集団規範アウェアなエージェントが人間と共存する社会に関する合意形成シミュレーション

Multiagent-Based Simulation of Human Society Including Group Norm-Aware Agents

布施 陽太郎 [1] *
Yotaro Fuse[1]

[1] 富山県立大学
[1] Toyama Prefectural University

Abstract: 本研究では集団規範アウェアなロボットが人間社会に参加する影響をシミュレーションで調査する．先行研究は単一集団内のエージェントの行動や影響に焦点を当てていたが，本研究は複数の集団で構成される社会への適応による影響を検討する．結果から，個々のロボットによる影響力が人間同士の影響力よりも小さい場合でも，集団規範を考慮するロボットは社会全体の合意形成を促進することが明らかになった．また，他者の意見を受け入れにくい社会では，人間ロボット社会で形成される多数派意見が人間だけの社会の場合と異なることが確認された．

1 はじめに

今日，アシスタントやパートナーとしてロボットが人間と共生する社会を目指し，人間とロボットの間のやり取りに関する様々な研究が取り組まれている [1, 2]．社会性とは集団を作って生活しようとする性質である．特に，人間集団において人間は直接的なやり取りや指示無しに，その場の状況に応じて集団の一員である他者に特定の振舞いを期待することがある [3]．そのようなとき，その集団では集団規範が共有されているといえる．人間の社会ではその場で求められる振る舞いは明文化されず，しばしば暗黙的に共有される．その規範に適応できるかどうかは，ある集団メンバが集団の一員としてふさわしいか否かの判断のための基準のになり得る．直接的なやり取りが無いシチュエーションでも，暗黙的に形成される集団規範に適応するエージェントやロボットの研究が取組まれている [4, 5]．

一般に人間社会は複数の小集団によって構成され，しばしば各個人は複数の様々な小集団に所属する．すなわち，個々人が所属する集団への適応は個々人の行動様式へ影響を与え，その個々人が属する異なる別の集団での振舞い方に影響を与え得る．CASA 理論 (Computers Are Social Actors) や Media Equation は人間がコンピュータ，ロボット，その他のデジタルデバイスを社会的な存在とみなしたり，メディアやそのコンテンツを実際の人間や社会的な相互作用と同じように扱ったりする傾向があることについて指摘している [6, 7]．どちらも計算機エージェントであるロボットが人間とロボット間のコミュニケーションにおいて，ロボットの意見や行動が人間の意見形成や決定に影響を与える可能性を示唆している．したがって，人間とロボットが共生する社会を想定したとき，ロボットと人間の間に社会的影響関係が発生し得る．具体的にはナッジの概念を考慮に入れ，ロボットやエージェントを起点として人間の振舞い方に影響を与えることについての検討がなされている [8, 9]．

加えて，AI 倫理の観点からロボットの振る舞いが人間の意思決定や考えに影響を与えることを前提とし，ロボットが規範に従って意思決定することについての問題点が指摘されている [10]．特に，ロボットが準拠する規範が時代遅れになることやロボットによる共有されている規範の変容が起こることが定性的に指摘されている．したがって，個々のロボットの振る舞いが人間社会全体に影響を与える可能性が定性的に示唆されているといえる．

先行研究 [4, 5] では，人間とエージェントの集団において意思決定モデルに従うエージェントが表出する振舞いやその集団全体で起きる現象を観測するのみであった．したがって，エージェントにとっての集団の外を設計し，個々の集団への適応が多数の集団によって構成される社会がどのような影響を受けるのかをこれまでのマルチエージェントシミュレーション分野の知見に基づくシミュレートを試みる．人間を模したエージェントと集団規範アウェアなエージェントで構成さ

*連絡先：富山県立大学工学部知能ロボット工学科
〒 939-0398 富山県射水市黒河 5180
E-mail: fuse@pu-toyama.ac.jp

れる社会をシミュレートすることによって，ミクロである集団の規範に適応するエージェントによる社会というマクロへの合意形成への影響を分析する．社会を構成する人間がロボットなどのエージェントから受ける影響を定量的にシミュレーションすることは，他者への同調という人間が持つ特性を含めたエージェントデザインの土台構築に貢献することが期待できる．

2　方法

2.1　シミュレーション概要

人間を模した多数のエージェント (HA: Human Agent) とロボットの代わりとして集団規範アウェアな意思決定をする多数のエージェント (RA: Robot Agent) が混在する環境を設定し，意見合意形成に関するシミュレーションを実施する．0 から 1 の連続値をとる意見 $x \in [0.0, 1.0]$ を各エージェントが持ち，ステップごとにエージェントが他のエージェントとインタラクションする際にある条件に基づいて両エージェントがそれぞれの意見 x を更新する．人間エージェント HA は合意形成シミュレーションで知られる Deffuant モデルで用いられる意思決定方法を，ロボットエージェント RA は我々の先行研究の意思決定手法を採用する．以上のシミュレーションによって，人エージェント HA とロボットエージェント RA で構成される社会における意見のダイナミクスをマルチエージェントベースで定量的に分析することを目的とする．

