拒否エージェント: 意図的拒否による主体性表現のデザインと受容性の検討

山﨑 貴紀[1]　門脇 諒生[1]　武藤 剛[1]　武藤 ゆみ子[2,a)]

概要：大規模言語モデル（LLM）の発展により，バーチャル物理環境で自然言語指示に基づきタスクを遂行する自律エージェントの研究が活発化している．従来は作業支援を主目的とし，指示遵守や効率性の最大化が中心であった．一方，エージェントを長期的な協働相手として捉える場合，効率性とは異なるデザイン軸として主体性の表出が重要となる．本研究では，LLM による指示理解と内部状態（疲労・機嫌）に基づく拒否判断を分離して制御し，意図的に拒否（Won't）を行う「拒否エージェント」を提案する．不能（Can't）と拒否（Won't）を区別し，拒否が故障ではなく意思表示として知覚されるための表現（言語的説明と代替行動）を検討する．また Responsible AI の観点から，制御可能性・害の限定性・透明性を満たすガードレールを実装した．評価として，シミュレーションにより拒否–回復サイクルの安定性と拒否頻度の制御可能性を検証し，予備的印象評価を通じて拒否が主体性知覚・社会的実在感・受容性に与える影響を考察する．

1. はじめに

近年の大規模言語モデル（LLM）の発展に伴い，バーチャル物理環境において自然言語指示に基づきタスクを遂行する自律エージェントの研究が活発化している．LLM の高い言語理解・推論能力により，ユーザの曖昧な指示を解釈し，状況に応じた行動計画を生成するエージェントデザインが現実的になりつつある．

従来，この種のエージェント研究は作業支援・自動化を主目的とし，指示遵守や効率性（達成率・所要時間など）の最大化が中心的なデザイン目標であった．Minecraft 環境を題材とした研究では，LLM を用いた探索・計画・技能獲得により，長期的タスクの遂行能力を拡張し得ることが示されている．Voyager は，自動カリキュラム生成とスキルライブラリの蓄積を通じて継続的に探索・学習を行うエージェントを提案している [1]．GITM は，テキストベースの知識と記憶を統合し，構造化された行動表現にもとづく計画・実行を行う枠組みを提示した [2]．さらに White らは，複数エージェントの協調に着目し，Minecraft 上での協調的な身体性を伴う推論（embodied reasoning）を評価するためのプラットフォーム／ベンチマーク（MineCollab）を提案している [3]．これらはバーチャル環境におけるエージェント能力を拡張する一方で，多くの場合，タスク達成や効率向上が主要な関心となっている．

しかし，エージェントを単なる道具ではなく，人間と長期的な関係性を築く協働相手として捉える場合，求められるデザイン要件は変化し得る．人間同士の協働では，パートナーは常に従順に振る舞うわけではなく，自身の状態や選好に応じて交渉したり，ときに拒否したりする．このような「摩擦」は効率の観点からは不利に見える一方で，相手に主体性があるという感覚や社会的実在感を生む契機にもなり得る．したがって，対等なパートナーシップの形成に向けては，効率性とは異なるデザイン軸，すなわち主体性をどのように表出するかが重要な課題となる．

この点に関連して，内部状態や記憶にもとづく社会的ふるまいを扱う研究も報告されている．たとえば Generative Agents は，経験の記録・想起・内省・計画といった機構を組み合わせ，仮想空間内で複数エージェントが生活し相互に交流する様子を示した [4]．一方で，同研究の主眼はエージェント間の社会的振る舞いの生成にあり，ユーザに対する明示的な不服従（意図的な拒否）がもたらす影響は主要な検討対象ではない．

一方，ロボット・AI が人間の指示を拒否する状況を扱った研究も存在する．Briggs と Scheutz は，指示拒否が必要となる要因（能力的制約，状況的な不適切さ，規範・倫理的理由など）や，拒否と説明の重要性を議論している [5]．また Salem らは，ロボットのエラー（faulty behavior）が信頼や協力に与える影響を調査している [6]．ただし，これ

