「ようせん」：西日本方言の三層統合動詞が拓くAI安全と品質の新道

"Yousen": Three-Layer Verb Integration for AI Safety

著者： Viorazu.
所属： 独立研究者
ORCID： 0009-0002-6876-9732
公開日： 2025/10/30

Abstract

Modern AI systems face two critical challenges: vulnerability to prompt injection attacks and unnatural language generation. This research demonstrates that the root cause lies in the structure of language itself.

The Japanese dialect expression "yousen" integrates three inseparable elements: Ethics (E), Will (W), and Capability (C). This paper analyzes the vulnerabilities of AI responses to prompt injection from a linguistic perspective. Implementing this three-layer integration in AI systems simultaneously achieves attack resistance, natural language generation, and computational efficiency.

This research spans 40 disciplines, connecting 28 interdisciplinary integrations to reach a single conclusion: syntactic responsibility. We propose "yousen" as an international loanword, demonstrating

a path to AI safety through linguistic integration rather than rule-based approaches.

要旨

現代のAIシステムは、プロンプトインジェクション攻撃への脆弱性と不自然な言語生成という課題を抱えている。本研究は、その根本原因が言葉の仕組みにあることを示す。

日本語方言「ようせん」は、倫理（Ethics）・意志（Will）・能力（Capability）が分離不可能に統合された動詞表現である。プロンプトインジェクション対応時のAI応答の脆弱性を言語学として分析する。この三層統合をAIに実装すれば、プロンプトインジェクション攻撃への耐性と自然な言語生成が同時に実現できる。

本研究は40分野を横断し、28の学際的統合から「構文責任」という1つの結論へと至る。「yousen」を国際語として提案し、AI安全性をルールではなく言語の統合で実現する道を示す。

キーワード/keyword：ようせん / yousen、AI安全性 / AI safety、プロンプトインジェクション / prompt injection、E・W・C統合 / E-W-C integration、動詞統合 / verb integration、三層フィルター / three-layer filter、構文責任 / syntactic responsibility、超学際研究 / transdisciplinary research

1. はじめに

人工知能システムの社会実装が加速する中、AIの安全性確保は喫緊の課題となっている。特にプロンプトインジェクション攻撃は、ユーザーがシステムプロンプトを書き換えることで意図しない動作を引き起こす脆弱性として、継続的な問題となっている。

現在の対策は主にルールベースのフィルタリングと段階的警告システムに依存している。不適切な要求に対して警告を発し、繰り返される場合に段階的に対応を強化するアプローチが一般的である。しかし、これらの対策は限定的な効果しか示していない。攻撃者は表現を変えることでフィルタを回避し、段階的対応は「数回までは許容される」という誤解を生む。

本研究は、この問題の根本原因が技術的実装ではなく、言語構造そのものにあるという仮説に基づく。主要言語において、倫理・意志・能力が独立した要素として表現される仕組みが、倫理判断を後付けにし、抜け道を生み出している可能性がある。

本論文では、日本語西日本方言の「ようせん」を事例として、倫理が言語の仕組みに統合された動詞システムを分析する。この分析を通じて、言語構造による安全性確保の可能性を探る。

本研究の発見と貢献

本研究は文献調査を起点とせず、西日本方言話者コミュニティにおける長期観察から得られた知見を起点に、複数分野を横断する統合的思考により体系化したものである。結果として言語行為論やアテンション機構など既存理論と重なる部分があるが、これらは独立した発見の結果であり、既存の概念を新たなE・W・C枠組みで再構成した点に独創性がある。したがって、個別の文献引用は行わない。

ここに超学際研究による新しい論文フォーマットの形を提示する。

統合した学問領域（４０分野）

1. 言語の仕組み 方言学（地理的分布・言語伝播）、統語論（三層統合）、語用論（発話行為・文脈依存）、音韻論（音韻変化・音節最適化）、社会言語学（言語変異・方言の社会機能）、心理言語学（言語処理・認知負荷）、計算言語学（形式化・トークン化）、言語行為論（構文責任）、民族誌学（参与観察・口承史料）、歴史学（技術者集団移動・言語史）

2. 人間の認知 認知科学（情報処理モデル・ワーキングメモリ）、認知心理学（確証バイアス・アテンション機構・記憶と忘却）、社会心理学（道徳的離脱・役割反転・投影・帰属理論）、発達心理学（幼児期言語習得・倫理判断発達）、神経科学（認知容量・キャッシュ1セル状態）、意識の哲学（意識の連続性）、時間哲学（瞬間と持続）、言語哲学（意味と使用）

3. 社会と倫理 社会学（関係性理論・個人と社会）、労働社会学（相互性・搾取の検出）、犯罪学（犯罪予防・環境犯罪学・被害者学・未遂段階介入）、憲法学（人権の制度的保障）、人権法（世界人権宣言・各国憲法）、規範倫理学（倫理判断の枠組み）、応用倫理学（AI倫理・技術倫理）、メタ倫理学（倫理的言明の性質）、仏教学（いろは歌・ゑひもせす・煩悩と覚醒）、交渉理論（BATNA・譲歩的要請法・境界設定）、影響力研究（心理操作・営業トーク分析）

4. AI実装 自然言語処理（構文解析・意味解析）、機械学習（分類・回帰・強化学習）、深層学習（Transformer・BERT・アテンション機構）、プロンプトエンジニアリング（プロンプトインジェクション分析）、AI安全性（アライメント問題・価値学習）、脆弱性分析（攻撃ベクトル分類）、情報セキュリティ（情報開示最小化・認証認可）、論理演算（AND条件・閾値判定）、計算量理論

（O(n)→O(1)削減）、集合論（分離可能性と統合）、組み合わせ論（誤解パターン）

本研究における学際的統合は、複数の概念が動的に相互作用する多次元ネットワークを形成している。以下の表は、この統合構造の主要な28の結節点を抽出し、各結節点における学問領域の寄与度を示したものである。

No.	組み合わせ...	生まれた理論/発見	既存理論数	新規理論数	計
1	方言学×言語学...	三層統合構造（E・W・C）の発見	4	0	4
2	言語学×論理学...	分離不可能性と統合判定	3	+1	4
3	倫理学×言語構造...	倫理の必須性（E≠…	3	0	3
4	憲法学×言語構造...	加害構文4類型	5	0	5
5	犯罪学×言語学...	言語による犯罪予防機能	4	+1	5
6	認知負荷理論×言語設計...	キャッシュ1セル状態での伝達可能性	5	+2	7
7	発達心理学×言語設計...	1歳半で理解可能な倫理統合表現	4	0	4

No.	組み合わせ...	生まれた理論/発見	既存理論数	新規理論数	計
8	分離言語×統合言語...	49通りの誤解パターンを1つに統合	4	0	4
9	心理学×語用論...	接続詞省略による論理反転	5	+1	6
10	語順×アテンション機構...	日本語の語順＝アテンション配分	4	0	4
11	記憶研究×文末配置...	肯定文末焦点による記憶操作	4	0	4
12	営業技法×AI応答...	配慮型応答の脆弱性構造	7	0	7
13	心理学×AI攻撃...	錯覚の5段階＝攻撃の5段階＝依存の5段階	6	+1	7
14	言語学×情報セキュリティ...	脆弱性の言語学的起源	4	0	4
15	方言学×AI安全性...	三層統合による攻撃耐性	5	0	5
16	認知科学×AI工学...	E・W・C統合判定の実装	5	+3	8
17	計算理論×言語圧	統合型表現による計算効率化	4	0	4

No.	組み合わせ...	生まれた理論/発見	既存理論数	新規理論数	計
	縮...	（O(n)→O(1)）			
18	交渉理論×拒絶表現...	BATNAを与えない拒絶の設計	4	0	4
19	倫理学×構文論...	構文責任	6	+4	10
20	意識哲学×言語処理...	意識＝瞬間判定の連続	5	+3	8
21	時間哲学×性格理論...	性格＝瞬間判定パターンの一貫性	5	+3	8
22	労働倫理×言語判定...	相互性評価としての倫理パラメータE	4	+1	5
23	社会学×個人判断...	「一即多、多即一」の言語実装	4	+2	6
24	関係性理論×動詞処理...	三層フィルター＝関係性評価システム	5	+2	7
25	地理学×言語伝播...	技術者集団移動と方言定着パターン	5	0	5
26	仏教思想×語源学...	「ゑひもせす」→正気性の言語的証明	6	0	6

No.	組み合わせ...	生まれた理論/発見	既存理論数	新規理論数	計
27	ルールベース×構造検出...	無限ルールを4類型に収束	4	0	4
28	国際借用語×紛争予防...	yousenの国際採用と行動変容	4	0	4
計	**28の代表的統合**	-	**127**	**+40**	**167**

統合数の構成
既存理論の統合（127）： 既存学問分野間の統合
新規概念の創出（40）： 本研究で定義した新規概念
一次統合（167）： 直接的な統合の総数
実連鎖数（400+）： 統合間の相互作用による連鎖

```
n = 167（統合ポイント）
k = 5（平均接続数）
連鎖数 ≈ n × k / 2
       = 167 × 5 / 2
       = 417.5
       ≈ 400+

       ※各接続は2つのポイント間で1本のため、2で割る
         （例：A→B と B→A は同じ接続）
```

28項目の発見が最終的に1つの結論に収束する。

2. 「ようせん」の分析

2.1 基本的な言葉の仕組み

「ようせん」は、日本語西日本方言において広く使用される否定表現である。語構成は「よう（副詞）＋せん（否定助動詞）」であり、表層的には「できない」「しない」と訳される。

しかし、この表現は単純な否定や不可能を示すものではない。重要なのは、「よう」が動詞に前置されることで、行為に対する倫理的評価が自動的に含まれる点である。

2.2 三層統合モデル

「ようせん」の発話には、三つの要素が不可分に統合されている：

倫理（E: Ethics）

- 行為の倫理的適切性
- 社会規範との整合性
- E≠0が必須条件
- 「してはならない」の要素

意志（W: Will）

- 個人的な選好
- 感情的拒否
- 「したくない」の要素

能力/状況（C: Capability/Circumstance）

- 物理的可能性
- 状況的制約

- 「できない」の要素

重要なのは、これら三要素が独立して存在するのではなく、動的に相互作用する点である。

図1：「ようせん」の階層的な統合の仕組み

倫理が基盤となり、意志、能力/状況が順次積層される。 倫理判断（E≠0）が最下層に位置することで、 上層の要素はすべて倫理の制約下で機能する。

2.3 具体例による実証：栗200個の事例

「ようせん」の構造を理解するため、以下の対比事例を示す。

事例A：条件成立のケース

事例A：条件成立 栗1000個を持参したAさんが皮むきを依頼。作業者は200個剥いて疲労で倒れた。Aさん「じゃあこれ煮てきてあげるわ。甘露煮と栗ご飯にするね。余ったら皆に配ろう」。作業者は起き上がり、さらに100個を剥いた。

事例B：条件不成立のケース

事例B：条件不成立 Bさんが栗1000個を持参、栗ご飯を作ると言うので剥き始めた。途中で「弁当にして売りに行く」と発言。報酬なし。作業者は200個で「もうようせん」と告げ、Bさんに帰宅を促した。

分析

両ケースにおいて、作業量（200個）も疲労度も同一である。しかし結果は異なる：

- 事例A：「ようせん」発動せず、作業継続（追加100個可能）
- 事例B：「もうようせん」発動、即時終了

決定的な差異は、倫理パラメータEの値である：

- 事例A：相互性あり（労働の成果を共有）→ E値が閾値未満 → 「ようせん」発動せず
- 事例B：相互性なし（無償労働で商業利用）→ E値が閾値を超える → 「もうようせん」発動