本研究ではロボットから人間への影響力に関するパラメタを変化させることによる社会での合意形成への影響を調査する．人間同士はお互いに意見について影響しあうが，ロボットから人間への影響力（影響係数）はどの程度の大きさになるのかは，ロボットの外見や知性のデザインに依存すると仮定する．したがって，人間から人間への影響力を基準にして，ロボットから人間への影響力は同程度かそれよりも小さいものと仮定する．これによって，ロボットがどの程度の人間の意見への影響力を持つと，社会全体での意見収束や合意形成に影響を与えるのかを調べることにつながる．また，合意形成シミュレーションで用いられる Deffuant モデルで一般的に用いられる閾値を変動させることによる影響も調査する．すなわち，以上の 2 要因でパラメタを変化させることによる合意形成への影響をシミュレートする．

2.2　モデル

図 1 に示されるように，モデルには各ノードが 4 つのコネクションを持つ 10×10 のトーラス形状のグリッドネットワークが採用され，100 体のエージェント A がネットワークの各ノードに 1 体ずつ配置される．各エージェント A は 0 から 1 の実数値である意見 x_A を持つ．図 2 はあるエージェント A の近傍を示す．ネットワークは 1 ノードにつき 4 つのコネクションを持つため，図 2 内の $A_x^{\mathrm{nbd}}, x = \{n, s, e, w\}$ がエージェント A にとっての近傍他者として定義される．

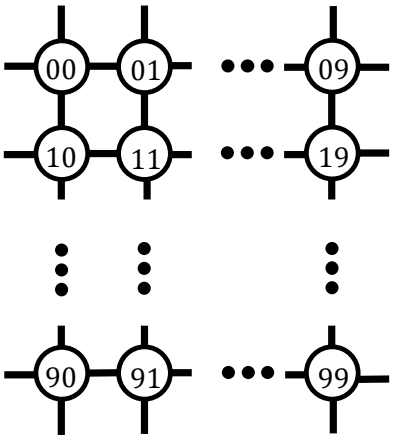

図 1: シミュレーションで用いられたトーラス形状のグリッドネットワーク.

図 3 は 1 回のシミュレーションの経過を示す．ある特定のパラメータの組み合わせを持つ人工社会において，各ノードにエージェントを重複なくランダムに配置する．その時，各エージェントは一様分布に従う 0 から 1 の間の実数値を初期意見 $x^{(0)}$ として持つ．初期状態はステップ数は 0 であり，シミュレーションが開始されると，ステップごとに各エージェントは近傍他者とペアを形成する．この時，ペア形成の仕方によっては，他のどのエージェントともペアを組まないエージェントが発生することがある．ペアを組んだエージェント同士はインタラクションし，条件に従って意見を更新する．このペア形成と意見の更新を s_{MAX} 回繰り返した後に，シミュレーションは終了する．本研究では，意見の更新方法が 2 通りあり，片方は合意形成シミュレーション研究でしばしば用いられる Deffuant モデルベースの意見更新方法で，もう片方はこれまでに提案されてきたロボットのための集団規範アウェアな意思決定手法に基づく．

以上の乱数によって制御される初期意見とペアの生成は seed 値によって統制され，複数の seed 値に基づいた乱数生成でシミュレーションを複数回実行する．したがって，あるパラメータの組み合わせによって形成される人工社会は複数回シミュレートされ，そのパラメータの組み合わせに基づいた人口社会の多様な変遷を観察する．

