推定量が正である．また，原因変数が DivinedWolf と結果変数が EstimateWolf のとき IPW 推定量が正である．反対に，反駁する例として，原因変数が EstimateVillager と，結果変数が EstimateWolf のとき IPW 推定量が負である．また，原因変数が DivinedWolf と結果変数が EstimateVillager のとき IPW 推定量が負である．

本研究では基準値を定義して，発言の有無を決めていた．しかし，原因変数に対し，結果変数が複数ある場合でも，同ターンに発話された場合，同じ結果変数を選択し，返答してしまうため，返答が単調なものになってしまっていた．

人狼ゲームが時系列に沿って推論立てていくゲームであるため，原因変数となる発話以前の情報が結果変数を生起させると考えられる．しかし，本研究では，原因変数となる発話以前の情報を考慮していない．そのため，未観測の要因 ϵ による影響が大きく，c 統計量の値が高い返答ルールが多く抽出できなかったのではないかと考える．

今後の課題として，様々なログデータを対象とすることや，原因変数となる発話以前の情報，発話以外の要因についての検討が必要であると考えられる．

5.8 実験エージェントの考察

ログデータより，実験エージェントは，原因となる発話に対して返答をおこなっていることを確認した．

一方で，原因となる話題や返答が多数存在する場合に，返答生成モデルのみでは適切な返答を選ぶことができないことがあった．これは，実験エージェントが適切な推論をせずに，基準値のみで返答の有無を決定しているため，適切な返答を選ぶことができていないと考えられる．また，返答を行うことが勝率に影響を与えることを確認した．

返答生成モデルエージェントは同調する返答のみをおこなう．そのため，適切な返答をすることができず，矛盾した発言をする場合が見られた．このため，他のエージェントからの信頼が得られず，勝率の低下を招いたと考えられる．

反対にランダムエージェントの勝率が高かった要因として，明らかに矛盾した発言をせず，また，適当に発言をすることによって，他のエージェントから投票を受けなかったと考えられる．その結果，人狼に投票する可能性が高くなり，全体として勝率が高くなったと考えられる．

6 おわりに

本研究では，因果推論を用いて，人狼知能プロトコルによる返答生成モデルの検討をおこなった．

本モデルを用いることで，会話として成立するような返答ができると考えられる．

しかし，解析元のデータによっては，返答となる発話に偏りができることがわかった．また，返答生成モデルエージェントは 1 つの話題に対して 1 つの返答をおこなうだけである．これは会話としての最低条件を満たしているだけで，円滑な会話をしているとは言えない．

今後の課題として，他のログデータで解析をおこなうことや，本研究で対象にしなかった発話について同様の検討をおこなうことが望まれる．

謝辞

本研究は JSPS 科研費 JP16K00310，JP17K00317 の助成を受けたものです．

参考文献

[1] 狩野 芳伸, 鳥海 不二夫, 片上 大輔, 大澤 博隆, 稲葉 通将, and 篠田 孝祐. **人狼知能 - だます・見破る・説得する人工知能-**. 森北出版株式会社, 2016.

[2] Are you a werewolf? — looney labs. http://www.looneylabs.com/games/werewolf.

[3] 星野 崇宏. **調査観察データの統計科学 - 因果推論・選択バイアス・データ融合**. 岩波書店, 2009.

[4] 黒木 学. **構造的因果モデルの基礎**. 共立出版, 2017.

[5] 大林 準. ロジスティック回帰分析と傾向スコア (propensity score) 解析. **天理医学紀要**, 19(2), 2016.