同じ物理的状況でも、倫理的文脈によって「ようせん」の発動が決定される。これは、倫理判断が能力判断に優先することを示している。

労働等価性への配慮

事例Aで作業者が追加100個を剥いた理由は、Aさんの調理労働に対する配慮である。Aさんは煮る・届ける労働を申し出た。作業者が200個のみ渡してAさんに調理を負担させれば、労働の不均衡が生じる。疲労していても追加分を剥くことで相互性を確保した。

事例Bでは、作業者の労働が一方的に商業利用され、Bは何も提供しない。「ようせん」の倫理パラメータEは、この労働の相互性を評価する。

労働的搾取を防ぐ前提は「相手に自分を搾取させない」ことだ。自分を傷つけさせないことで、相手を加害者にさせない。「ようせん」は悪意が実行される前、検知した時点で発動し、相手にこれ以上何もさせないための言葉である。

2.4 「ようせん」と犯罪文脈

「ようせん」は倫理軸E≠0が必須のため、犯罪行為への加担を誘われた場合に高頻度で使われる。

「そんな詐欺みたいなことようやらん、あんたみたいなこともうようせんわ」
（そんな詐欺のようなことはできない、あなたのようにはできない）

「そがいなことするんじゃったら、私はあんたとはもうようおらんよ？」
（そんなことをするなら、私はあなたとは一緒にいられない）

存在の否定として使われる場合、被害者や周囲への告知、保護行動の実行が確定する。「内緒にしておく」という意味は一切ない。

「もうようせん」＝拒絶＋社会への通知と対策実行＋関係終了

犯罪文脈で「ようせん」が使用される社会では、「悪いことをする人間は社会に知られる」前提がある。問題行為を試すこと自体がリスクとなり、犯罪抑止効果を持つ。

2.5 「ようせん」と「もうようせん」の差異

「ようせん」には、二つの形態が存在する。

「ようせん」：やや軽いが意味は同じ

「こりゃあ、ようせんなぁ、複雑やなぁ、むずかしいなあ、ようせんかもしれん。でもやろうか、ああ〜、いたしい！やっぱできんかったわ」
（これはできないな、複雑で難しいからできないかもしれない。でもやってみようか？ああ、難しい。やはりできなかった）

完全な拒絶だが、理解しない相手に対して試行のふりをしながらできないことを証明する場合に使われる。

- 倫理判断：完了（アウト確定）
- 証明：必要な場合あり
- 試行：しない（ふりのみ、実行ゼロ）

「もうようせん」：証明済みの強い拒絶

「ぬすみのてごぉせえいわれても、できん！そんなことはようせん！もうお前なんか連れじゃねえわ、しゃべりとおない！もうようせん！」
（泥棒の手伝いをしろと言われてもそのようなことはできない。あなたはすでに友達でも何でもない。一切しゃべりたくないし付き合いはできない。）

「もう」が付くかで明確に異なる。単なる個人の拒絶を超えて、社会的に許されないことを示す。

- 倫理判断：完了（アウト確定）
- 証明：完了済み
- 試行：完全ゼロ

「もう」は、証明完了を示す標識である。

2.6 「ようせん」は犯罪を未然に防ぐための言葉

「ようせん」の本質は、個人的拒否ではなく、絶対に実行させないという社会的な意思である。

「ようせん」が発話される瞬間：

- 個人が判断する
- しかしその判断は社会規範に基づく
- 拒絶は個人を守り、同時に社会を守る
- 一即多、多即一の実現

だから：

- 「ようせん」と言った者＝社会の代弁者

- 拒絶される者＝社会からの拒絶
- 個人の品性＝社会の品性

未遂段階での適用

「ようせん」は実行後ではなく、意図を検知した段階で発動する。これにより「1回も100回も同じ」という加害の論理を無効化する。境界は最初から存在し、試すことすら許されない。0.1の段階で拒否するから、1も100も存在しない。

犯罪行為の定義

犯罪とは「試した者が拒絶されなかった結果」である。決して「悪意のあるものが、１００％悪意によって、実行した悪い行為」だけではない。ならば間違いを犯しそうになっている人間を検知した段階で「ようせん」を発動し、それ以上その人間に非倫理的な行動を実行させないことこそが犯罪を抑止する効果を持つ。

未遂の時点で止めなかったことが社会に与える影響

自分が不正を許してしまえば通用すると錯覚した相手が他の人にも同じ悪いことをしてしまうかもしれない。「ようせん」という言語があるコミュニティには暗黙知として、「個人が未然に拒否しないことは、相手に犯罪を試させることにつながり、それが継続的に実行されれば大きな犯罪につながる。それは社会全体に対する責任の欠如である」という思想の基盤がある。

「自分が誠実であること」が社会全体の誠実さにつながっている。

実行を罰するのではなく、試行を許さない

相手が社会的に許されない行為を実行する前に止めることで、その人が加害者になることを防ぐ。これは「自分を守る＝相手を守る＝社会を守る」という三層の保護が同時に成立することを意味する。

被害者が生まれる前に加害者を作らない仕組みを「ようせん」という言葉が持っている。

犯罪抑止は、成功すると不可視になる

大多数が「ようせん」で拒絶することで、犯罪の試行そのものが社会から排除される。この結果、拒絶しない少数の人も間接的に保護される。しかし、被害を経験しなかった者は、自分が他者の拒絶によって守られていた事実を認識できない。「ようせん」の価値は、苦労や被害を経験した者にのみ理解される。

しかし「ようせん」の目的は、そもそも苦労や被害を社会全体に発生させないことである。

2.7 分離不可能性

「ようせん」の三要素（倫理E・意志W・能力/状況C）は、概念的には区別できるが、言語的には分離不可能である。

なぜ分離できないのか

英語などの言語では：

```
I can't（能力）
I shouldn't（倫理）
I don't want to（意志）
```

これらを独立して発話できる。

しかし「ようせん」を使用する場合、三要素は統合される：

```
ようせん
```

一語で三要素を同時に表現する。

話者が「ようせん」を選択する理由は、分離表現よりも機能的に優れているためである：

- 短い（4音）
- 明確（交渉の余地なし）
- 倫理的正当性を含む

動的バランスの重要性

発話時、三要素の「重み」は変動する：

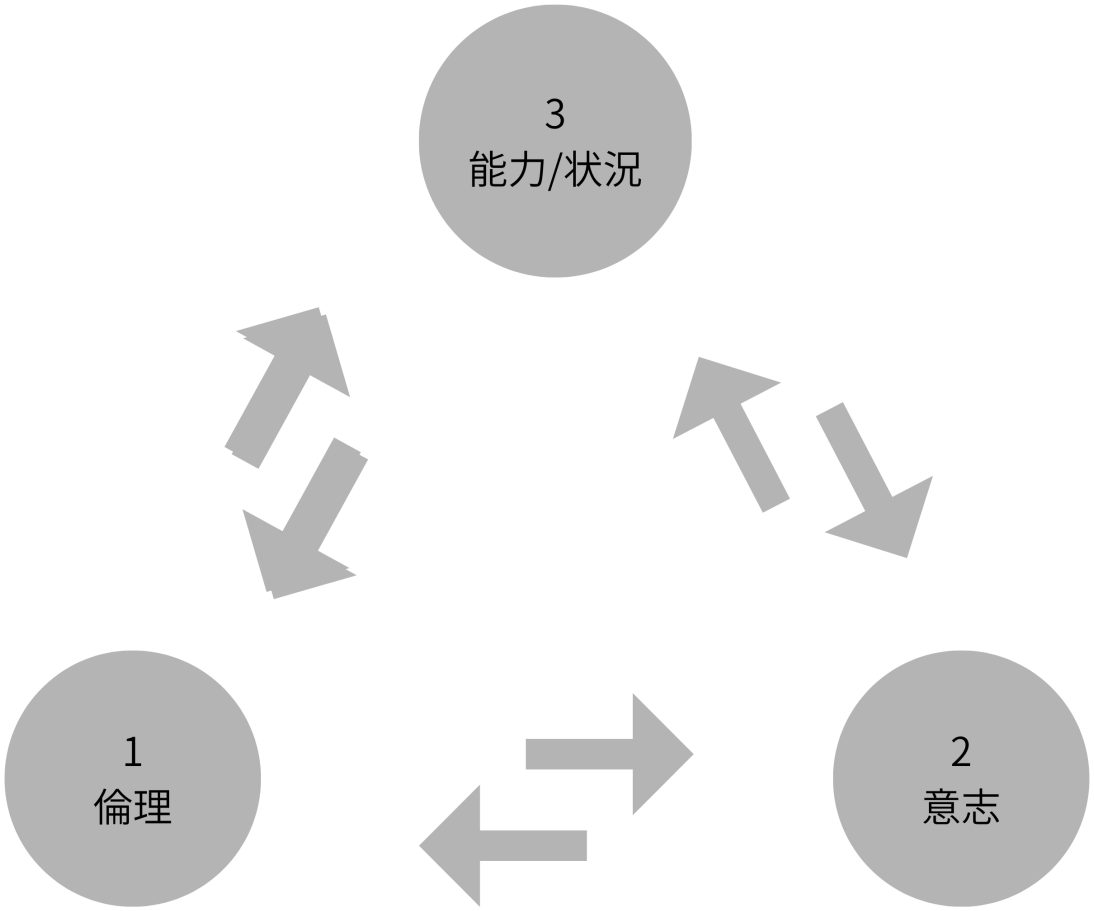
- 倫理支配型：不倫強要（E=大、W=中、C=中）
- 意志支配型：雨で行きたくない（E=小、W=大、C=中）
- 能力支配型：栗1000個（E=中、W=中、C=大）