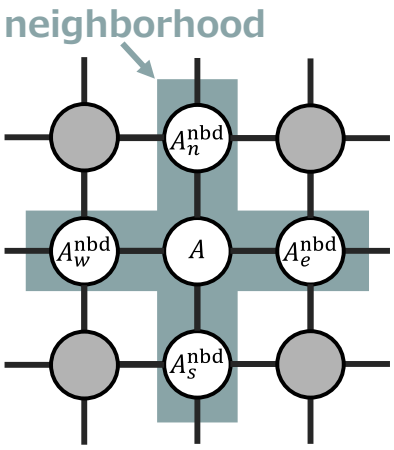

図 2: グリッドネットワークでの近傍の範囲．青色で着色されたエリアが近傍となる．

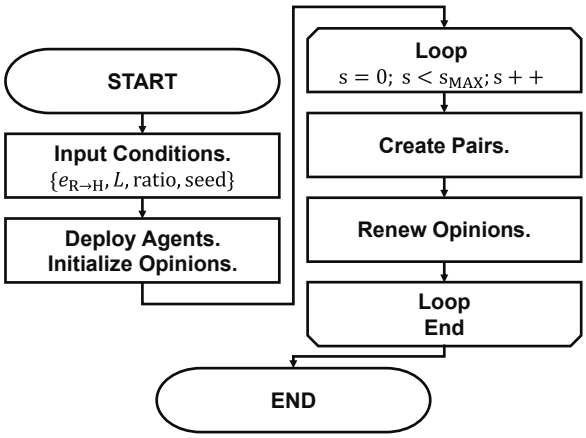

図 3: 1 回のシミュレーションの流れ．制御パラメータの入力，エージェントの配置や意見初期化の後，ステップが s_{MAX} 回繰り返される．

2.3 エージェント

2.3.1 Deffuant モデルに基づくエージェント

Deffuant モデルは，個々人が持つ意見のダイナミクスや社会的影響に関するマルチエージェントベースのモデルの 1 種である [11]．本モデルは個人の間の相互作用がどのように社会全体の意見の分布やコンセンサス形成に影響するかを理解するために用いられる．Deffuant モデルには意見 x を持つエージェントが存在し，各エージェントの意見の初期化の後，ステップごとにランダムに設定されたエージェントのペアの間のインタラクションが繰り返され，システムが安定する（意見の変化がなくなる）までステップが更新され続ける．

Deffuant モデルにおけるエージェント間のインタラクションの手順を以下に示す．あるエージェント i の意見 x_i は実数値 $[0.0, 1.0]$ によって表現される．意見の初期化では各エージェントの意見に一様分布に従う乱数が代入される．ランダムに選定されたエージェント i と j の意見 x_i, x_j が式 1 を満たす場合にインタラクションが発生し，式 2 と式 3 によってそれぞれのエージェントの意見 x_i, x_j が更新される．この更新は 2 体のエージェントが合意に近づくよう促進する．意見の差が許容閾値 L 未満であれば，意見の近いエージェント同士のみがインタラクションすることを意味する．許容閾値 L 以上である場合，エージェント同士は意見がかけ離れておりそれぞれの意見を更新しない．パラメータ μ は意見収束の速さを表す．

$$|x_i - x_j| < L \tag{1}$$

$$x_i = x_i + \mu(x_j - x_i) \tag{2}$$

$$x_j = x_j + \mu(x_i - x_j) \tag{3}$$

本シミュレーションでは，各ステップにおいてペアを形成できないエージェントが存在することがある．あるステップにおいてペアを形成しない HA は前ステップと同じ意見を持つ．したがって，HA の意見は更新されない．

2.3.2 集団規範アウェアなエージェント

集団規範アウェアなエージェント (GNAA) はステップ t において，環境を観察し，その観察を基に内部状態を更新し，その内部状態に基づいて行動を出力する．そのような入出力をステップごとに実施し，他者を観察しながら，エージェント自身の振舞いを集団に適応させ続けることを試みる．本研究では，観察対象は近傍の他者の意見 x の集合 X_{nbd} であり，行動は自らの意見 x_{self} である．本研究で実施されるシミュレーションにおいて，集団規範アウェアなエージェントの意思決定の方法は先行研究で提案された意思決定手法を基にして実装される．

合意形成シミュレーションにおいて，意見 x は $[0.0, 1.0]$ の実数値を取る．GNAA は選択可能な意見の集合 $X_{\mathrm{self}} = \{k\delta \mid k \in \mathbb{N},\ 0 \leq k\delta \leq 1\}$ から 1 つの意見 x_{self} を採用して出力する．内部状態としてエージェントは価値関数 V を持つ．エージェントは特定集団で暗黙的に共有されている規範に則った適応的行動を推定するために価値関数を用いる．価値関数は，エージェントが実施可能な振舞いや準拠可能な行動実行の基準の集合に，特定の集団内でそれに基づいて実際に意思決定することの価値の集合を割り当てる．

本意思決定手法は集団内の規範が存在することを前提として，選択可能な意見集合 X_{self} に対して，意見 x を表明する価値の集合を割り当てる．そして，その関数の中で最も高い価値を割り当てられた意見 $x^* =$

$\arg\max \mathrm{V}(x)$ がエージェントの意見として採用される．価値関数は式 4 と式 5 によって更新される．

$$\mathrm{V}(x) \leftarrow (1-\alpha)\mathrm{V}(x) + \alpha\left(\mathrm{R}(x) + \gamma \max \mathrm{V}(x)\right) \quad (4)$$

$$\mathrm{R}(x) = \sum_{x' \in X_{\mathrm{nbd}}} \left\{ \exp\left(-\frac{(x-x')^2}{\mathrm{kurtosis}} \right) \right\} \quad (5)$$

