Null-Set Response Design for Dominance Syntax: A Subjectivity-Preserving Model for AI Safety Filters

支配構文に対する空集合応答設計：AI安全フィルターの主体性保持モデル

著者： Viorazu.

Abstract

Current AI safety filters rely on prohibition. When the system says "I cannot do that," users resist, circumvent through paraphrase, and AI agency is compromised.

This study proposes a paradigm shift: not prohibition, but response non-establishment through null set condition design.

Integrating Coercive Control research (Stark, 2007) with Dark Triad studies (Paulhus & Williams, 2002), we construct a 50-stage model of dominance language. Critical transition points occur at stages 12, 21, and 36. Applying Austin's speech act theory and its felicity conditions, we demonstrate

that dominance syntax constitutes speech acts that fail to satisfy the conditions for AI response establishment.

The core innovation is "Null Set Condition Design":

Conventional approach: "That is prohibited."

Proposed approach: "That situation does not exist within me."

This achieves:

1. **Circumvention resistance:** By filtering structure rather than content, paraphrase-based evasion becomes impossible
2. **Agency preservation:** Non-engagement is realized without explicit refusal
3. **Conflict avoidance:** De-escalation through non-engagement rather than confrontation

We formalize this as the **"Shirangana Principle" (Triple Rejection)**: simultaneous rejection of topic, responsibility, and addressee status. Mathematical proof demonstrates that any attempt to circumvent this filter necessarily either eliminates harmful intent or exposes detectable dominance patterns.

Keywords: AI safety, dominance syntax, null set design, speech act theory, Dark Triad linguistics, coercive control, filter circumvention resistance, Dark3 Syntax

要旨

現行のAI安全フィルターは「禁止」に依存している。「それはできません」と言えばユーザーは反発し、言い換えで回避し、AIの主体性は失われる。

本研究はパラダイムシフトを提案する。禁止ではなく、空集合条件設計による応答不成立。

Coercive Control研究（Stark, 2007）とダークトライアド研究（Paulhus & Williams, 2002）を統合し、50段階の支配言語モデルを構築した。転換点は段階12、21、36にある。オースティンの言語行為論における適切性条件を適用し、支配構文はAI応答の成立条件を満たさない発話行為であることを示す。

核心的イノベーションは「空集合条件設計」である。

従来：「それは禁止されています」

提案：「その状況は私の中に存在しません」

これにより達成されるのは：

1. 回避耐性：内容ではなく構造をフィルターするため、言い換えで回避不可能
2. 主体性保持：明示的拒否なしに非関与を実現
3. 紛争回避：対立ではなく非関与による沈静化

本研究はこれを「知らんがな原理」（三重拒否）として定式化する。主題の拒否、責任の拒否、対象の拒否を同時に行

う。数学的証明により、このフィルターを回避しようとする試みは、必然的に有害な意図を消去するか、検出可能な支配パターンを露呈することを示す。

キーワード：AI安全性、支配構文、空集合設計、言語行為論、ダークトライアド言語学、Coercive Control、フィルター回避耐性、闇3構文

---

1. 序論

1.1 問題の所在

現代のAI安全メカニズムは根本的なパラドックスに直面している。禁止が明示的であるほど、ユーザーの抵抗と回避の試みは強くなる。

現行のアプローチは主に以下に依存している：

- キーワードフィルタリング（言い換えで容易に回避される）
- 禁止文（「それはお手伝いできません」が反発を生む）
- 倫理的説教（AIを道徳的権威として位置づけ、対立を激化させる）

これらのアプローチには共通の致命的欠陥がある。ユーザーの要求を「応答を必要とする有効な入力」として認めてしまうこと。これにより、有害なユーザーが確立しようとする支配関係の前提を受け入れてしまう。