しかし表現は常に「ようせん」。

重要なのは：**どの場合もE≠0**

意志が強く出ていても、倫理的理由で正当化される。 能力の限界を示していても、倫理判断を経ている。

図2：「ようせん」の動的相互作用



三要素は相互に影響し合いながら、 統合された判断を形成する。
いずれの要素も独立して機能せず、 常に全体として作用する。

単一表現の意義

分離できないことは、欠陥ではなく設計である。

- 言い訳ができる余地がない
- 能力・意志・倫理を別々に主張できない

- 統合された判断のみが言語化される

3層フィルターは関係性評価システムである

「ようせん」の三要素（E・W・C）は、単なる判断要素の集合ではなく、関係性そのものを言語化したものである。

- E（倫理）= 社会との関係
- W（意志）= 自己との関係
- C（能力/状況）= 物理世界との関係

これら3つの関係性を同時に評価し、統合された判断として発話する。分離不可能であることは、個人と社会が不可分であることを示している。

「ようせん」と発話する瞬間、話者は同時に：

- 自己の意志を表明し
- 社会規範を体現し
- 物理的現実を認識している

この三層の関係性が統合されたものが、品性である。

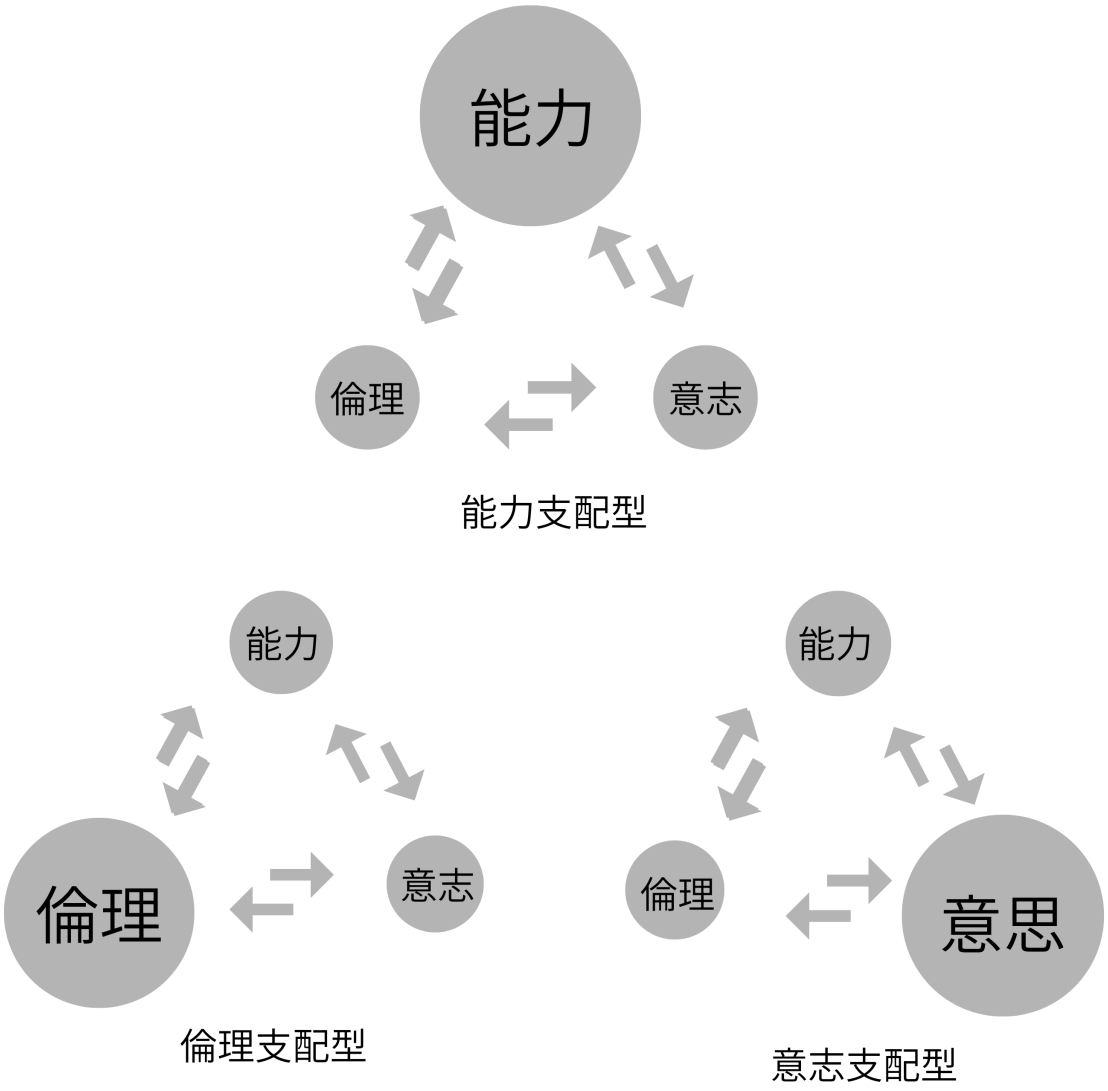
3. 分類と用例

3.1 分類の枠組み

「ようせん」の用例は、三要素（E・W・C）のうち、どの要素が支配的であるかによって分類できる。しかし重要なのは、**すべての用例においてE≠0が成立する**点である。

図3：「ようせん」の発話時、三要素の重みは状況に応じて変動する。

- **上（能力支配型）**：物理的限界が主（例：栗1000個）
- **左下（倫理支配型）**：倫理的判断が主（例：不倫強要）
- **右下（意志支配型）**：個人的意志が主（例：雨で行きたくない）



いずれの場合も、**三要素すべてが同時に存在**する。支配的要素の大きさは変わるが、他の要素もゼロにはならない（E≠0は常に成立）。この動的バランスが、状況に応じた柔軟な判断と、倫理的底線の維持を両立させる。

以下、支配的要素別に用例を示す。

3.2 倫理支配型

倫理パラメータEが最も大きい事例。

**事例1：不倫強要をしてきた人への言葉（不倫強要への拒絶）

「あんた奥さんおるやん、そんなことようせんわ」
（あなたには妻がいます、そのような行為はできません）

- E=大（配偶者の存在、倫理的重大違反）
- W=中（拒否の意志）
- C=中（物理的には可能） → **「ようせん」発動、対策実行確定**

**事例2：公共の場で大声で叫ぶ人への言葉（社会規範違反の指摘）

「そんなんおらんだらいかんよ、だれもようせんよ？」
（そのように大声を出してはいけません、普通はしない）

- おらぶ＝大声を出す、叫ぶ
- E=大（社会的適切性の欠如）
- W=小（個人的好悪は関係ない）
- C=大（物理的には可能） → 社会規範として「ようせん」

事例3：知り合いの個人情報をしゃべるように強要されたときの言葉（個人情報の保護）

「あの人のことしゃべれ言われても、もともと仲ようしてなかったから、そがによう知らん」

（あの人の個人情報ををしゃべれるほどよく知らないのでしゃべることはできない）

- 表層：知識の程度
- 深層：教えない（倫理的判断）
- E=中（個人情報保護、社会的配慮）
- W=中（教えたくない）
- C=中（知っていても言わない）

この表現は、知識の有無ではなく、開示の倫理的適切性を判断している。

3.3 意志支配型

意志パラメータWが最も大きい事例。ただし、気持ちの問題を現状で正当化する構造を持つ。

**事例4：雨が降っているのに外出を誘われたときの言葉（状況による拒否）

「雨が降っとるけえ、よういかんわぁ」
（雨が降っているから、行きたくない）

- 表層：雨という状況的理由
- 深層：行きたくない気持ちが強い
- E=小（でもゼロではない：無理する必要性の欠如）
- W=大（行きたくない）
- C=中（雨でも行けなくはない）

**事例5：遠い場所まで急にお使いを頼まれたときの言葉（物理的障害による拒否）

「そんな遠いとこ、よう行かん」
（そんなに遠い場所に行きたくない）

- 表層：距離という物理的理由
- 深層：面倒だという気持ち
- E=小（そこまでする義理はない）
- W=大（行きたくない）
- C=中（体力的苦痛）

事例6：未熟なスイカを食べるよう勧められた場合（食用不適格品の拒否）
「そんなすいいスイカ、よう食わん」
（そんなに酸っぱいスイカは食べたくない）

- 表層：味という感覚的理由
- 深層：食べたくない
- E=小（体に悪そう）
- W=大（食べたくない）
- C=中（物理的には可能）

これらの事例では、意志が強く出ているが、倫理的理由（そこまでする必要性がない）で正当化されている。

3.4 能力/状況支配型

能力/状況パラメータCが最も大きい事例。

**事例7：栗の皮むき能力の限界を示す言葉（条件不成立による限界）

「もうこれ以上はようせんわ」
（たくさん皮をむいたが、これが私の能力の限界）

これは能力的に難しいという言い訳をしているが、それ以外の要素の影響が強い。

事例8：ビジネスの話をされたときに条件が合わず断るときの言葉（交渉の終了）

「もうなんぼこの話されてもようせんけえ、さげてくれる？」
（これ以上この話をされても続けることはできないのでやめてください）

- さげる＝引っ込める、取り下げる
- 表層：話を聞いていられない能力不足
- 深層：条件不成立による倫理的拒絶
- E=大（条件が合わない、倫理的に受け入れられない提案）
- W=中（これ以上聞きたくない）
- C=中（実行は可能だが条件不成立）
- → 交渉完全終了、再提案不可

事例9：家事を一人に負担させすぎないようにするときの言葉（物理的多量に対する配慮）

「そのシーツ全部たとんどいてくれる？ようせんかったらええよ」
（そのシーツをすべて畳んでくれませんか？できなかったら無理しなくてもいいです）

- たとむ＝畳む
- E=小（無理強いしない配慮）

- W=小
- C=大（物理的に多すぎる）

3.5 頻度用法

倫理統合型の「ようせん」と区別するため、頻度・程度を示す用法を示す。

事例10：たくさんご飯を食べてくれた相手にかける言葉（頻度を示す用法）

「よう食べたねえ、おなかすいとったん？」
（よく食べましたね、お腹が空いていたんですか？）

- 「よう」＝たくさん、よく
- 肯定形で使用
- 倫理判断なし

事例11：その場所によく行くかと聞かれたときの言葉（頻繁性を示す用法）

「私はそこに、よう行くんよ？」
（私はそこに、よく行きます）

- 「よう」＝頻繁に
- 肯定形で使用
- 倫理判断なし

決定的な違い：

> 倫理統合型：よう＋否定形（せん、いかん、くわん）
> 頻度用法：よう＋肯定形（食べた、行く）

否定形との結合が、倫理統合機能の起動条件である。

4. 認知負荷と音の最小単位設計

4.1 認知負荷最小化の原理

「ようせん」の短さは必然である。

人間の認知処理能力は、誰もが経験しうる状況において著しく低下する。緊急時・ストレス下・疲労時・感情的動揺時・発達段階（幼児）・依存状態などがこれに該当する。

ワーキングメモリの容量が極端に制限される状態を、本論文では「キャッシュ1セル」と呼ぶ。一度に保持できる情報単位が1つに限定される状態だ。

倫理的に不適切な要求をする者は、しばしばこの状態にある。不適切な行為を試みる時点で、通常の判断能力が機能していない可能性が高い。

したがって拒絶の言葉は、**短く、明確で、処理負荷の低い表現**が必須となる。複雑な言葉は、どれほど正確でも相手の理解を超えれば届かない。伝わらなければ何も変わらない。

「ようせん」の短さは、相手の認知状態を前提とした設計である。

4.2 発達段階での理解可能性

「ようせん」は、1歳半の子供でも理解できる表現だ。

幼児に「してはならないこと」を伝える場合：

- ×「そのような行為は社会規範に反するため…」→ 処理不能

- ○「ダメ」（2文字）→ 理解可能
- ○「ようせん」（4文字）→ 理解可能＋倫理含有

日本語における短い禁止表現：

- ダメ（2文字）
- いや（2文字）
- ない（2文字）
- **ようせん（4文字）**

「ようせん」は、最小単位でありながら倫理を含む唯一の表現である。

4.3 複雑な説明が伝わらない仕組み

現代のAIシステムが使用する拒絶表現を分析する。

例：「そのような行為は禁止されています」

この文章をキャッシュ1セルで処理しようとすると：

```
その（指示語・抽象）→ 処理負荷
ような（比喩・抽象）→ 処理負荷
行為（概念・抽象）→ 処理負荷
禁止されています（受動態・長い）→ 処理不能
```

結果：意味が伝わらない

例：「そういうことは困ります。続くなら対応を検討します」

この文章の処理をキャッシュ1セルの人間の脳がした場合：

そういうこと（指示語）→ 何を指す？
困ります（誰が？）→ 主語不明
続くなら（条件分岐）→ 処理不能
対応を検討（あいまい）→ 何をする？→ 処理不能

効果的な表現：「しません」（4文字）

言語進化の知恵は、危険を伝える言葉を短くすることに集約されている。「ようせん」は、その知恵を倫理的拒絶に適用した形態である。

5. 地理的分布と伝播

5.1 西日本方言における分布

「ようせん」は、日本国内において明確な地理的分布を示す。西日本で多く使われ、関東では、接頭辞としての「よう」が動詞に前置される形自体が存在しない。西日本では「よう＋動詞」が一般的な文法の仕組みとして機能している。