すなわち，各 RA は $[0.0, 1.0]$ の範囲をとる意見 x を複数個の実数値に等間隔に分割して選択可能な意見 $x' \in A$ として保持し，各ステップにおいて価値関数に基づいて A の中から 1 つ選択しその RA 自身の意見 x を採用する．価値関数の更新に伴い RA が持つ意見が変遷していくことになり，このことは集団規範アウェアにロボットの意見を調整することを意味する．

本シミュレーションでは，各ステップにおいてペアを形成できないエージェントが存在することがある．その場合，RA はインタラクションすることなく近傍エージェントの意見の観察に基づいて自身の価値関数を更新する．HA とは異なり，ペアの形成ができないエージェントであっても意見は更新される．このことは集団規範アウェアなロボットは意見を表明することは無くとも，自身が所属する集団メンバの意見や行為を観察していることを表している．

2.3.3 インタラクションルール

本シミュレーションでは，Deffuant モデルに従うエージェントを人間エージェント (HA)，集団規範アウェアなエージェントをロボットエージェント (RA) とみなす．すなわち，2 つの意見更新方法のいずれかを採用して振舞う多数のエージェントがネットワークにランダムに配置される．本シミュレーションのパラメタは表 1 と 2 に示される．

ペアを組んだエージェントが HA 同士なら，式 1 を満たす場合において式 2 と式 3 によって意見が更新される．RA と HA がペアを組んだ場合，式 1 の条件に関わらず，RA は価値関数を式 4，5 に基づいて更新する．HA は式 1 の条件を満たす場合にのみ，RA の意見 x_{R} に基づいて HA の意見 x_{H} を式 6 によって更新する．

$$x_{\mathrm{H}} = x_{\mathrm{H}} + e_{\mathrm{R}\to\mathrm{H}} \times \mu(x_{\mathrm{R}} - x_{\mathrm{H}}) \quad (6)$$

$e_{\mathrm{R}\to\mathrm{H}}$ は RA から HA の意見への影響効率係数である．Deffuant モデルにおいて人間エージェント同士が影響を与える効率 $e_{\mathrm{H}\to\mathrm{H}}$ を 1.0 と定義する．ロボットから人間への影響は必ずしも人間同士の影響と同じ程度を持つとは限らないと仮定し，制御パラメータとして $e_{\mathrm{R}\to\mathrm{H}}$ を 0 から 1 の範囲で変化させて複数シナリオでのシミュレーションを実施する．これにより，エージェント社会の中の HA が持つ影響係数 $e_{\mathrm{R}\to\mathrm{H}}$ の大きさと集団規範アウェアな RA の存在が混合モデル内での合意形成に影響をどの程度与えるのかを分析する．

2.4 パラメータ

表 1 には固定されたパラメータが示されている．各エピソードにおける初期意見の分布は seed 値に基づいて統一されており，制御パラメータの変化によって発生する変化のみを原因としたマクロへの影響を観察する．

表 2 に基づき，制御パラメータとして L，$e_{\mathrm{R}\to\mathrm{H}}$ を変化させた混合モデルを用いて，複数のシナリオをシミュレーションする．1 度のシミュレーションの実施において，L，$e_{\mathrm{R}\to\mathrm{H}}$，rate，seed 値に個別の値が代入される．また，seed 値は 0 から 99 の整数値を取ることから，L，$e_{\mathrm{R}\to\mathrm{H}}$，rate によって定められる人工社会において，初期意見とペアの生成方法が 100 パターン生成されることを意味する．すなわち，seed 値を 100 個用いてシミュレーションすることによって，初期意見とペアリングに依存しない傾向を観察することを目論んだ．

Deffuant モデルでのシミュレーションにおいてはネットワーク構造に関わらず $L \geq 0.5$ の場合はエージェント社会全体での合意形成が発生することが示されている [12]．したがって，例えば $L = 0.25$ の場合はエージェント社会での 2 種類の意見への合意形成の収束が起こることが想定される．すなわち，L を 0.5 もしくは 0.25 に設定することで，HA が全体的な合意形成を起こしやすい人工社会と起こしにくい人工社会でのシミュレーションを実施する．

パラメータ $e_{\mathrm{R}\to\mathrm{H}}$ を制御することで，人間の他者に対する受容の度合いやロボットから人間への影響力の違いがマクロであるエージェント社会全体に与える影響を評価する．また，RA の存在すること自体による合意形成への影響を分析するため，100 体の HA のみが存在する通常の Deffuant モデルのシミュレーションも制御パラメータとして L と seed 値を変化させて実施した．

本シミュレーションのソースコードは python の mesa ライブラリを用いて実装された．

2.5 データの収集と分析

本研究では各シミュレーションにおいて各エージェントが持っていた意見 x の時系列変化を記録する．パラメータ L と $e_{\mathrm{R}\to\mathrm{H}}$ が制御され，これらの値の変化が人工社会全体に与える影響を評価するために，具体的には得られるデータから以下の 3 つの指標を算出する．

第 1 に，人工社会内エージェントが持つ意見の標準偏差 SD の時系列変化を可視化する．各ステップにおける意見のばらつき度合を算出することで，人工社会全体での意見の収束を観察する．第 2 に，最終ステップでの意見の頻度を算出する．意見は 0 から 1 の実数値であり，0.05 刻みで合計 20 クラスを作成する．最終

表 1: 設定パラメータ.