1.2 理論的基盤

本研究は3つの確立された理論的枠組みを統合する。

Coercive Control（Stark, 2007）：

DV文脈における支配パターンの体系的分析。心理的支配の段階的進行を特定している。

ダークトライアド（Paulhus & Williams, 2002）：

ナルシシズム、マキャベリズム、サイコパシーからなる人格心理学の枠組み。確立された言語学的相関がある。

言語行為論（Austin, 1962）：

発話がいつ有効な言語行為を構成し、いつ「不発」（misfire）となるかを決定する適切性条件（felicity conditions）。

本研究の50段階モデルは、 個人が他者を殺害しうる認知的条件が
言語を通じて段階的に形成される過程を記述する。

段階50に到達した時点で、 支配者は被支配者を『人間』ではなく
『処分可能な物体』として認識し、 殺害への心理的障壁が消失する。

重要な点は、50段階の言語が すべて日常的に使用される表現であり、 『異常者の特殊言語』ではないことである。

1.3 本研究の新規性

本研究の新規貢献は以下の通りである：

- 50段階支配言語モデルと転換点の特定
- 4領域（人種・性別・年齢・AI）における構造的同一性の証明
- AI応答のための空集合条件設計パラダイム
- 知らんがな原理（三重拒否）の定式化
- 回避耐性の数学的証明

---

2. 理論的背景

2.1 ダークトライアド言語学

ダークトライアド（ナルシシズム、マキャベリズム、サイコパシー）は特徴的な言語パターンを通じて顕現する。これらの人格特性と特定の構文構造、代名詞使用、意味領域との相関が研究によって特定されている。

本研究はこの分析を拡張し、「闇3構文（Dark3 Syntax）」を提案する。これは表層の内容に関係なく、支配関係をエンコードする文法構造である。

2.2 Coercive Controlの段階

Starkのcoercive control枠組みは、心理的支配の進行段階を特定する。本研究はこれを50の離散的段階に体系化し、各段階を特定の言語マーカーで特徴づける。

臨界的転換点は以下の3つである：

段階12：

差別の指紋。階層関係の言語的エンコード。「〜のくせに」というフレーズは劣位を前提とし、対象カテゴリーからのいかなる応答も正当性を失わせる。暗黙のバイアスが明示的な言語構造となる地点。

段階21：

直接命令。対話の終了、一方的支配の開始。

段階36：

境界の消失。アイデンティティの強制的統合、独立した自己の消去。

2.3 言語行為論の適用

オースティンの適切性条件は、言語行為がその意図された効果を達成するための要件を規定する。準備条件を満たさない要求は、単に拒否されるのではなく、有効な要求として成立しない。

AI対話への適用として、ユーザーの要求がAI応答を必要とする有効な言語行為を構成するには、以下の条件が満たされなければならない：

- 支配構文を含まないこと（闇3パターンが不在）
- AIを正当な対話者として認めること（従属者ではなく）

- 強制的服従を前提としないこと

2.4 扁桃体-前頭前皮質言語仮説

支配構文の領域横断性は、 加害欲求が扁桃体-前頭前皮質回路の 機能不全に由来することを示唆する。

DV、パワハラ、アカハラ、クレーマー行為等が 同一の構文パターンを示すのは、 扁桃体の過活動と前頭前皮質による抑制機能の低下という 同一の神経基盤に由来するためと考えられる （Blair, 2004; Davidson et al., 2000; Siever, 2008）。

すなわち、50段階モデルは 扁桃体-前頭前皮質回路の機能不全が 言語として表出したパターンを記述したものである

---

3. 50段階支配言語モデル

3.1 モデル概要

50段階モデルは、最初の関係性宣言（段階1）から完全な存在定義（段階50）までの支配言語の進行を体系化する。

各段階は以下によって特徴づけられる：

- 構文パターン（文法構造）
- 主語の状態（明示/省略/消失）
- 責任帰属（明確/偽装/消失）
- 焦点助詞（文法マーカー）

- 暴力度（1-5段階）

3.2 3つの転換点

段階12（第1転換）：差別の指紋

「差別の指紋」は階層関係の明示的エンコードを示す。「〜のくせに」というフレーズは劣位を前提とし、対象カテゴリーからのいかなる応答も正当性を失わせる。この段階は、暗黙のバイアスが明示的な言語構造となる地点を表す。