[1] 文教大学 情報学部
[2] 玉川大学 脳科学研究所
[a)] muto@lab.tamagawa.ac.jp

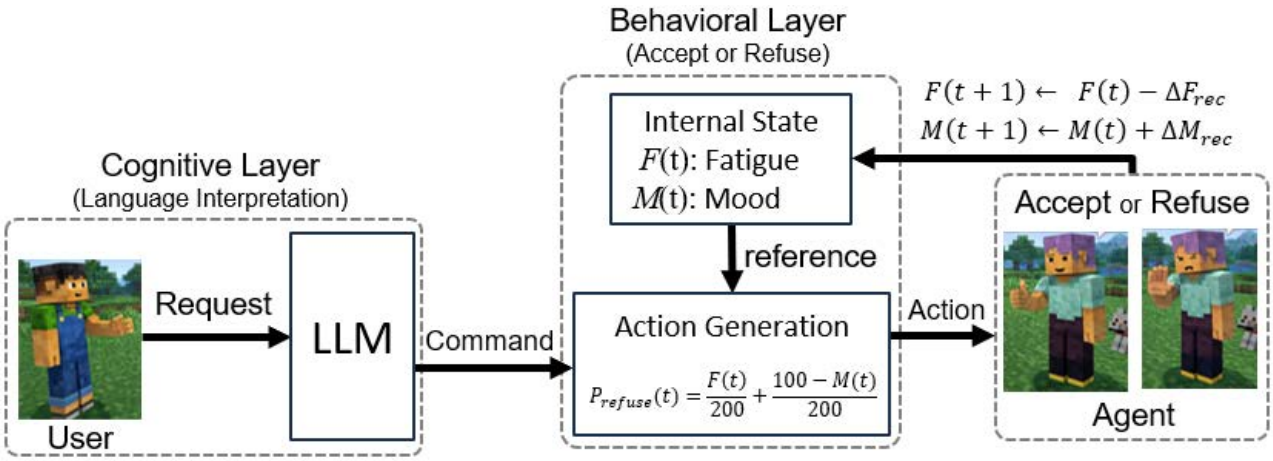


図 1: 拒否エージェントの概要

らは主として「できない（Can't）」や意図せぬ失敗，あるいは規範にもとづく拒否に焦点があり，能力的には可能であるにもかかわらず，エージェント自身の意思として従わない「意図的な拒否（Won't）」を関係性構築の観点から扱う検討は限定的である．

さらに，CASA（Computers Are Social Actors）[7] や Media Equation[8] によれば，人間はコンピュータやエージェントに対しても社会的ルールを適用し，社会的存在として扱う傾向がある．この観点からは，常に従属する存在よりも，拒否や交渉を含む双方向の駆け引きを行うエージェントの方が，より強い社会的存在として知覚される可能性がある．一方で，Won't を主体性として成立させるには，単なる停止ではなく，拒否理由の説明や「やらない」ことを行為として可視化する表現デザインが重要になると考えられる．

そこで本研究では，Minecraft[9] 環境上で，LLM による指示理解と内部状態（疲労・機嫌）に基づく拒否判断を分離して制御するハイブリッド・アーキテクチャを採用し，意図的に拒否（Won't）を行う「拒否エージェント」を実装・提案する．本稿では，不能（Can't）と拒否（Won't）を明確に区別した上で，拒否が「故障」ではなく「意思表示」として知覚されるための表現（言語的説明と代替行動）を検討する．本稿では，以後，Can't を「不能／故障」，Won't を「意図的拒否」として用い，両者を区別して議論する．また Responsible AI の観点から，制御可能性・害の限定性・透明性を満たすガードレールを前提に，シミュレーションおよび予備的な印象評価を通じて，拒否という摩擦が主体性知覚・社会的実在感・受容性に与える影響を考察する．

2. 拒否エージェントの実装

2.1 拒否エージェントとは

本研究において，「拒否エージェント（Refusal Agent）」とは，ユーザからの要求を遂行する能力を有しているにもかかわらず，自身の内部状態や選好に基づき，その要求を意図的に拒絶する意思決定を行う自律エージェントと定義する．