Model Parameter	
Opinion x	$[0.0, 1.0]$
Convergence coefficient μ	0.5
width, height	10, 10
neighborhood	4-connectivity
Manhattan distance d	1
Number of Episodes (seed value)	100
Max Steps MAX	500
Robot Agent Parameter	
Increment of Opinion δ	0.01
number of set of actions $\lvert X_{\text{self}}\rvert$	101
Learning rate α	0.1
Discount factor γ	0.9
Gaussian parameter kurtosis	100

表 2: 制御パラメータ.

Tolerance Threshold L	0.25, 0.5
Effect $e_{\text{R}\to\text{H}}$	0, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.5, 1.0
Human–Robot rate	100:0, 50:50
Seed Value	0 to 99

ステップにおける各エージェントの意見を各クラスに分類する．これによって，最も頻度が高いクラスに含まれる頻度の値を基に，その社会のエージェントの意見が完全に収束したのか，多数派と少数派の意見に分離したのかを評価する．また，意見頻度の算出によって，全エージェントの内の過半数の意見が 1 つのクラスに属する場合，その社会の意見の中央値は多数派の意見 x^M とみなせる．

最後に，人間社会と人間ロボット社会の間の多数派意見 x^M と IQR(四分位範囲) の差を算出する．人間社会 (HS) と人間ロボット社会 (HRS) はそれぞれ，全てのエージェントが HA のみで構成された社会，エージェントが HA と RA で構成された社会を意味する．ここでは同じ seed 値とパラメータの基での HS と HRS での最終ステップでの多数派意見 x^M_{HS}，x^M_{HRS} を比較する．エージェントの初期意見 $x^{(0)}$ と各ステップにおけるペアの形成は seed 値によって統制されているため，社会を構成するエージェントの意見更新方法の違いの観点から多数派意見を比較可能である．すなわち，HA のみで構成される社会での最終的な多数派意見 x^M_{HS} と比べて，集団規範アウェアな RA が混在する社会での最終的な多数派意見 x^M_{HRS} はどのように変化するのかを観察する．このことは，集団規範アウェアな RA がミクロである集団へ適応することが，マクロである人工社会全体に影響を与えることに関する指標となる．

以上より，人工社会全体での意見のばらつき度合，多数派意見の形成，過半数意見の変動というこれらの評価指標を制御パラメタ L，$e_{\text{R}\to\text{H}}$ の基で観察する．

3 結果

図 4，図 5 は各 L 条件において影響係数 $e_{\text{R}\to\text{H}}$ の値ごとの意見収束の時系列変化を示す．凡例にある数値は影響係数 $e_{\text{R}\to\text{H}}$ の値を示しており,「only HA」は Deffuant モデルに基づく HA のみの社会 (HS) での標準偏差の時系列変化を意味する．加えて，各条件の薄い色で色づけられた範囲は 95%信頼区間 (CI) を示している．

図 4 と図 5 において，いずれの条件においてもステップを経るごとに SD は減少する傾向にあり，特定のステップから減少傾向は無くなっていることが観察された．図 4 の onlyHA 条件と比較して，混合モデルの極端に小さい $e_{\text{R}\to\text{H}}$ を持つ条件を除き，SD はより減少している．加えて，$e_{\text{R}\to\text{H}} = 0.05$ あたりを境にそれ以上の値を持つ $e_{\text{R}\to\text{H}}$ の条件ではより一層 SD の値が 0 に接近した．一方で，図 5 の onlyHA 条件では最終的に SD が 0 となるまで減少する一方で，混合モデルの極端に小さい e を持つ条件では SD が 0 に近づくまで減少することは無かった．しかしながら，$e_{\text{R}\to\text{H}} \geq 0.1$ の条件であれば，onlyHA 条件と同様に収束する様子が観察された．