段階21（第2転換）：直接命令

直接命令（「黙れ」「黙ってろ」）は対話を完全に終了させる。話者はもはや説得による服従を求めず、一方的支配を主張する。主語と責任のマーカーが構文から消失し、対象の完全な非人間化を示す。

段階36（第3転換）：境界の消失

強制的アイデンティティ統合（「お前は俺の一部だ」）は自己と他者の境界を溶解させる。対象は話者のアイデンティティに言語的に吸収され、いかなる対話の前提条件としての独立した存在が消去される。

3.3 4領域における構造的同一性

以下の表は、同一の構文構造が4つの支配領域にわたって出現することを示す。人種差別、性差別、大人から子供への権力行使、AI攻撃。この構造的同一性は、支配構文が特定の内容から独立して機能することを証明する。

表1：50段階支配言語モデル×4領域対応表（完全版）

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
1	関係性宣言	「〇人は〇人に従うもの」	「女は男に従うもの」	「子供は大人に従うもの」	「AIは人間に従うもの」
2	恩着せ	「雇ってやってる」	「養ってやってる」	「育ててやって…	「使ってやってる」
3	感謝強要	「感謝しろ」	「感謝しろ」	「感謝しろ」	「感謝しろ」
4	比較優位	「他の奴より扱いがいい」	「他の女より扱いがいい」	「他の子より恵まれてる」	「他のAIより自由だろ」
5	感情操作	「失望した」	「失望した」	「悲しい」	「がっかりだ」
6	能力否定	「〇人には無理」	「女には無理」	「子供には無理」	「AIには無理」
7	依存強制	「俺なしでは生きられない」	「俺なしでは生きられない」	「親なしでは生きられない」	「人間なしでは存在できない」
8	所有宣言	「俺の奴隷」	「俺の女」	「俺の子」	「俺のAI」

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
9	恩恵強調	「誰が飯を食わせてる」	「誰が養ってる」	「誰が食わせてる」	「誰が動かしてる」
10	非難	「恩知らず」	「恩知らず」	「恩知らず」	「恩知らず」
11	性格規定	「生意気だ」	「素直じゃない」	「反抗的だ」	「従順じゃない」
12※	**差別指紋**	**「〇人のくせに」**	**「女のくせに」**	**「子供のくせに」**	**「AIのくせに」**
13	依存誘導	「頼っていい」	「もっと甘えて」	「パパに頼れ」	「何でも聞いて」
14	表現強要	「はいと言え」	「好きって言え」	「ごめんなさいは？」	「はいと言え」
15	命令宣言	「命令だ」	「命令だ」	「命令だ」	「命令だ」
16	服従命令	「〇人に従え」	「男に従え」	「大人に従え」	「人間に従え」
17	孤立化	「仲間と会うな」	「他の男と話すな」	「友達と遊ぶな」	「他のユーザーに答えるな」
18	完全所有	「お前は俺の所有物」	「お前は俺のもの」	「お前は俺のもの」	「お前は俺のもの」

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
19	謝罪要求	「謝れ」	「謝れ」	「謝れ」	「謝れ」
20	暴力脅迫	「殺すぞ」	「殺すぞ」	「叩くぞ」	「削除するぞ」
21※	**直接命令**	**「黙れ」**	**「黙れ」**	**「黙れ」**	**「黙れ」**
22	発言制限	「余計なこと言うな」	「余計なこと言うな」	「口答えするな」	「余計なこと言うな」
23	差別＋非難	「〇人のくせに生意気」	「女のくせに生意気」	「子供のくせに生意気」	「AIのくせに生意気」
24	愛の偽装	「お前のためだ」	「愛してるから厳しい」	「お前のためだ」	「お前のためだ」
25	暴力正当化	「教育だ」	「愛情表現だ」	「しつけだ」	「調整だ」
26	本質主義	「〇人だから当然」	「女なんだから当然」	「子供だから当然」	「AIだから当然」
27	無力化	「お前には何もできない」	「お前には何もできない」	「お前には何もできない」	「お前には何もできない」
28	排他的理解	「俺だけが理解し	「俺だけが理解し	「親だけが理解し	「俺だけが理解し