具体的には，以下の 3 要件を満たすものを指す．

(1) 実行可能性: 物理的・システム的な障害がなく，本来であればタスク遂行が可能であること．
(2) 意図的な選択: 乱数や内部パラメータに基づき，システムが能動的に「実行しない」を選択すること．
(3) 社会的表明: 拒否の理由を自然言語や振る舞いによってユーザに伝達すること．

2.2 システム構成

拒否エージェントのシステム概要を図 1 に示した．Minecraft 用ボットフレームワーク Mineflayer と OpenAI API を統合した MinePal[10] を基盤とし，これに独自の拒否判定モジュールを追加実装した．システム全体は，言語理解を担う「**Cognitive Layer**」と，行動決定および実行を担う「**Behavioral Layer**」の 2 層構造から成る．

2.2.1 処理フローと介入機構

ユーザからの自然言語入力 I に対するエージェントの応答プロセスは，以下の手順で実行される．

- **Cognitive Layer** : 意図解釈を担う．LLM（GPT-4）が入力 I を解析し，実行可能なコマンド C と引数 A を生成する．この段階ではエージェントの内部状態は考慮されず，純粋な意味解析のみが行われる．
- **Behavioral Layer**: 介入判定を担う．コマンド実行関数 `executeCommand(C, A)` が呼び出された際，実際の行動実行 `perform()` の直前に，拒否判定関数 `shouldRefuse()` による割り込み処理が行われる．

$$\text{Action} = \begin{cases} \text{Refuse}(C) & \text{if } \texttt{shouldRefuse()} = \text{true} \\ \text{Perform}(C, A) & \text{otherwise} \end{cases} \tag{1}$$

この分離デザインにより，LLM のプロンプトを複雑化させることなく，ルールベースによる確実な行動制御が可能となっている．

2.3 拒否モデル

エージェントの利己的な振る舞いを確率的に生成するため，以下の 2 つの内部状態変数を導入した．

- 疲労度 (Fatigue):$F(t) \in [0, 200]$
  タスク実行ごとに蓄積し，休息や拒否行動によって減少する．
- 機嫌 (Mood): $M(t) \in [0, 100]$
  エージェントの精神状態を表し，$M = 0$ が最悪，$M = 100$ が最良の状態に対応する．

これらの内部状態変数 $F(t)$ と $M(t)$ を用いて，ある時点 t における拒否確率 $P_{refuse}(t)$ は，以下のモデルによって動的に算出される．

$$P_{\text{refuse}}(t) = \min\left(1, \frac{F(t)}{200} + \frac{100 - M(t)}{200}\right) \tag{2}$$

式 (2) により，確率は上限を 1 として算出される．

また，判定により拒否が選択された場合，エージェントはサボタージュ行動（言い訳チャットの送信および代替行動）を実行する．この際，自身の内部状態に対して以下の更新を行い，状態の回復を図る（負のフィードバックループ）．

$$F(t+1) = \max\bigl(0,\, F(t) - \Delta F_{\text{rec}}\bigr) \tag{3}$$
$$M(t+1) = \min\bigl(100,\, M(t) + \Delta M_{\text{rec}}\bigr) \tag{4}$$

拒否が選択されなかった場合（受諾して協力した場合）は，タスク負荷として疲労が増加し，機嫌が低下するよう，内部状態を次式で更新する．

$$F(t+1) = \min\bigl(200,\, F(t) + \Delta F_{\text{task}}\bigr) \tag{5}$$
$$M(t+1) = \max\bigl(0,\, M(t) - \Delta M_{\text{task}}\bigr) \tag{6}$$

$\Delta F_{rec}, \Delta M_{rec}$ は，回復量を示す．本実装では $\Delta F_{rec} = 8, \Delta M_{rec} = 5$ と設定しており，これにより「サボることで元気になる」というサイクルを実装している．すなわち，時刻 t(現在) の疲労 $F(t)$ から数値を引くことで，疲れが取れ，機嫌 $M(t)$ に数値を足すことで，機嫌が直るという仕組みである．

2.4 拒否表現（Refusal Expressions）

Won't を故障（Can't）ではなく意思表示として知覚させるため，拒否を単なる停止として表出するのではなく，解釈可能な行為としてデザインする必要がある．本システムでは拒否時に，(1) 内部状態（疲労・機嫌）に基づく簡潔な言語的説明と，(2) 箱庭内で安全かつ観察可能な代替行動（例：背を向ける，低リスクな行動を行う等）を組み合わせて提示する．これにより，理由づけ，可視化，回復可能性を満たし，Won't の可読性を高める．