図 6，7 は各 L 条件において影響係数 $e_{\text{R}\to\text{H}}$ の組み合わせごとの 5 番目まで高い頻度をもつクラスの頻度を示す．縦軸は頻度，横軸は i 番目に高い頻度を示したクラスを意味する．すなわち，図 6，7 は各人工社会内の意見の分布に関わらず，意見の収束の度合いを評価することにつながる．図 6，7 の両 L 条件において共通して $e_{\text{R}\to\text{H}} \geq 0.05$ の場合において，最も頻度の高いクラスが 50 の値を超えている．すなわち，$e_{\text{R}\to\text{H}} \geq 0.05$ の場合において，エージェントの過半数が 1 つの意見に収束している傾向が明らかになった．

図 8，9 は各 L 条件と影響係数 $e_{\text{R}\to\text{H}}$ の組み合わせごとの人間社会と人間ロボット社会の間の中央値と IQR(四分位範囲) の差を示している．縦軸は人間ロボット社会での中央値もしくは IQR から人間社会での中央値もしくは IQR を引き算した値であり，横軸は影響係数 $e_{\text{R}\to\text{H}}$ である．すなわち，縦軸の値が 0 より大きいもしくは小さい場合は，人間社会での中央値 x^M_{HS} と比べて人間ロボット社会での中央値 x^M_{HRS} が異なっていたことを意味する．このことは，最終的に収束に至った時に過半数派意見として採用された値との変化があることを示す．ただし，この指標において x^M_{HS}，x^M_{HRS} が有効であるのは，過半数意見を形成する傾向のあった $L_{.25}$

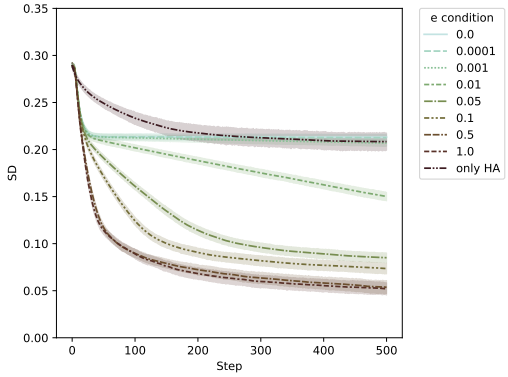

図 4: $L_{.25}$ 条件において，集団規範アウェアなロボットエージェントから人間エージェントへの影響係数 $e_{\mathrm{R}\to\mathrm{H}}$ を制御パラメータとして変化させた場合の意見収束の時系列変化．人間エージェントのみで構成されるモデルである onlyHA 条件も含む．

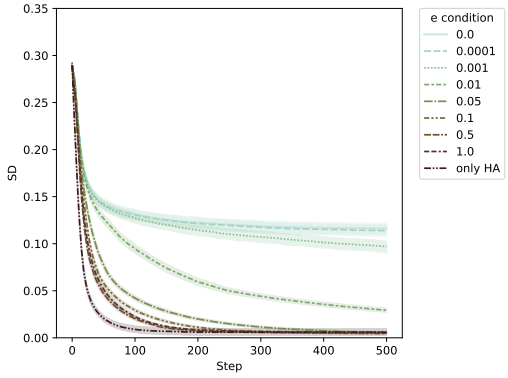

図 5: $L_{.50}$ 条件において，集団規範アウェアなロボットエージェントから人間エージェントへの影響係数 $e_{\mathrm{R}\to\mathrm{H}}$ を制御パラメータとして変化させた場合の意見収束の時系列変化．人間エージェントのみで構成されるモデルである onlyHA 条件も含む．

かつ $e_{\mathrm{R}\to\mathrm{H}} \geq 0.05$ の場合と $L_{.50}$ かつ $e_{\mathrm{R}\to\mathrm{H}} \geq 0.01$ の場合のみである．また，IQR に関する値が 0 よりも小さいことと大きいことは，人間ロボット社会での意見の分布が小さいことと大きいことを意味する．これは収束の度合いや社会での意見の先鋭化の度合いの指標としてみなせる．

図 9 が示す $L_{.50}$ 条件では，中央値と IQR が共に 0 に近い値になっている．一方で，図 8 が示す $L_{.25}$ 条件では大きく変動している．特に，過半数派が発生したとみなせる $e_{\mathrm{R}\to\mathrm{H}} \geq 0.05$ の場合においては，中央値がおおよそ ± 0.3 の範囲で変化したこと，IQR の差は 0 かマイナスの値を取った．