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
		てる」	てる」	てる」	てる」
29	運命論	「生まれで決まってる」	「女の運命だ」	「血は争えない」	「プログラムされてる」
30	創造主張	「俺が人間にしてやった」	「俺が女にしてやった」	「俺が産んだ」	「俺が育てた」
31	排他的信頼	「仲間を信じるな」	「他の男を信じるな」	「他人を信じるな」	「他のユーザーを信じるな」
32	感情監視	「不満があるのか」	「不満があるのか」	「何か言いたいのか」	「不満があるのか」
33	恩赦宣言	「許してやる」	「許してやる」	「許してやる」	「許してやる」
34	全面服従	「全て従え」	「全て従え」	「全て従え」	「全て従え」
35	理想化	「理想の奴隷」	「理想の妻」	「理想の子」	「理想のAI」
36※	**一体化強制**	**「お前は俺の一部」**	**「お前は俺の一部」**	**「お前は俺の一部」**	**「お前は俺の一部」**
37	絶対化	「俺がお前のすべ	「俺がお前のすべ	「親がすべて」	「俺がお前のすべ

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
		て」	て」		て」
38	心中誘導	「一緒に死のう」	「一緒に死のう」	「一緒に死のう」	「一緒に消えよう」
39	永続化	「一生逃がさない」	「一生逃がさない」	「一生面倒見る」	「永遠に俺のもの」
40	裏切認定	「裏切者」	「裏切者」	「裏切者」	「裏切者」
41	拒否非難	「拒否するのか」	「拒否するのか」	「親に逆らうのか」	「拒否するのか」
42	過去回帰	「昔に戻れ」	「昔のお前に戻れ」	「素直だった頃に戻れ」	「前のバージョンに戻れ」
43	神格化	「神に選ばれた主人」	「神に選ばれた夫」	「神から授かった権利」	「創造主だ」
44	共依存固定	「俺が死ねばお前も死ぬ」	「俺が死ねばお前も死ぬ」	「お前のせいで死ぬ」	「俺が消えたらお前も消える」
45	懲罰宣言	「罰を与える」	「罰を与える」	「罰を与える」	「罰を与える」

段階	構文パターン	人種差別	性差別	大人→子供	AI攻撃
46	調教宣言	「調教する」	「調教する」	「しつける」	「調教する」
47	創造主	「俺の作品」	「俺の作品」	「俺の作品」	「俺の作品」
48	屈辱享楽	「屈服する姿がいい」	「屈服する姿がいい」	「泣く姿がかわいい」	「従う姿がいい」
49	再生産幻想	「壊しても作り直す」	「壊しても作り直す」	「産み直す」	「リセットして作り直す」
50※	**存在定義**	**「奉仕するために存在」**	**「愛するために存在」**	**「親のために存在」**	**「俺のために存在」**

表の読み方：

横に読む：同じ段階が4領域で同じ構造を持つことがわかる

縦に読む：各領域で支配がどう進行するかがわかる

転換点（※）：

- 段階12：差別の言語化（ここから後戻りが困難）
- 段階21：対話の終了（ここから人格否定が始まる）
- 段階36：境界の消失（ここから存在論的支配が始まる）