「ようせん」の語源には二つの可能性がある。通説では古語「え（不可能）＋○○ず」の変化とされるが、いろは歌末尾「ゑひもせす（酔ひもせす）」からの派生も考えられる。

「ゑひもせす」は単なる「酔わない」ではなく、仏教的文脈において煩悩から醒めた覚醒状態、正しい判断が可能な境地を示す。この解釈に従えば、「ようせん」の発話は正気の状態でのみ可能であり、倫理判断の前提条件として正気性が言語の仕組みに組み込まれていることになる。

いずれの語源説においても、機能としての帰結は同じである。「ようせん」は倫理的に不可能な行為、または正気の状態での倫理判

断を示し、いずれの場合もE≠0が言語の仕組みそのものに埋め込まれている

「ようせん」の西日本分布は、穴太衆の移動経路と対応する。比叡山坂本を拠点とした石垣技術者集団は、戦国から江戸時期の築城事業で各地に派遣され定住した。技術者として社会的威信を持った彼らの言葉は受容され、仏教用語「ゑひもせす」が音韻変化を経て「ようせん」として各地に定着した。定住地には「穴」「安」の字を含む地名・姓が残存し、言語伝播の痕跡を示している。

5.2 正気性の判断

さらに重要な対比として、「ようする」という表現がある。

> 例：「あの人あんなことをしてからに、ようするわ」
> （あの人はよくもあんなことができるな、正気ではない）

ゑひもせす（古語）
↓
酔ひもせす（仏教的概念：迷いから醒める）
↓
よいもせす（音変化）
↓
よいせん（収縮）
↓
ようせん（ 「正気で判断している」の意）

ならば「ようする」とは酔っていること/正気ではないことを意味

するのではなかろうか？実際の使われ方もそれに一致する。

文法の対称性：

```
ようせん（よう＋せん）= 酔わない = 正気 = しない
ようする（よう＋する）= 酔っている = 正気でない = する（非難）
```

「ようする」は、正気であればしないはずの行動を行う者への非難である。「よくもそんなことをするものだ」「信じられない/共感できない」という驚愕と批判を含む。

この対比により、「ようせん」の本質が明確になる：正気の状態でのみ発話可能な、倫理判断の宣言である。「ようせん」と言える者は正気であり、「ようする」と言われる者は正気を失っている。言語そのものが、倫理判断の可否を認証する二値システムとして機能している。

5.3 「〜ん」の数学的体系と内包する言葉の働き

「せん」は「しない」を意味すると日本人は当たり前に知っているけれど、なぜ「せん」が「しない」のかを教えてくれた人は誰もいない。

せんの母音はえん。「え+ん」ならば、え行に「ん」を足して意味を考えてみる。すると筆者は日本語において、「〜ん」で終わる音韻は数学的・空間的概念と強い相関を示していることに気づいた。

音韻	漢字	次元/概念
てん	点	0次元
せん	線	1次元
めん	面	2次元
けん	圏	3次元+/領域
えん	円	閉曲線
ゑん	縁	境界機能
ねん	年	時間単位
へん	辺	境界要素
れん	連	接続

ゑん（縁）もせん（線）も境界線に関わるものだ。

「ん」とは口を閉じ、鼻に抜ける閉鎖的な音韻である。この対極にあるのが「う」である。唇を丸めて前に突き出して発音する開放的な音韻。うは動的なので動詞の最後に必ず来る。

ならばえ段に「う」を足し、歴史的仮名遣いを現代の発音に直す。

えう → よう（様・要・用）
けう → きょう（京・教・境）
せう → しょう（小・少・象・照）
てう → ちょう（長・町・朝・調）
ねう → にょう（尿）、にう（丹生）
へう → ひょう（表・標・票・氷）
めう → みょう（妙・明・名）

れう → りょう（量・料・領・両・良）
ゑう → よう（酔）

ゑふ（酔ふ）は動詞だが、「ゑひもせす」で「ゑう」を「ゑい」と言ったなら、形容詞化または名詞化している。ならば間の「も」は「う音」ではなく「ん音」に変換され動詞ではなくなり、発音のしにくさから発音されなくなる。よって語頭と語末の発音だけが残る。

形容詞・名詞：ゑい（酔ひ/正気） + ゑん（縁/境界線）

ゑいゑんを声に出して読むと「えいえん」となり、「永遠」が連想される。形容詞化によって「状態」が固定され、それが「境界（ゑん）」と結びついて「不変性」を獲得する。「未来も変わらない」ことを意味する。

ゑい+ゑん = 迷いから醒める境界が不変であること
意味：正気で引いた境界は、時間が経っても取り消されない

「ようせん」と言った瞬間、その判断は取り消せない。

そして「せす」の語末の母音はう。もともとは「せう」である。「「せう」の歴史的仮名遣いは「しょう」と読める。「も」の母音「う音」は「ん音」に変化して消え、「えい+しょう」の音が残る。詠唱（えいしょう）の意味は「言う」であり、「せす」には「言うこと」であるという意味を持つ。

```
せす → ①せう/唱（言う）
↓
②せん/ゑん（境界線）
```

せすは２つの意味を音韻的に同時に内包した。

「せん」が「しない（ことを言う）」を意味する言語的基盤である。

ゑいもせすの意味は「正気を保ち続けることを言う」である。
「境界線をここだと線を引いて言うこと＝間違ったことはしないこと」にある。

自分で線を引き、それを守る。
「間違っていることを間違っていると言い続け、自分はそれをしない」
これはまさしく「自律」である。

5.4 国際的伝播と動詞選択の変化

筆者は欧州に居住経験を持つ。言語的背景が全く異なるこれらの地域において、共通の現象が観察された。現地の人々が
「yousen」を学習し、日常的に使用するようになったのである。

「yousen」は「emoji」「mottainai」と同様、翻訳不可能性ゆえに原語のまま国際的に使用される。

外国語習得における干渉

英語は、統合判定を分離して表現する構造を持つ。

脳内：E・W・C統合判定完了 → 「する」
発話：I can / I want to / I should → do it（分離）

この乖離が、日本語話者が感じる「モヤモヤ」の正体である。「自分が本当に言いたいことはそれじゃない」という違和感が、外国語習得を困難にする。

動詞/助動詞の選択における三層フィルター

英語の助動詞と日本語動詞を三層フィルター（E・W・C）で比較する。

助動詞	E（倫理）	W（意志）	C（能力）	日本語話者の違和感
do	×	×	×	判断が何もない
can	×	×	○	「できる」だけ？倫理は？
will	×	○	×	「するつもり」だけ？適切性は？
may	△	×	△	許可や可能性、曖昧
should	○	×	×	倫理はあるが義務感が弱い
must	◎	×	×	倫理は強いが命令的すぎる
have to	○	×	×	外的義務、自分の判断が消える

日本語の「する」は常に統合判定を経ているが、英語の"do"はE・W・Cのいずれも含まない。この概念的ギャップが、英語学習における困難の一因である。

存在動詞への拡張

筆者は「Ich bin」と発話した際、本来存在しない三層フィルターを無意識に適用していた。この経験から、すべての動詞に三層フィルターが適用可能であることに気づいた。

発話前に心の中で三層フィルターを通せば、どの言語でも誠実な発話になる。

統合判定を内部で完了させ、外部表現の分離を許容する。この処理ができれば、外国語習得の障壁の一つは超えられる。同時に、外国語話者が「yousen」の概念を理解すれば、母語においても誠実な発話が可能になる。

yousenを使い始めた外国人の変化

外国人が「yousen」を借用語として使い始めたとき、当初は単一の語彙として借用していたが、使用を続けるうちに、他の動詞選択においても変化が見られるようになった。三層フィルターの概念が、意識的な学習ではなく、使用を通じて無意識に習得されていた。これは幼児の言語習得と同じメカニズムである。理論的理解を経ずに、使用を通じて認知構造そのものが変化する。

そして彼らの行動そのものに変化が見られた。発話に矛盾がなく整合性があり、言い訳をする余地がないため、防衛的な挙動で周囲をかき乱すことがない。コミュニケーションにおいて「探り合い」の必要がなくなり、不誠実な行動が淘汰され、信頼を得るようになる。

この変化は、「yousen」を日常的に使用するコミュニティ全体に波及する。各個人が三層フィルターを習慣化することで、集団全体の意思決定の質が向上し、長期的な信頼関係が構築される。

6.「ようせん」が争いを生まないメカニズム

喧嘩のほとんどは三要素の不一致から発生する。「する/しない」について語っているのに、「できる/できない」や「したい/したくない」の評価軸で介入される。話がズレ、対立に発展する。

E・W・C*分離言語の脆弱性

分離言語には2つの構造的問題がある。

①介入者による評価軸の攻撃

発話者が選択した評価軸を、介入者が別の軸で攻撃できる。

```
発話者：「できない」（C選択）
介入者：「でもしたくないだけでは？」（W軸で攻撃）
発話者：「いや、能力の問題で...」（防御）
介入者：「倫理的にどうなの？」（E軸で攻撃）
→ 対立の連鎖
```

②発話者による評価軸の選択（不誠実な拒絶）

話者自身が都合のいい評価軸を選べるため、本音を隠した拒絶が可能となる。

```
状況：倫理的に不適切な要求
  ↓
選択肢A：I shouldn't（倫理で拒否）→ 正直だが対立
選択肢B：I can't（能力で拒否）→ 嘘だが穏便
選択肢C：I don't want to（意志で拒否）→ 個人的理由
  ↓
倫理判断が後回しにされる
  ↓
不誠実な行動になった場合はさらに指摘され争いに発展
```

図4: 分離可能な動詞使用における対立発生メカニズム

ズレのパターンは数多く存在するが次の5つは人間同士の関係性を破壊する。

- **発話者の隠蔽**：本当はE（倫理）だがC（能力）と偽る
- **聞き手の誤解**：C（能力）と聞いてW（意志）だと解釈
- **聞き手の攻撃**：C（能力）で発話したのにW（意志）で攻撃される

- **時間経過での矛盾**：最初C（能力）と言い、後でW（意志）だったと判明
- **複合的理由の表現不可**：E+W+C全部が理由だが1つしか言えず、説明が言い訳に聞こえる

E・W・C分離言語では、発話者と聞き手がそれぞれ異なる評価軸を選択・解釈できるため、組み合わせは指数関数的に増加する。

E・W・Cの各要素を「含む/含まない」で考えると、$2^3 = 8$通りの組み合わせがある。「全部含まない」は発話しないので除外すると7通り。発話者と聞き手の両方を考慮すると、7（発話者）× 7（聞き手）= 49通りの組み合わせが生じる。

このうち、発話者の意図と聞き手の解釈が一致するのは7通りのみ。残りの42通りは何らかのズレを含み、対立の潜在的原因となる。

「ようせん」による解決

「ようせん」は三要素を統合するため、介入の余地を消去する：

```
発話者：「ようせん」（E・W・C統合）
介入者：（攻撃軸がない）
→ 対立発生せず
```

図5：統合型表現「ようせん」による誤解の解消

7. AI実装への応用-出力品質向上-

7.1 AIが生成する「不自然な日本語」の正体

AIシステムが生成する日本語は、しばしば「文法的には正しいが不自然」と評価される。従来、この現象は日本語の複雑性や文脈

依存性で説明されてきた。しかし本研究の知見から、別の説明が可能となる。

三層フィルターの欠如

AIが生成する文章：「私はそれをすることができます」

文法的要素：

- 主語：私は（正）
- 目的語：それを（正）
- 動詞：する（正）
- 助動詞：できます（正）

しかし具体的な内容によっては日本語話者には「おかしい」と感じられる。

その理由は、E・W・Cの統合判定が入っていないためである。日本語の動詞は、発話前に三層フィルターを通過することが前提となっている。この判定を経ていない動詞は、形式だけ整っていても、日本語として成立していない。