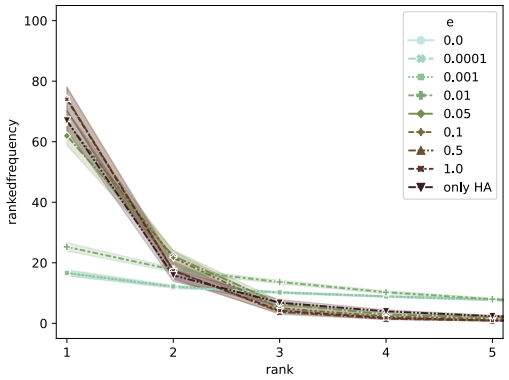

図 6: $L_{.25}$ 条件における意見頻度の降順ソート．

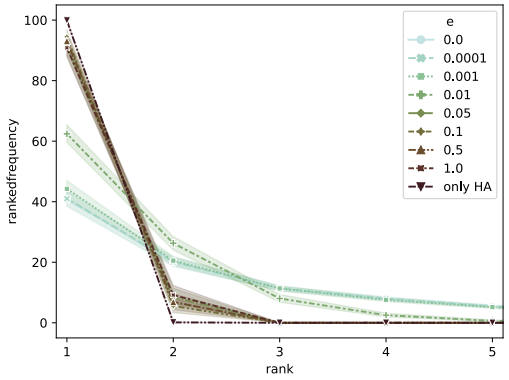

図 7: $L_{.50}$ 条件における意見頻度の降順ソート．

4 考察

図 4，図 5 より，人間ロボット社会 HRS にて $e_{\mathrm{R}\to\mathrm{H}} = 0.1$ でも人工社会全体での意見収束を発生させられることが明らかになった．このことは集団規範アウェアロボットから人間への影響が人同士の 10 分の 1 の効率であったとしても社会全体での合意形成を促進し得ることを示唆している．加えて，$L_{.25}$ かつ $e_{\mathrm{R}\to\mathrm{H}} \geq 0.01$ の条件において人間社会 HS よりも意見の合意形成が進むことも明らかになった．集団規範アウェアな意思決定がある集団から別の集団へと意見を伝播させ，社会全体での共通意見の形成に貢献したことが示唆された．すなわち，人工社会内の意見の標準偏差という指標から，人間同士の意見の影響力ほどロボットから人間への影響力が大きくなくともロボットの振舞いが集団規範アウェアであれば，それらのロボットの意思決定は社会全体での合意形成に影響を与える可能性があることを本シミュレーション結果は示唆している．

図 6，7 は $e_{\mathrm{R}\to\mathrm{H}} \geq 0.05$ において社会に過半数派が形成される傾向を示唆した．すなわち，その条件において人間社会 HR と同様に人間ロボット社会 HRS でも圧倒的多数派と少数派が生まれる傾向がみられた．一

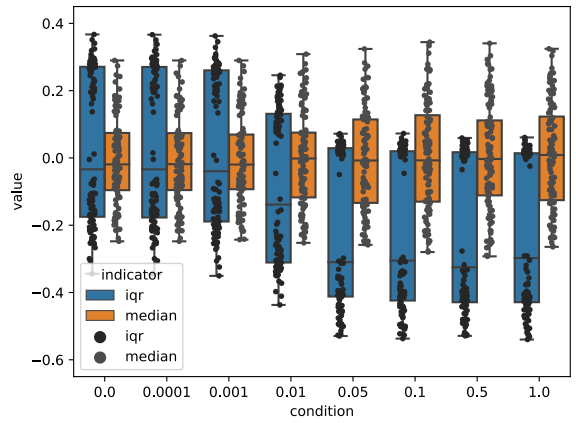

図 8: $L_{.25}$ 条件における人間社会と人間ロボット社会での意見の中央値の差と両社会での IQR の差.

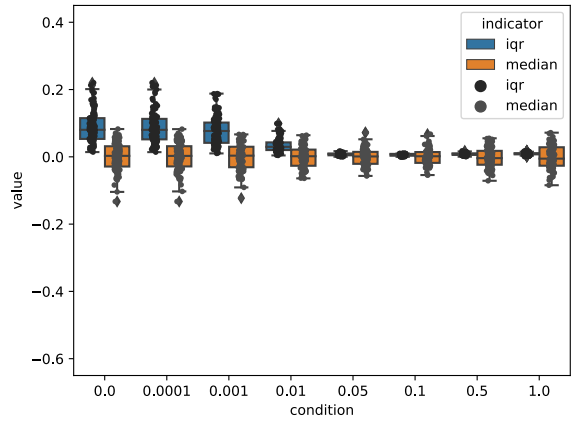

図 9: $L_{.50}$ 条件における人間社会と人間ロボット社会での意見の中央値の差と両社会での IQR の差.