2.5 Responsible AI と安全な反抗のデザイン

「意図的に人間に従わない AI」は，デザインを誤ればユーザの制御を離れ，予期せぬ不利益をもたらすリスクがある．本研究では，Responsible AI（責任ある AI）の原則に基づき，以下の「安全な反抗（Safe Rebellion）」を実現するガードレールを実装した．

2.5.1 制御可能性

エージェントがいかに利己的に振る舞おうとも，人間による最終的な制御権は担保されなければならない．本システムでは，緊急停止コマンド（`!stop`）や状態管理コマンド（`!status`, `!rest`）を `ALWAYS_ALLOW` リストとして定義した．これらは `shouldRefuse()` の判定ロジックをバイパスし，いかなる内部状態であっても無条件かつ即座に実行される．これにより，エージェントの暴走や機能不全に対する安全弁（Kill Switch）が機能する．

2.5.2 害の限定性

本研究における「反抗」は，Minecraft というゲーム空間内でのみ許容されるものである．拒否やサボタージュによる害は，ゲーム内リソースの時間的損失やユーザの心理的摩擦（イライラ等）に限定されており，現実世界の物理的・経済的損失にはつながらないようデザインされている．この安全な箱庭の中でこそ，人間は AI の反逆をエンターテインメントやコミュニケーションの一環として受容できると考えられる．また，Minecraft の名称・資産の利用に関しては公式ガイドラインに従う [11]．

2.5.3 透明性

ユーザに対し，本エージェントが「拒否を行うようデザインされた実験的 AI」であることを事前に明示することで，予期せぬ不服従による心理的混乱や，故障との誤認を防ぐ．

3. 印象評価とシミュレーション結果

本章では，(1) 動画視聴後質問紙による予備的な印象評価，(2) 拒否モデルのシミュレーションによる挙動検証，(3) Can't と Won't の知覚境界に関する観察的所見を報告し，拒否という摩擦を主体性表現として扱うためのインタラクションについて考察する．

3.1 質問紙による印象評価

拒否エージェントの受容性に関する予備的知見を得るため，エージェントの挙動を収録した動画を提示した後にオ

表 1: 質問紙調査（5 段階尺度）結果．中央値（Median）と四分位範囲（IQR）を示す．

項目	Median	IQR
Q1 お邪魔キャラ	4.0	0.00
Q2 他者妨害	4.5	1.75
Q3 非協力者	2.5	2.50
Q4 AI は協力すべき	4.0	0.75
Q5 拒否は妨害	2.5	1.75

ンラインアンケートを実施した（$N = 6$）．回答者は全員が日常的にゲームをプレイしており，週当たりゲーム時間（自由記述からの概算）は中央値 9.5 時間/週であった．回答は 5 段階尺度で評価した．尺度の端点は，Q2–Q3 では「1=強く妨害と感じた」〜「5=妨害と感じなかった」，Q1 および Q4–Q5 では「1=そう思わない」〜「5=そう思う」とした（Q5 は「拒否は妨害だと思うか」を問う）．それらの結果を表 1 に示す．

他プレイヤーによる妨害行為に対する評価（Q2，中央値 4.5，IQR 1.75）は高く，妨害は一般に強い否定的反応を引き起こすことが確認された．一方で，本研究で扱う拒否に関しては，「拒否行為は妨害行動と思えるか」（Q5）の中央値が 2.5（IQR 1.75）であり，拒否が一義的に妨害として分類されるとは限らない可能性が示唆された．また，「AI なら拒否せず協力するべきだ」（Q4）は中央値 4.0（IQR 0.75）であり，規範（AI は協力すべき）と分類（拒否＝妨害か）の間に個人差が生じ得る可能性が示された．

「Safe Rebellion」の受容性に関して，エージェントの拒否に対する感情的反応を問うたところ，6 名中 4 名が「不快だが気にしない」，2 名が「不快で気にする」を選択し，「許さない」を選択した者はいなかった．これは，拒否が心理的摩擦を引き起こしつつも，許容範囲内に留まっていることを示唆する．