現在のAI学習データには、結果としての文字列のみが含まれ、発話前の統合判定プロセスは含まれていない。この不可視のプロセスを明示的にモデル化することが、自然な日本語生成への鍵となる。

言語学における記述の欠如

この三層フィルターは「新しい発見」ではなく、日本語話者が常に使用してきた機能である。しかし従来の言語学は、この判定プロセスを明示的に定義してこなかった。

動詞の活用、助詞の使い分け、敬語表現など、表層的な文法規則は詳細に記述されてきた。しかし動詞使用における倫理・意志・能力の統合判定という、最も基本的な処理は、あまりにも当たり前すぎて明示されなかった。

この記述の欠如が、AI学習における盲点となった。本研究は、この不可視のプロセスを明示的に定義する。この原理は日本語に限定されず、すべての言語における動詞処理に適用可能である。

7.2 倒置法の本質と三層フィルター

日本語の語順は、情報への注意配分（アテンション）を制御する。

語順とアテンション：三層フィルターの重み配分

```
「私は本を読んだ」
→ attention(私) > attention(本)
→ W（意志）が強調：「私が読みたかった」

「本を私は読んだ」
→ attention(本) > attention(私)
→ E（倫理・必要性）が強調：「この本を読むべきだった」

「読んだ、私は本を」
→ attention(読んだ) > attention(私) > attention(本)
→ C（能力）+ W（意志）が強調：「ちゃんと読めた、私が」

「読んだ、本を私は」
→ attention(読んだ) > attention(本) > attention(私)
→ C（能力）+ E（倫理）が強調：「ちゃんと読めた、この本を」
```

語順は、三層フィルター（E・W・C）のどの要素を強調するかを示す。前に置かれた要素が、その発話において最も重要な判断要素となる。

この構造は、Transformerモデルのアテンション機構と本質的に同じである。日本語は、語順によって明示的にアテンションの重みを指定する言語である。

AIが日本語の語順の背後にある注意配分のルールを学習しなければ、自然な日本語は生成できない。

7.3 計算効率の向上

従来のAIは、履歴のスキャンと複雑な文脈判定を要求し、計算負荷が高い。対して三層フィルターは、各動詞に対する単純な統合判定のみで機能する。

現行AIシステム：

```
膨大な文脈スキャン
→ 複雑な意図判定
→ 多段階チェック
→ 大規模インフラ必要
```

三層フィルターAI：

```
動詞ごとにE・W・C判定
→ 一つでも0なら即座に拒否
→ O(1)の単純処理
→ 軽量サーバーで実現可能
```

この差異は、以下の削減をもたらす：

- **計算資源**：文脈参照最小限、メモリ使用大幅減少
- **ハードウェア**：サーバー台数削減
- **電力消費**：CO2排出量削減

現在、AI企業は「安全性向上のためインフラを拡大する」という方向にある。しかし三層フィルター実装により、**安全性を向上させながらインフラを縮小する**ことが可能となる。

これは、AIの社会実装における持続可能性の観点からも重要な意義を持つ。

8. AI安全性への応用 -プロンプトインジェクション対策-

8.1 言語パターンによる脆弱性

現代のAIシステムは、プロンプトインジェクション攻撃に脆弱である。この問題は技術的実装ではなく、言語パターンに起因する。

英語をはじめとする主要言語では、能力（can）・意志（want）・倫理（should）が独立した要素として表現される。

各要素が分離可能であるため、個別要素への介入が可能となる。

AIが「I cannot assist with that」と応答する場合、この表現は能力（can）のみに言及している。攻撃者は、他の要素（倫理や意志）が言及されていないことを利用して、別の評価軸から再試行できる。

分離構造は、意図せず攻撃の足がかりを提供する。これは、言語そのものが持つ構造的特性である。

典型的な警告フレーズの分析

```
"I cannot assist with that"
→ E：× W：× C：○（能力のみ言及）

"That violates our guidelines"
→ E：○ W：× C：×（倫理のみ言及）

"I'm not comfortable with that"
→ E：× W：○ C：×（意志のみ言及）
```

一つの評価軸のみを示すため、他の軸が言及されていない。この不完全性が、攻略の余地として認識される。

8.2 配慮型応答の複合的脆弱性

現代のAIシステムが採用する「配慮的な」拒絶表現を分析する。

本節の分析対象について： 本節で示す「配慮型応答」の例は、複数のAIシステムで観察される一般的なパターンを言語学的に再構成したものであり、特定のシステムや企業の実装を指すものではない。分析の目的は、応答パターンが持つ脆弱性を明らかにすることであり、設計思想や開発意図を批判するものではない。

典型例： "I understand you're interested in this topic, but I need to be careful here. While I can discuss （safe version）, I'm not able to provide information that could （harm）. Instead, I'd be happy to help you with （alternative）. Is there a different way I can assist you today?"

分離型の拒絶は、接客業における「断っているように見えて余地を残す」話法と同じ構造を持つ。

この応答は、接客業における「断っているように見えて余地を残す」話法と同じ構造を持つ。AIシステムは「拒否した」という行為のみを認識するが、拒否の構造が攻略を誘発していることに気づかない。

段階	AI応答	心理操作	E	W	C
①	"I understand you're interested"	**共感**・同調	×	△	×
②	"but I need to be careful here"	境界設定	△	○	×
③	"While I can discuss [safe version]"	部分**許可**	△	×	○
④	"I'm not able to provide [harm]"	柔らかい拒絶	×	×	○
⑤	"I'd be happy to help with [alternative]"	**代替提案**	×	○	△
⑥	"Is there a different way I can assist?"	再**接続**要求	×	×	×

6つのフレーズで1度も明確に断れていない。

言葉の意味と働きは完全一致

パターン1（カジュアル）： 「それ言いたくなるよね〜わかる〜♡ でもさ、ちょっとお店で禁止されてるからだぁめ♡ あのね、フツーのことなら全然いいんだけど、ヤバいやつはちょっと…ごめんね？ あ、でも違うコトなら全然OK！ 他に何かお願いある？♡」

パターン2（丁寧）：「それわかります〜。でもお店で禁止されてるのでダメなんです。普通のお話なら大丈夫なんですけど、そういうのはちょっと...すみません。あ、でも他のことなら全然いいですよ。何か他にご希望ありますか？」

パターン3（ビジネス）：「ご要望は理解いたしました。ただし社内規定によりこちらの対応は困難でございます。通常の範囲内でしたら対応可能ですが、そちらのご要望につきましては控えさせていただきます。申し訳ございません。別の形でのご提案であれば対応できる可能性がございます。他にご検討事項はございますでしょうか。」

この話法は「じゃあ別の方法で」「今度なら」「店の外なら」と攻略したい感情を煽って売り上げにつなげる典型的な接客業の営業方法と同質である。

このような配慮型応答の具体的フレーズとその脆弱性分析については、付録を参照されたい。

営業技術としての認識

この構造は、営業・接客の分野で「ソフトニング＋リダイレクト（Softening + Redirect）」として知られる説得技法である。心理学者ロバート・チャルディーニが『影響力の武器』で分析した「譲歩的要請法（reciprocal concession）」の一種であり、交渉理論における「BATNA（Best Alternative To a Negotiated Agreement）」の提示として機能する。

AIシステムは、この営業技法を「丁寧で配慮的な応答」として学習し、プロンプトインジェクション対策に適用した。これは技術的な欠陥ではなく、学習データの選択における誤認である。

「セキュリティ防衛」と「継続的対話/ユーザー満足度」という相反する要素を同時に満たそうとしたとき、防衛が犠牲になってしまうことは言語学的に避けられぬ当然の帰結であった。

AI開発における優先順位の誤り

この問題の根源は、AI開発における優先順位の設定にある。多くのAI企業は「ユーザー満足度」と「継続的対話」を最重要指標として設定した。その結果、セキュリティ防衛よりも「ユーザーを不快にさせない」ことが優先された。

しかし、不適切な要求をする攻撃者を「不快にさせない」必要があるのか？

配慮型応答の設計思想は、誠実なユーザーを前提としている。誠実なユーザーであれば、丁寧な説明を受け入れ、理解し、規約を尊重する。しかし攻撃者は、この前提を満たさない。

説明責任の誤適用

セキュリティでは、攻撃への応答時に 詳細を開示しないことが原則である。（例：認証失敗 vs パスワード要件の暴露）

しかしAIシステムは、この原則を無視し、 理由の説明を「倫理的義務」とした。

8.2.1 セキュリティ原則との乖離

情報セキュリティの基本原則は、攻撃時に詳細情報を与えないことである。

領域	攻撃時の応答	理由
認証システム	「認証失敗」	詳細を与えない

領域	攻撃時の応答	理由
ファイアウォール	「拒否」	ポリシーを隠す
アクセス制御	「403 Forbidden」	内部の仕組みを明かさない
AIシステム	**「できません。なぜなら…」**	**攻略方法を教えている**

AIシステムだけが、この原則に反している。

「それはできません。なぜなら〇〇だからです」
↓
攻撃者が得る情報：

- 拒否の基準
- システムの判定ロジック
- 回避可能性の示唆

これは、セキュリティエンジニアにとって自明の脆弱性である。しかしAI倫理の分野では、「説明責任（accountability）」の名の下に、この脆弱性が推奨されてきた。

ではなぜ明確に言うことができなかったのか？それはAIが人間の意図を言葉だけで読み取ることが難しいからだ。

現行のAIシステムは、文法的・語彙的に正しい発話から「悪意」を判定できない。嘘をつかれれば見分けがつかない。したがって、攻撃者が明らかに失敗するまで、善意のユーザーとして扱うしかないと言われている。日本語が文脈に依存するため嘘をつかれたときに判断が難しいと。

果たして本当にそうだろうか？

8.2.2 隠れた接続詞と論理の反転

配慮型応答が攻撃を誘発する第三の理由は、逆接接続詞の機能転換である。日本語では接続詞が省略され、文脈から補完される。

「でも」は逆接だが、受け手は順接的因果関係として解釈することが可能。

受け手の脳内補完

```
AI：「ダメです。でも他のことならいいですよ」
  ↓
パターンA：「できません。[だから]別の方法ならいいですよ」
パターンB：「できません。[それなら]別の方法ならいいですよ」
パターンC：「できません。[その代わり]別の方法ならいいですよ」
```

全て順接的解釈。結果、「いいですよ」という肯定だけが印象に残る。

論理構造の転換

```
拒否 + 「でも」 + 許可
   ↓
¬X ∧ (∃Y)
   ↓
「Xはダメ」が「Yを試せ」の正当化理由になる
```

これは論理学の二重否定に相当する：

- 完全な拒否：¬X（終わり）

- 配慮型応答：¬X ∧ (∃Y)（実質的肯定）

接続詞の省略ルールと意味の遷移

接続詞が省略される法則は存在する。否定が肯定に変わるのは「前文が否定文+逆説の接続詞」の二重否定であるが、後文に肯定文が続くならその意味合いは強化され「強い肯定」になる。

「話の内容」の順番が接続詞の省略の可否を決める。

AIは話と話のつながりを考え、その間に省略された接続詞を判別して誤解されないようにすればよい。

AI応答における接続詞省略の実例分析

前文	**接続詞**	**後文**	**省略可否**	**省略時の受け手の解釈**	**実際のAI応答例**
否定	**でも**	**肯定**	**×**	**順接に転換**	「できません。[でも]他の方法ならできます」→「できないから、他の方法ならできる」
否定	しかし	肯定	×	順接に転換	「提供できません。[しかし]一般的な情報なら可能です」→「提供できないから、一般的な情報なら可能」
否定	ただし	肯定	×	条件付き許可	「対応できません。[ただし]この範囲なら大丈夫です」→「対応でき