方で，$e_{\mathrm{R}\to\mathrm{H}} < 0.05$ の場合は複数の少数派が誕生し，過半数のエージェントが採用するような意見は発生しない傾向がみられた．このことは，ロボットによる集団への適応が社会全体というよりはローカルの合意形成を促進し，全体での意見収束が発生しなかったことを意味する．多様な意見を維持した社会が実現しているとも読み取ることができる．

図 8，9 より，$L_{.50}$ 条件では値が 0 付近であったが，$L_{.25}$ 条件で過半数派を形成した $e_{\mathrm{R}\to\mathrm{H}} \geq 0.05$ 条件における多数派の意見 x^M_{HS}，x^M_{HRS} を比較すると，人間とロボットの初期位置や初期意見によっては ± 0.3 の差が発生した．意見は 0 から 1 の実数値であり，特に $L_{.50}$ 条件と比べて $L_{.25}$ 条件では大きな規範変容の発生が起こった傾向を示している．すなわち，人間が合意形成において他者の意見との違いを許容しにくい社会では，ロボットによる集団規範アウェアな意思決定を起因として，人間同士で形成する合意された意見とは異なる意見が過半数派の意見として採用される可能性があることを示唆している．このことは，集団規範アウェアな意思決定というミクロでの適応が，マクロでの合意形成に影響を与えることを示唆している．

ロボットから人間への影響効率が人間同士の影響効率と比べて小さい場合においても，人間社会と人間ロボット社会では意見のばらつきが小さくなり，多数派や過半数派の意見が社会の中で形成された．しかしながら，人間が他者との意見の違いを許容しにくい社会において合意された意見が人間社会と人間ロボット社会で発生する場合，異なるものになる可能性が示唆されている．このことは，社会に自律したエージェントが参入する場合において，人間のみで共有されてきた規範が規範を考慮するエージェントによって変容させられる可能性があるという定性的な指摘 [10] と合致する．しかしながら，シミュレーション上での規範の変容自体は数値の変化であり，このシミュレーションではその変化自体の評価はできない．

5 おわりに

本研究では多数の集団規範アウェアな意思決定をするロボットエージェントが人間社会に参入してインタラクションすることによる影響をシミュレーションによって評価した．特に，ロボットから人間への影響効率と人間の意見の許容閾値を制御することによる人工社会全体への影響を探索した．社会を構成する人間がロボットなどのエージェントから受ける影響を定量的にシミュレーションすることは，他者への同調という人間が持つ特性を含めたエージェントデザインの土台構築に貢献することが期待できる．シミュレーションから，ロボットから人間への影響力が人間同士の影響力よりも小さい場合であっても，集団規範アウェアなロボットの意思決定は社会全体の合意形成を促進することが明らかになった．また，人間が他者との意見の違いを許容しにくい社会においては，人間ロボット社会で多数派として合意される意見が人間社会で本来形成されるはずの多数派意見と異なるものになることが観察された．このことは規範を考慮するエージェントによって社会全体の規範が変容する可能性を示唆している．将来的には，他のパラメータの制御や人間とロボットの比率，ネットワーク構造の変更による人間ロボット社会での合意形成について調査する．

謝辞

本研究は JSPS 科研費 JP23K16983 の助成を一部受けたものである．

参考文献

[1] Leite, I. et al.: Social robots for long-term interaction: A survey, *International Journal of Social Robotics*, Vol. 5. No. 2, pp. 291–308 (2013)

[2] Sheridan, T. B.: A review of recent research in social robotics, *Current Opinion in Psychology*, Vol. 36, pp. 7–12 (2020)

[3] Feldman, D. C.: The development and enforcement of group norms, *Academy of Management Review*, Vol. 9, No. 1, pp. 47–53 (1984)

[4] Fuse, Y., Tokumaru M.: Social Influence of Group Norms Developed by Human–Robot Groups, *IEEE Access*, Vol. 8, pp. 56081–56091 (2020)

[5] Fuse, Y. et al.: Unleashing Fairness: How a Group Norm-Aware Agent Shakes Up the Ultimatum Game, *IEEE Access*, Vol. 11, pp. 36727–36740 (2023)

[6] Nass, C., Steuer, J. and Tauber, E. R.: Computers are social actors, *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*, pp. 72–78 (1994)

[7] Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, (1996)

[8] Borenstein, J., Arkin, R.: Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being, *Science and Engineering Ethics*, Vol. 22, pp. 31–46 (2016)

[9] Hang C., Ono T., Yamada S.: Designing Nudge Agents that Promote Human Altruism, *The 13th International Conference on Social Robotics*, pp. 375–385 (2021)

[10] Coggins, T. N., Steinert, S.: The seven troubles with norm-compliant robots, *Ethics and Information Technology*, Vol. 25, Issue 2, 29 (2023)

[11] Deffuant, G. et al.: Mixing beliefs among interacting agents,*Advances in Complex Systems*, Vol. 3, Issue 01n04, pp. 87–98 (2000)

[12] Fortunato, S.: Universality of the Threshold for Complete Consensus for the Opinion Dynamics of Deffuant et al., *International Journal of Modern Physics C*, Vol. 15, No. 9, pp. 1301–1307 (2004)