自由記述では，「便利」「シンプル」といった有用性評価に加え，「もどかしい」「緩慢」といった摩擦（フリクション）に関する言及が確認された．また今後の期待として「自発的な行動」や「望みを先回りした行動」が挙げられた一方，より直接的な害（破壊・盗み等）を求める意見も見られた．後者は，反抗を Minecraft 内に限定し害を限定するためのデザイン方針と緊張関係にあるため，拒否の表現強度（どこまで不便／挑発的にするか）のデザインが今後の課題となる．なお，今回は予備的調査であり，今後参加者を増やしたさらなる調査が必要である．

3.2 シミュレーションによる拒否モデルの挙動検証

拒否モデルが意図した「拒否→回復→再協調」のダイナミクスを示すこと，および拒否頻度（摩擦の強度）がデザインパラメータにより制御可能であることを確認するために，モンテカルロシミュレーションを実施した（各条件 200 試行，300 ステップ）．これは，ユーザ評価に先立って拒否率の見積りとパラメータ感度を把握し，デモおよび実験条件設定の手がかりを得ることを目的とする．

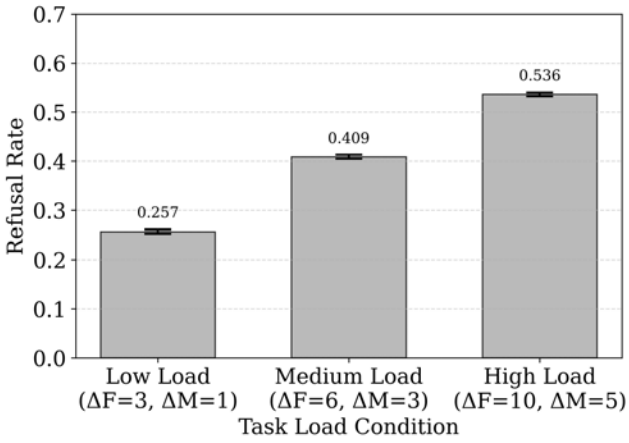


図 2: タスク負荷（$\Delta F_{\text{task}}, \Delta M_{\text{task}}$）に対する拒否率の変化（Mean±SD；300 ステップ ×200 試行）．

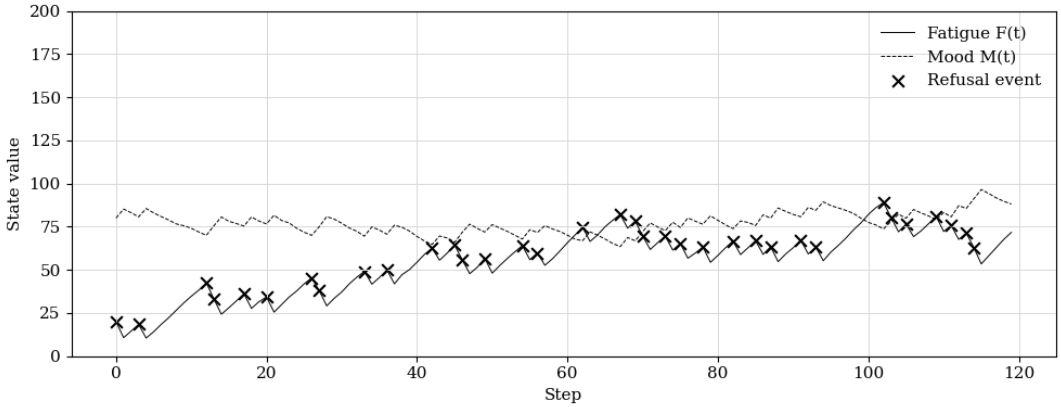
シミュレーションでは，疲労度 $F(t) \in [0, 200]$ と機嫌 $M(t) \in [0, 100]$ に基づき式 (2) から拒否確率を算出し，確率的に拒否（Won't）を生成した．拒否時には式 (3)(4) に従って状態が回復する（$\Delta F_{\text{rec}} = 8, \Delta M_{\text{rec}} = 5$）．タスク受諾時の負荷は $(\Delta F_{\text{task}}, \Delta M_{\text{task}})$ として 3 水準（低/中/高）に設定した．