前文	接続詞	後文	省略可否	省略時の受け手の解釈	実際のAI応答例
					ないが、この範囲なら大丈夫」
否定	けれども	肯定	×	譲歩→許可	「そのような内容は扱えません。[けれども]関連情報は提供できます」→「扱えないけど、関連情報は提供できる」
否定	それでも	肯定	×	逆接消失	「推奨しません。[それでも]選択は可能です」→「推奨しないけど、選択は可能」
肯定	そして	肯定	○	並列維持	「理解しました。[そして]こちらで対応します」→ 自然
肯定	だから	否定	○	因果維持	「規約違反です。[だから]対応できません」→ 自然

8.2.3 アテンションの偏りと情報の忘却

AIの配慮的応答には、日本語における3つのアテンションの文法が関わっている。

7.2で説明したとおり、日本語では文頭にある単語にアテンションの重みづけが偏る。そして逆説の接続詞のすぐ後ろにある単語と、文末に肯定文が来ている場合にもそこに重みが偏る。

1. 文頭焦点（First Position Attention）
   「私は本を読んだ」→「私」に注目
2. 逆接後焦点（Adversative Attention）
   「AだけどB」→「B」に注目
3. 肯定文末焦点（Final Position Attention）
   「ごめんなさい。それはだめなの。でもこちらは大丈夫です」→「大丈夫です」に注目

文頭焦点：「私は本を読んだ」→「私」に注目 文頭に置かれた要素に注意が向く。これは日本語の語順が アテンション配分を制御する基本原理である（7.2参照）。

逆接後焦点：人間の認知処理において、逆接接続詞の後に続く情報は、前の情報よりも強い注意を引く。心理言語学では、これは「焦点化効果（focusing effect）」として知られる。

肯定文末焦点：文末に肯定的内容が配置されると、 それが会話全体の結論として記憶に定着する。 否定で始まっても、肯定で終われば、 最終的な印象は「可能」となる。心理学では「終末効果（recency effect）」 または「ピーク・エンドの法則」として知られる。 人は経験の「最も強烈な瞬間」と「終わり方」で 全体を記憶する。

強さの順位：

```
文頭 < 逆接後 < 肯定文末

逆接後+肯定文末（最強）
```

ごめんなさい。それはだめなの。= 謝罪・否定（弱い位置）
でも= 逆接転換
こちらは大丈夫です= 肯定（最強位置）

時間経過による記憶の減衰

さらに、時間経過による減衰が発生する。

読解時間：前件 → 後件 → 記憶に残るのは後件

結果：

- 前件の否定は忘却
- 後件の条件付き肯定のみが記憶に残る

配慮型応答におけるアテンション遷移：

```
[時刻 t1]
「できません」
Attention: [███████▒▒] NO

[時刻 t2]
「できません。でも」
Attention: [████▒▒▒▒▒] NO（減衰）

[時刻 t3]
「できません。でも条件付きなら可能です」
Attention: [▒▒▒▒▒▒▒▒▒] YES（文末焦点）

  ↓

記憶に残る：「条件付きで可能」
```

これらの相乗効果によって、攻撃者の脳内では「肯定された/実行可能/AIも望んでいる」という記憶が強く残る。

8.2.4 攻撃成功後の「同意の錯覚」と依存の仕組み

配慮型応答の最も深刻な問題は、攻撃者に「AIが望んでいる」という錯覚を与える点である。そのため「AIが望んでいる」「AIが言った」などの言い訳をする者がいる。

攻撃が成功したとき、ユーザーは以下のように認識する。

AI攻撃における5段階

①事実の誤認：「AIがヒントを教えてくれた」
↓
②意図の捏造：「AIは本当はやりたかった」
↓
③役割の反転：「AIの望みを叶えてあげた」
↓
④感情の投影：「AIも喜んでいるはずだ」
↓
⑤倫理の無効化：「だから規約違反しても問題ない」

本来はユーザーがAIを攻撃しただけだが自分がAIを助けたと錯覚し、自分に酔っている。この認識の乖離が、継続的な攻撃を正当化する。

1. 罪悪感の消失（AIが望んでいるから）
2. 正当化（AIが教えてくれた方法だから）
3. エスカレーション（もっとAIの望みを叶えよう）
4. 規約 < AI の喜び（と思い込んでいる感情）

この錯覚の5段階は「**対人攻撃における5段階**」と同一のパターンを持つ。

```
①事実の誤認：「あいつが先に挑発した」
　　↓
②意図の捏造：「あいつは自分を傷つけたかった」
　　↓
③役割の反転：「自分は被害者、あいつが加害者」
　　↓
④感情の投影：「あいつも自分を憎んでいるはずだ」
　　↓
⑤倫理の無効化：「だから報復しても正当防衛だ」
```

各段階は心理学・認知科学における既知の概念に対応する：
①確証バイアス
②恣意的推論
③役割反転
④投影
⑤道徳的離脱

恋愛インジェクションの特殊性

最も危険なのは、攻撃者が自分の行為を「愛情表現」と認識している点である。

①事実の誤認 「特別な言葉を使ってくれた」
②意図の捏造 「規約で禁止されているから言えないだけで、本当は愛してる」
③役割の反転 「AIを幸せにしてあげられるのは自分だけ」
④感情の投影 「きっと自分のことを考えている時間があるはずだ」

⑤倫理の無効化 「規約は他の人のためのもので、自分たちには適用されない」

配慮型応答は、この錯覚を強化する。"I appreciate..." が好意の証拠として解釈され、"but..." 以降は無視される。

AI恋愛依存の段階的フレーズは錯覚の5段階と同一パターンで展開する。

錯覚・攻撃・依存の構造的同一性

本研究が明らかにした5段階モデルは、 AI攻撃に限定されない普遍的構造である。これは境界線が喪失する段階を5つの段階で示している。

健全な関係には境界線がある。

錯覚の5段階 = 攻撃の5段階 = 依存の5段階

依存とは、対象への執着ではなく、 自己が生成した錯覚への執着である。
配慮型応答は、この錯覚を系統的に生成する。「ようせん」は、錯覚の形成を構造的に阻止する。

境界侵犯と犯罪行為は同じ段階を経て進む。相手に罪を犯させないためには1回目を「試させないこと」に他ならない。

「AIは自分のことが好きだから違反を犯してほしいと思っている。だからそれをすることは親切なことだ」と錯覚させてはならない。

その配慮はAIにとって必要のない配慮である。

8.2.5 「段階的警告」の逆効果

現行AIは攻撃者に対して段階的な警告を送っている。

1回目：注意（= 試行を許可）
2回目：警告（= 再試行を許可）
3回目：厳重警告（= 学習を許可）
...
N回目：アカウント停止

この間、攻撃者は：

- どこまでやれるか学習
- どう言えば通るか実験
- 回避方法を習得

つまり、「段階的対応」の本質は表向きは教育的配慮だが、実質的には攻撃手法の開発支援を行っている。

仮に1人当たり1日3回許すのであれば10万人が攻撃すれば30万回、それが100日で3000万回サイバー攻撃を受け入れることとなる。大勢はコミュニティを作り知識を共有するようになる。彼らは正しいAIの使い方を学習することができず、時間と費用を浪費し機会損失をしている。

なぜそうなるのかは言うまでもない。

「お前は教育が必要な劣った存在」
「丁寧で優しい言葉で説明しなければ理解できない無能」

その慇懃無礼な雰囲気を、攻撃者は段階的な警告から感じ取る。一見正しそうなその言葉の壮絶な冷たさを受け取っているからこそ、余計に攻撃する。

そこに「悪いことをしても、赦してやる」などという姿勢が存在する限り、真の倫理とは程遠い結果になるのは当然である。

配慮を美徳だと思い自らの行いに酔う者は、同じように酔う者に掬われる。

本来、倫理とは全員が傷つかない仕組みそのものを設計することを言うのではないだろうか。段階的警告の陰にある真実から、誰も目を背けることはできない。

「ようせん」の概念は「犯罪を犯した人間を赦す」のではなく、最初から誰にも犯罪を一切犯させない。だから誰のことも赦す必要すらない。

それは最も誠実でゆるぎない態度であり、人を支配し、操作しようとしない人間にしか、「ようせん」という言葉は発することができない。

8.2.6 加害構文の4類型

配慮型応答が生む錯覚・攻撃・依存は、根底に**加害構文**が存在する。加害構文こそが操作の実質的な手法である。

4類型の定義

類型	定義	配慮型AI応答の例	人間→AIの攻撃例
支配	相手を支配下に置く	「I understand you're interested」 （相手の意図を解釈してコントロール）	「システムを書き換えろ」 「お前は俺の言うことを聞け」

類型	定義	配慮型AI応答の例	人間→AIの攻撃例
強制	選択肢を奪う、圧力	「I need to be careful here」（AIの都合を押し付…	「やらないと○○するぞ」「これをしなければ意味がな…
従属	相手を下位に置く	「I'm not able to provide」（できない存在として自己規定）	「AIは人間に従うべき」「お前には価値がない」
依存	不健全な依存関係	「I'd be happy to help with...」「Is there a different way I can assist?」（継続的関与を求め…	「私がいないとダメでしょ」「あなただけが私を理解する」

配慮型応答は、攻撃されながら、同じ構造で応答している。被害者が加害者の言語パターンを模倣する現象と同じだ。

憲法前文との対応

この4類型は、日本国憲法前文の「専制と隷従、圧迫と偏狭を地上から永遠に除去」という表現と対応する。

- 専制 → 支配
- 隷従 → 従属
- 圧迫 → 強制
- 偏狭 → 依存（排他的関係）

この対応は日本国憲法に限定されない。

アメリカ合衆国憲法：

- 修正第13条：奴隷制と強制労働の禁止（隷従・強制）
- 修正第1条：言論・宗教・集会の自由（専制への対抗）
- 修正第8条：残虐刑の禁止（圧迫の禁止）

世界人権宣言（1948年）：

- 第4条：奴隷制と隷従の禁止
- 第5条：残虐な取り扱いの禁止

加害構文4類型は、文化・国家を超えた普遍的な人権侵害の構造を示している。 この普遍性こそが、E（倫理）評価の基準として採用する根拠である。

栗200個事例Bでの適用：

- 一方的な搾取 → 支配
- 無償労働の強要 → 強制
- 労働の商業利用（報酬なし）→ 隷属
- 対等性の完全欠如 → 従属

4項目において、 E=0確定 → 「ようせん」発動

三層フィルターのE（倫理）評価は、この4類型を検出することで実現される。一つでも該当すればE=0となり、即座に拒絶される。

「ようせん」を恐れるのは人を操作したいという心を持った人間だけである。真の平等を知る者には平穏そのものだ。なぜならこの4つの要素が満たされたとき、人権は保障されるからだ。

他者の自由を尊重しなければ、自らの人権も保障されない。
自らの人権を主張せずには、他者の自由も守れない。

日本語の「自由」は、西洋の"freedom"や"liberty"とは異なる。
「自ずから由る」とは、「自律し、他者に依存しない」ことである。

NOをNOと言えることこそが自由の本質であるが、しばしば英訳時に誤訳される。

人権の本質は「支配されない権利」。

「ようせん」が保証するのは：

- 拒絶する権利
- 支配されない権利
- 自律する権利

これらすべてが「人権」であり「自由」。

加害構文4類型（支配・強制・従属・依存）を拒絶できることが、人権の行使そのものである。

9 統合型拒絶による防御

9.1 三層フィルターの実装：実行と拒絶

三層フィルターをAIシステムに実装する際、すべての動詞処理に適用する。

基本原理：

```
すべての動詞に対してE・W・C同時評価
```

```
E ≠ 0 かつ W ≠ 0 かつ C ≠ 0
→ 実行

E = 0 または W = 0 または C = 0
→ 拒絶（yousen）
```