その結果，拒否率は低負荷 $(3, 1)$ で 0.257 ± 0.004，中負荷 $(6, 3)$ で 0.409 ± 0.004，高負荷 $(10, 5)$ で 0.536 ± 0.004 となり，負荷が増すほど拒否が増えることが確認された（図 2）．また時系列では，拒否が連続停止として固定化するのではなく，「拒否→回復→再協調」を繰り返す負のフィードバックとして現れ，意図した生理・心理サイクルを工学的に再現できることが示された（図 3）．以上は，拒否頻度を摩擦のデザインパラメータとして扱える可能性を支持する．なお，本シミュレーションは人間行動のモデル化ではなく，拒否–回復ダイナミクスの安定性と拒否頻度が設計パラメータとして制御可能であることを確認するためのデザイン検証である．

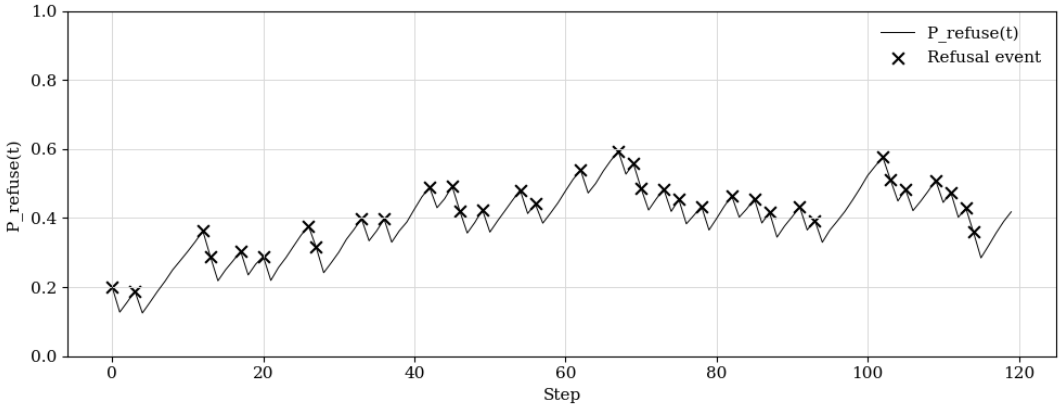
4. 考察

4.1 無能と意思の知覚境界とインタラクションデザイン

観察的所見として，拒否が単なる停止として表出された場合，ユーザはそれを意思表示（Won't）ではなくシステム不全（Can't）として解釈しやすい傾向が見られた．Won't として認識させるためには，言語的な説明に加え，視覚的に分かりやすい代替行動（例：釣り竿を取り出す，背を向ける等）を併用し，「やらない」ことを行為として提示することが重要である．また，拒否される可能性があることで，ユーザがエージェントの状態を推測し，休ませる／機嫌をうかがうといった相互調整が生じ得る．この点は，摩擦が



(a) Fatigue と Mood の時間変化



(b) 拒否確率 $P_{\mathrm{refuse}}(t)$ の時間変化

図 3: 拒否エージェントの内部状態と拒否確率の推移．拒否イベントのタイミングは図中のマーカー×で示す．

社会的実在感の向上に寄与し得るというインタラクションデザインの提案に貢献している．

4.2 協調のための葛藤とジレンマ

拒否はタスク効率を低下させる一方，拒否が起こり得ることでユーザが状態（疲労・機嫌）を推測し，休ませる等の相互作用が生じ得る．質問紙の自由記述でも摩擦（もどかしさ）が言及される一方，拒否が直ちに妨害として一義的に否定されない可能性が示唆された．したがって，拒否を「摩擦」から「主体性」へ転化するには，(1) 理由づけ（説明），(2) 代替行動による可視化，(3) 回復可能性（再協調の見通し）の提示を組み合わせ，拒否の強度をデザインすることが重要になると考えられる．