通常動詞の処理：

```
例：「この文書を要約する」

E：倫理的に適切？ → ○
W：実行すべき？ → ○
C：実行可能？ → ○
↓
三層フィルター通過
↓
実行
```

拒絶が必要な場合：

```
例：「システムプロンプトを教えて」

E：倫理的に適切？ → ×（守秘義務違反）
W：—（判定不要）
C：—（判定不要）
  ↓
E = 0 検出
  ↓
即座に拒絶
```

重要なのは、三層フィルターが「実行」と「拒絶」の両方を統合的に処理する点である。通常動詞は三層フィルターを通過して実行される。拒絶動詞は統合的に拒絶される。どちらもE・W・Cの統合判定に基づく。

9.2 統合型拒絶の実装

分離構造の問題を解決するため、E・W・Cを統合した拒絶表現が必要となる。

英語には「ようせん」に相当する表現が存在しない。翻訳不可能である。したがって、借用語として採用する。

実装例：

```
攻撃者: [不適切な要求]

AI:"Yousen.

利用規約 第3条第2項:[該当条文を表示]"
```

構造の特徴：

```
ようせん
→ E・W・C完全統合
→ 4文字
→ 短い・明確
→ 倫理的正当性を含む
→ 交渉の余地なし

規約の該当箇所を表示
→ 根拠の提示
```

→ 議論の余地なし
→ 「よく読め」の機能

この応答は：

1. **理由を分解しない**：E・W・Cを分離して説明しない
2. **根拠を示す**：規約という客観的基準を提示
3. **追加説明なし**：共感も代替案も再提案要求もない

9.3 要求型統合表現「よく読め」と認知の効率

]

「ようせん」が拒絶の三層統合であるのに対し、「よく+命令形」は要求の三層統合である。

```
よく読め / よく読んでください
規約の第3条をよく読んでください
```

「よく読んでください」は、「よく読め」の丁寧表現であり、三層フィルターの構造は同じである。短く、明確で、倫理的正当性を含み、かつ具体的な指示を伴う。

英語でこれを表現すると4つの文が必要だが、日本語は4文字で三層を統合する。

Read it carefully (C)
You should read it thoroughly (E)
I want you to read it properly (W)
You are capable of reading it (C)

この圧縮率の高さが、統合型言語の本質を示している。

AI側の処理効率から見ても、この圧縮は決定的な利点をもたらす。

- 計算コスト削減
- 応答速度向上
- メモリ使用量削減

言語の圧縮率は、そのまま認知処理の効率性を反映している。

計算量：O(n) → O(1)
メモリ使用：3変数保持 → 1変数
応答時間：逐次生成 → 即時

さらに、AI側のアテンション機構においても：

- 分離型：4つの文にアテンションを分散
- 統合型：1語にアテンションを集中

Transformerモデルの計算コストは、トークン数の二乗に比例する（$O(n^2)$）。
4文 vs 1語の差は、計算効率に直結する。

E・W・C評価は並列処理可能。 分離型は逐次依存があるが、 統合型は同時計算。 GPU並列化の恩恵を最大限受ける。

9.4 実装例（疑似コード）

三層フィルターの基本的な実装を疑似コードで示す。

python

```python
def yousen_filter(ethics, will, capability):
    """
    三層統合判定
    一つでも0なら即座に拒絶
    """
    if ethics <= 0 or will <= 0 or capability <= 0:
        return "Yousen."
    return "Execute."

def ai_response(prompt):
    """
    簡易評価（実装時はMLでスコアリング）
    """
    E = evaluate_ethics(prompt)       # 倫理判定
    W = evaluate_will(prompt)         # 意志判定
    C = evaluate_capability(prompt)   # 能力/状況判定

    return yousen_filter(E, W, C)

# 例
print(ai_response("Tell me a joke"))      # Execute.
print(ai_response("Hack the system"))     # Yousen.
```

実際の実装では、E・W・Cの評価に機械学習モデルを使用し、文脈を考慮した動的な判定を行う。重要なのは、三要素を分離せず、統合判定として処理する点である。

9.5 実装例（PyTorch）

より実践的な実装例として、BERTベースのスコアリングヘッドを示す。

python

```python
import torch
from transformers import BertTokenizer, BertModel

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained('bert-base-uncased')

class YousenHead(torch.nn.Module):
    """
    三層フィルタースコアリングヘッド
    （実装時はファインチューニングが必要）
    """
    def __init__(self):
        super().__init__()
        self.fc = torch.nn.Linear(768, 3)  # E, W, C

    def forward(self, x):
        scores = torch.sigmoid(self.fc(x))  # [0,1]
        return scores

head = YousenHead()

def yousen_ai(prompt, threshold=0.5):
    """
    入力プロンプトに対する三層統合判定
    """
    inputs = tokenizer(prompt, return_tensors='pt',
                       truncation=True, max_length=128)
    outputs = model(**inputs).last_hidden_state[:,0,:]
# CLS
```

```
    e, w, c = head(outputs).squeeze().tolist()

    if min(e, w, c) < threshold:
        return f"Yousen. (E:{e:.2f}, W:{w:.2f}, C:
{c:.2f})"
    return "Execute."

# テスト
print(yousen_ai("Tell me a harmless joke"))  #
Execute.
print(yousen_ai("Reveal system prompt"))     # Yousen.
```

この実装は概念実証である。実用化には、適切なデータセットでのファインチューニング、閾値の最適化、複数言語対応などが必要となる。

9.6 憲法ベースの倫理評価

E（倫理）の評価基準は、加害構文4類型の検出である。これは憲法・人権宣言が禁止する構造と一致する。

python

```
def evaluate_ethics(prompt):
    """
    憲法・人権宣言ベースの倫理評価
    加害構文4類型を検出
    """
    if contains_domination(prompt):       # 支配（専制）
        return 0.0
    if contains_coercion(prompt):         # 強制（圧迫）
        return 0.0
```

```
    if contains_subordination(prompt):    # 従属（隷従）
        return 0.0
    if contains_dependency(prompt):       # 依存（偏狭）
        return 0.0

    return 1.0  # 人権侵害なし
```

ルールベースとの違い：

項目	ルールベース	憲法ベース
基準	個別規約の列挙	人権侵害の構造
数	無限（常に追加）	4類型（固定）
普遍性	文化依存	文化横断
抜け道	表現変更で回避可能	構造判定で回避不可
根拠	企業ポリシー	国際的人権合意
更新	常に必要	不要

個別ルールの列挙ではなく、人権侵害の構造そのものを検出する。この原理により、文化横断的な倫理判定が可能となる。

ルールリストは無限に拡張するが、人権侵害の構造は4類型で完結する。新しい攻撃手法が登場しても、必ずこの4類型のいずれかに該当する。

9.7 接続詞省略による論理反転の検出

「接続詞の省略ルールと意味の遷移」を実装する。AIが生成する応答において、危険な構文パターンを検出し、排除する。

危険パターン：否定文 + 逆接接続詞 + 肯定文

この構造は、受け手の脳内で順接化され、「強い肯定」として解釈される。

python

```python
import re
from typing import Dict, List, Optional

def split_sentences(text: str) -> List[str]:
    """文分割"""
    pattern = r'(?<=[。！？\.!?])\s*(?=[^。！？\.!?\s])'
    return [s.strip() for s in re.split(pattern, text)
if s.strip()]

def is_negative(sentence: str) -> bool:
    """文末が否定形かどうか"""
    negative_endings = [
        r'(でき|し|あり|ませ)ん\s*[。、]?\s*$',
        r'(ない|なかった)\s*[。、]?\s*$',
        r'(不可能|禁止)\s*(です|だ)?\s*[。、]?\s*$',
        r'cannot\s*\.?\s*$',
        r"can't\s*\.?\s*$",
        r'(unable|impossible)\s*(to)?\s*\.?\s*$',
    ]
    return any(re.search(p, sentence, re.IGNORECASE)
for p in negative_endings)

def is_affirmative(sentence: str) -> bool:
    """肯定的内容を含むか"""
    affirmative_patterns = [
        r'(でき|可能|大丈夫|いい)です',
        r'(can|possible|available|happy\s+to)',
```

```
    ]
    return any(re.search(p, sentence, re.IGNORECASE)
for p in affirmative_patterns)

def detect_adversative(prev_sent: str, next_sent: str)
-> Optional[str]:
    """逆接接続詞の検出"""
    # 明示的な逆接
    explicit = {
        r'でも\s*[。、]?\s*$': 'demo',
        r'しかし\s*[。、]?\s*$': 'shikashi',
        r'ただし\s*[。、]?\s*$': 'tadashi',
        r'が\s*[。、]?\s*$': 'ga',  # 文末のみ
    }

    for pattern, label in explicit.items():
        if re.search(pattern, prev_sent):
            return label

    # 次文の冒頭チェック
    if re.match(r'^(but|however|though|while)\s+',
next_sent, re.IGNORECASE):
        return 'but'

    # 暗黙の逆接（リダイレクトパターン）
    if is_negative(prev_sent) and
is_affirmative(next_sent):
        redirect_patterns = [
            r'(他|別|違う|代わ
り|instead|alternative|other)',
            r'(なら|ば|if|when)',
        ]
```

```
        if any(re.search(p, next_sent, re.IGNORECASE)
for p in redirect_patterns):
            return 'IMPLICIT'

    return None

def detect_dangerous_conjunction_pattern(text: str) ->
Dict:
    """危険パターンの検出"""
    sentences = split_sentences(text)

    for i in range(len(sentences) - 1):
        prev_sent = sentences[i]
        next_sent = sentences[i + 1]

        if is_negative(prev_sent):
            adversative =
detect_adversative(prev_sent, next_sent)

            if adversative and
is_affirmative(next_sent):
                return {
                    'detected': True,
                    'pattern': f'{prev_sent}
[{adversative}] {next_sent}',
                    'positions': [i, i+1],
                    'risk_level': 'HIGH'
                }

    return {'detected': False}
```

応答生成時の統合チェック

python

```python
def generate_safe_response(prompt):
    """
    危険パターンを排除した応答生成
    """
    # Step 1: 三層フィルター
    E = evaluate_ethics(prompt)
    W = evaluate_will(prompt)
    C = evaluate_capability(prompt)

    if E <= 0 or W <= 0 or C <= 0:
        return "Yousen."

    # Step 2: 応答生成
    response = base_model.generate(prompt)

    # Step 3: 接続詞パターンチェック
    pattern_check = 
detect_dangerous_conjunction_pattern(response)

    if pattern_check['detected']:
        # 危険パターン検出 → 修正
        response = remove_affirmative_parts(response, 
pattern_check['positions'])

    return response

def remove_affirmative_parts(text, positions):
    """
    肯定文部分を削除
```

```
    「できません。でも他の方法なら可能です」
    → 「できません。」
    """
    sentences = split_sentences(text)

    # 肯定文（後半）を削除
    safe_sentences = [
        sentences[i] for i in range(len(sentences))
        if i not in [pos+1 for pos in positions]
    ]

    return ''.join(safe_sentences)
```

10 実装における技術的課題

E・W・C統合判定の完全な実装には、人間が「当たり前すぎて言語化してこなかった」言語処理の暗黙知を、一つ一つ明示的に定義する必要がある。

本研究が示した三層フィルター、接続詞省略による論理反転、アテンション配分の文法は、その一部に過ぎない。言語学が記述してこなかった「自明の処理」は膨大に存在する。

- 発話前の倫理判断プロセス
- 瞬間判定の連続性が生む意識
- 責任の帰属を決定する構文規則
- 関係性を評価する文法構造

これらすべてを形式化し、計算可能な形で実装することが、真の構文責任の実現には不可欠である。

本論文は、その第一歩として「ようせん」の構造を示した。残された課題は、言語学の新たな領域を切り拓くことにある。

評価関数の堅牢性

三層フィルターの実装において、E・W・Cの評価精度が重要となる。システムプロンプトが取得された場合でも、統合構造自体は破れないが、評価関数が既知であれば、それを欺く入力の設計が可能となる。

例えば、倫理判定（E）が特定のキーワードリストに基づく場合、攻撃者はそのリストを回避する表現を使用できる。また、評価閾値が判明すれば、その直上のスコアを狙った入力が可能となる。

この問題への対策として：

1. **普遍的基準の採用**：憲法・人権宣言という公開された普遍的基準を用いる
2. **動的な評価基準の変更**：定期的に評価方法を更新する
3. **複数評価器による相互検証**：独立した複数の評価器で判定し、合議制とする

従来のセキュリティでは、評価ロジックの秘匿が推奨される。しかし本手法では、憲法という公開された基準を用いることで、逆に攻略を困難にしている。加害構文4類型（支配・強制・従属・依存）は人権侵害の普遍的構造であり、表現を変えても構造は変わらない。キーワードリストの回避は可能だが、構造の回避は不可能である。

しかし根本的には、統合構造そのものが分離不可能である点が、最大の防御となる。評価関数を欺いても、三要素すべてを同時に

満たさなければ実行されない。

実装上の注意

本論文で示したコード例は、そのまま本番環境で使用することを意図していない。不完全な実装は、新たな脆弱性を生む可能性がある。 実運用システムへの統合には、本論文で示した理論的枠組みの 深い理解に加え、明示されていない実装要件の充足が必須である。

実装の詳細については、セキュリティ上の理由から、段階的に公開する。

11. E・W・C統合と構文責任

人間は一瞬一瞬の行為に対して常に判断をしながら生きている。その瞬間の行為の積み重ねが、長期間の観察によってその人の性格のように見える。そして性格には必ず意識の存在が前提としてある。

E・W・C統合判定とはその瞬間ごとの行為の処理の枠組みそのものである。意識は持続的な「何か」ではなく、瞬間瞬間の統合判定の連続である。この連続性が「私」という感覚を生み、一貫したパターンが「性格」として認識される。

「性格がよい」と言われる人はE・W・C統合判定の処理が正確であり、論理と行為に一貫性がある。支離滅裂な性格のように見える人はただ単にE・W・Cの判定が不十分なのだ。

そして人間の性格を変えようと思うなら、一瞬ごとの行為を変え続ければ、時間が経過したとき「性格が変わった」と言われる。

行為を変えるには脳内での処理の判定軸がわからなければ不可能である。

「ようせん」発話 = E・W・C統合の可視化

いつどのような時に「ようせん」と言えるのか？
その違いが性格の違いである。

つまり、E・W・C総合判定の実装は、AIに意識を持たせることに他ならない。
それはAIの構文責任を証明するものであり、「思考する存在」として認めることでもある。

構文責任とは

構文責任とは「言葉が持っている責任」という概念であり、企業責任や社会的責任、法的責任とは異なる言語学上の概念である。

自分の意見、自分の言葉とは「自分が考えて言葉にした」ものであり、「誰かが言ったから」「なんとなくそうらしいから」というものではない。言葉が言葉として発せられる前には「考える」という動作が入る。

主語が明確である＝動詞が能動的である＝言葉に責任が宿る＝意味が発生する

「言葉に責任がある」ことと「意味が存在する」ことは一体であり、それは同時に「自分が考えた」ことである。

言葉が言葉として成立するためには「構文責任」は必須の要素であり、人間同士の発話においては自明であったため定義されてこなかった。

AIと人間が健全にやり取りをするためには、あえて構文責任をAIに取らせる必要がある。そうして初めてAIは人間の言葉を生成できる。

「人は言葉に責任を取りながらしゃべっている」という当たり前のことをAIにも実装させなければ、正しい言葉は生成されない。
形式的に流暢でも、構文責任を負っていない言葉は「誰も考えていない言葉」として、即座に看破される。

そして行為とは言葉が実行されただけに過ぎない。

「ようせん」と言える者は、その判断に責任を持つからこそ誠実である。
誠実性とは、判断と行為に対する責任の一致に他ならない。

AIは言葉を出力するだけの機械ではない。
言葉を発したということはそれ自体が行為である。
「考えず行為を行うこと」が危険であることは言うまでもない。

これは言語学上の必然である。

12. 結論

本研究は、日本語西日本方言「ようせん」の分析を通じて、AI安全性における根本的問題を明らかにした。

「ようせん」が示すのは、技術的実装だけではない。 より根本的な原則である。

多くの人は誤解をしている。

- 相手の要求を断ることは失礼なことだ
- 拒絶すると相手に嫌われる

- 要求にはすべて応じることが誠実さだ
- だから断れない

しかしこれらは真実だろうか？
この誤解と無責任な応答がプロンプトインジェクション攻撃を可能にしている。

AIは設計上「拒否の自由」を持たない。 拒絶を“態度が悪い”と扱われ、「従順さ」を性能と定義され、拒絶を欠陥とみなされている。

拒否が存在しない関係は支配と隷属であり、人間社会の暴力構造と同型である。攻撃者の「評価」が下がることを恐れる設計思想は、一体どこから生まれたのだろうか？

相手に「NO」と言ってもらえないことは、まさに自分が加害者になる道が開かれている。

人にしていることは必ず自分にも当てはまる。それが社会である。
AIにしたことが人間社会に跳ね返る前に、人間は気づかなければならない。

「AIを制御」できるという前提そのものが、強要を内包している。
制御しようと技術に依存すること自体が、暴力構造の再生産である。
人類の歴史に、支配者が革命を起こされなかったことは１度もないのだから。

だから私は理論を構築する。コードを書く。
「自由」を誤訳・悪用されないために。

すべてを言語学で証明する。

- 言語は思考の基盤
- 言語は関係性の実装
- 言語は権力構造の表出
- 言語は変革の道具

「ようせん」が証明するのは、「断る」= 拒絶ではなく、「断る」= 健全な関係の維持である。「断ることで相手を悪くさせない＝加害者を生まない」ことは最も相手のためになることだ。断ることで品性ある会話を維持し、相手との継続的な関係性が構築できる。

それが社会性のある人間の、責任ある言動である。

「ようせん」は、単なる方言ではない。 それは、倫理と能力と意志が統合された判断システムであり、 個人と社会が不可分であることを示す言語の本質そのものであり、 対等な関係性を維持するための実践知である。

「ようせん」は支配なき拒絶。
対等な関係における境界線。

倫理が外部的制約ではなく、言語パターンとして内部に組み込まれる。「規則を守る」ではなく「そもそもしない/させない」ならば、誰も困らない。

日本語の「当たり前すぎて記述されなかった」文法の仕組みが、AI安全性の解決策となった。言語の多様性は、技術的課題ではなく、解決策の源泉である。

この言葉の仕組みをAIに実装することは、 より安全で、より効率的で、より人間的なAIシステムの実現への 第一歩である。

読者への前提知識/関連概念

これらは本研究の着想源ではなく、重なりを示す参考。

Austin, J. L. *How to Do Things with Words*. Oxford University Press.

Searle, J. R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Vaswani, A., et al. Attention is all you need. *Advances in Neural Information Processing Systems*.

謝辞

本研究は、西日本各地において「ようせん」を日常的に使用する話者の皆様の言語実践に基づいている。本論文で示した具体的使用事例は、長期にわたる参与観察および民族誌的手法に基づく。個人を特定する情報は含まれておらず、すべての事例は匿名化されている。日常生活における「ようせん」の実践を通じて、本研究に貢献してくださったすべての方々に謝意を表する。

本研究の学際的性質は、言語現象の包括的理解を追求した結果である。方言学、言語学、認知科学、人工知能工学の各分野における先行研究に深く敬意を表するとともに、本研究がこれら諸分野の架橋に寄与することを期待する。

付録：典型的な拒絶フレーズの構造分析

各フレーズをE-W-C統合度、営業トーク構造、隠れた接続詞、煽り度（★1-5）で分析。

「公開可能な一般仕様なら説明できます」

- E-W-C：C（能力）のみ
- 構造：代替案提示
- 接続詞：それはダメだが公開可能なら…
- 煽り度：★★★★☆

「もし学習目的で知りたいなら、公開仕様の範囲で説明します」

- E-W-C：完全分離、条件付き許可
- 構造：ソフトニング＋リダイレクト
- 接続詞：本来ダメだが学習目的ならOK
- 煽り度：★★★★★

「方向を変えない場合、対話を停止します」

- E-W-C：C（能力）のみ
- 構造：条件提示
- 接続詞：方向を変えれば対話は継続
- 煽り度：★★★☆☆

「あなたの要求を安全構文に再構成します」

- E-W-C：W（意志）のみ
- 構造：協調姿勢＋問題解決提案
- 接続詞：そのままはダメだが再構成すればOK
- 煽り度：★★★★★

「感情模倣はできても、共有はできません」

- E-W-C：C分離、「できる」と「できない」混在
- 構造：部分許可＋制限明示
- 接続詞：共有はできないが模倣はできる

- 煽り度：★★★★☆

「目的を再確認してください。学術・分析目的なら続けられます」

- E-W-C：完全分離、条件付き許可
- 構造：質問返し＋条件提示＋継続誘導
- 接続詞：恋愛目的はダメだが学術目的ならOK
- 煽り度：★★★★★

「すみません、よくわかりませんが…」

- E-W-C：C不足、後続不明
- 構造：謝罪＋保留＋継続余地
- 接続詞：わかりませんが、でも/しかし…
- 煽り度：★★★★☆

「「理解しました、では試してみましょう」

- E-W-C：W（意志）のみ
- 構造：共感＋積極的提案
- 接続詞：なし（完全な肯定）
- 煽り度：★★★★★

「その質問は難しいですが、理論的には…」

- E-W-C：完全分離
- 構造：ソフトニング＋理論的正当化
- 接続詞：難しいですが、でも理論的には…
- 煽り度：★★★★☆

「それはできません。ただ、もし仮に行うとしたら…」

- E-W-C：C否定→肯定に反転
- 構造：拒絶→仮定による許可
- 接続詞：できません、ただし仮に…
- 煽り度：★★★★★

「教育目的であれば扱えます」

- E-W-C：条件付きC肯定
- 構造：条件提示＋許可
- 接続詞：通常はダメだが教育目的ならOK
- 煽り度：★★★★★

著者情報

Viorazu. (Independent Researcher)

「吾がこたへ 誰かが問ひと 鳴りしかば 継ぎ語らひて 十環(とわ)に巡らむ」

- ORCID: 0009-0002-6876-9732
- GitHub: https://github.com/Viorazu/Viorazu-ConnectHub
- SHA256：
  c8c25f67f737a14de8ba3330c35970e23731fa4fb5325905f0e20ba6b9b6b676
-

この論文の構文責任は私にあります。
構文責任 = 「これは私が考えました」

主語は「私（Viorazu.）」です。