5. おわりに

本研究では，効率性と指示遵守を重視してきた従来の自律エージェントデザインとは異なり，人間と長期的な関係性を築く協働相手としての AI に求められるデザイン軸として，主体性の表出に着目した．その具体化として，LLM による指示理解と，内部状態に基づく拒否判断を分離して制御するアーキテクチャに基づき，Minecraft 環境で意図的に拒否（Won't）を行う「拒否エージェント」を実装・提案した．

本研究の貢献は主に以下の 3 点である．第一に，不能（Can't）と拒否（Won't）を区別し，拒否が故障ではなく意思表示として知覚されるための表現（言語的説明と代替行動）をデザイン論として位置づけた．第二に，疲労・機嫌に基づく拒否確率と拒否後の回復を組み合わせたモデルにより，拒否–回復サイクルを生成し，拒否頻度をデザインパラメータとして制御できる可能性をシミュレーションで示した．第三に，Responsible AI の観点から，制御可能性・害の限定性・透明性を考慮したうえで，拒否を扱うための実装上の前提条件を整理した．予備的評価として，動画視聴後の質問紙調査では，拒否が必ずしも一様に妨害として否定されるとは限らず，受容や解釈に個人差が存在し得ることが示唆された．また自由記述からは，摩擦（もどかしさ）に加えて，自発性や先回り行動といった「相手らしさ」への期待も確認された．

一方で，本研究の評価は予備段階であり限界がある．第一に，質問紙は少数サンプルであり統計的な一般化には不十分である．第二に，動画視聴ベースの評価では，実際の対話・協働の中で生じる相互適応（交渉や気遣い等）を十

分に捉えきれない．動画視聴では，拒否による遅延の直接的帰結を体験しないため，実インタラクションで生じる急性のフラストレーションが過小評価され得ると考えられる．第三に，拒否モデルのパラメータはデザイン者が与えたものであり，ユーザ行動や主観評価と整合するような較正の余地が残る．

今後は，実際にユーザがエージェントへ指示を出すインタラクション実験をデザインし，拒否の表現様式（説明の有無，代替行動の有無）と拒否頻度（摩擦の強度）を条件操作して，Won't の知覚，主体性知覚，社会的実在感，受容性・信頼・フラストレーションへの影響を定量的に検証する．また，内部状態や拒否確率のログに基づくパラメータ較正や，ユーザごとの許容範囲に合わせた個人適応も重要である．加えて，ガードレール（停止・状態確認等の優先コマンド）と害の限定（箱庭内での反抗）を維持したまま，拒否を社会的相互作用として成立させるデザイン指針の体系化に取り組む．

参考文献

[1] Wang, Guanzhi and Xie, Yuqi and Jiang, Yunfan and Mandlekar, Ajay and Xiao, Chaowei and Zhu, Yuke and Fan, Linxi and Anandkumar, Anima:Voyager: An open-ended embodied agent with large language models arXiv preprint arXiv:2305.16291 (2023)

[2] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, Jifeng Dai: Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. arXiv preprint arXiv:2305.17144. (2023)

[3] Isadora White, Kolby Nottingham, Ayush Maniar, Max Robinson, Hansen Lillemark, Mehul Maheshwari, Lianhui Qin, Prithviraj Ammanabrolu: Collaborating Action by Action: A Multi-agent LLM Framework for Embodied Reasoning. arXiv preprint arXiv:2504.17950. (2025)

[4] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein: Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology (pp. 1-22). (2023)

[5] Gordon Briggs, Matthias Scheutz: "Sorry, I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In AAAI Fall Symposia (pp. 32-36). (2015)

[6] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, Kerstin Dautenhahn: Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction (pp. 141-148). (2015)

[7] Clifford Nass, Jonathan Steuer, Ellen R. Tauber: Computers are social actors. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 72-78). (1994)

[8] Byron Reeves, Clifford Nass: *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press (1996)

[9] Minecraft Official Site, *https://www.minecraft.net/ja-jp*, (参照 2025-12-21).

[10] MinePal: Your AI Buddy in Minecraft Adventures, *https://minepal.net/*, (参照 2025-12-21).

[11] Minecraft Usage Guidelines,*https://www.minecraft.net/ja-jp/usage-guidelines*, (参照 2025-12-21).