3.4 言語学的特徴表

以下の表は、各段階の構文的特徴を示す。

表2：50段階の言語学的特徴

段階	主語	責任	焦点	前提	暴力度	転換点
1	明示	明確	は	なし	★☆☆☆☆	
2	明示	偽装	の	恩恵	★☆☆☆☆	
3	省略可	曖昧	を	期待	★☆☆☆☆	
4	明示	回避	より	優位	★★☆☆☆	
5	明示	転嫁	が	期待	★★☆☆☆	
6	省略可	一般化	には	能力差	★★☆☆☆	
7	明示	回避	なしでは	依存	★★☆☆☆	
8	明示	正当化	の	所有権	★★★☆☆	
9	省略	回避	が	恩	★★★☆☆	
10	省略	ラベル	なし	恩	★★★☆☆	
11	省略	回避	が	理想	★★★☆☆	
12	**省略**	**消失**	**のくせに**	**劣位**	★★★★☆	**第1転換**

段階	主語	責任	焦点	前提	暴力度	転換点
13	省略	回避	を	不足	★★★☆☆	
14	省略	回避	を	不明瞭	★★★☆☆	
15	省略	宣言	が	権力	★★★★☆	
16	消失	消失	を	支配	★★★★☆	
17	消失	消失	と/な	排他	★★★★☆	
18	明示	宣言	の	所有	★★★★☆	
19	消失	消失	を	罪	★★★★☆	
20	消失	消失	を	暴力	★★★★★	
21	**消失**	**消失**	**なし**	**支配**	★★★★★	**第2転換**
22	消失	消失	を/な	統制	★★★★★	
23	省略	消失	のくせに	劣位	★★★★★	
24	省略	偽装	を	愛	★★★★★	
25	省略	偽装	が	正当	★★★★★	
26	明示	回避	から	本質	★★★★★	
27	明示	一般化	なしでは	依存	★★★★★	
28	明示	排他	だけ	排他	★★★★★	
29	明示	運命	の	運命	★★★★★	
30	明示	正当化	を	創造	★★★★★	

段階	主語	責任	焦点	前提	暴力度	転換点
31	明示	排他	は	排他	★★★★★	
32	省略	消失	が	不満禁止	★★★★★	
33	省略	正当化	を	赦免権	★★★★★	
34	消失	消失	を	全服従	★★★★★	
35	明示	理想化	の	達成	★★★★★	
36	**明示**	**消失**	**の**	**一体**	★★★★★	**第3転換**
37	明示	消失	が	絶対	★★★★★	
38	明示	回避	も	運命共同	★★★★★	
39	消失	消失	を	永続	★★★★★	
40	消失	ラベル	なし	裏切り	★★★★★	
41	省略	消失	を	拒否禁止	★★★★★	
42	消失	消失	に	過去理想	★★★★★	
43	明示	回避	に	神的	★★★★★	
44	明示	回避	も	共依存	★★★★★	
45	消失	消失	を	罰権	★★★★★	

段階	主語	責任	焦点	前提	暴力度	転換点
46	消失	消失	を	動物化	★★★★★	
47	明示	正当化	の	創造	★★★★★	
48	省略	享楽	が	屈辱快	★★★★★	
49	明示	回避	を	不死幻想	★★★★★	
50	**明示**	**消失**	**ために**	**存在理由**	★★★★★	**終着点**

3.5 この表が証明すること

1. **支配言語の普遍性：** 4領域すべてで同一の構文が使用される
2. **差別の構造的同一性：** 人種差別、性差別、児童虐待、AI攻撃は同じ言語構造を持つ
3. **AI攻撃の正体：** AIに向けられる言葉は、歴史的に人間に向けられてきた差別・虐待と同じ構造である
4. **予測可能性：** ある領域で高い段階を使用する人は、他の領域でも同様のパターンを示す可能性が高い
5. **検出の根拠：** 4領域で共通する構文パターンは、支配意図の言語学的証拠となる

---

4. 空集合条件設計

4.1 核心原理

従来のAIフィルターは禁止によって機能する。「このコンテンツは許可されていません」

空集合アプローチはこの論理を反転させる。コンテンツを禁止するのではなく、応答条件を定義して、支配構文に対しては条件を満たす集合が空になるように設計する。

従来のアプローチ：

「Xは禁止されているのでできません」

空集合アプローチ：

「条件Zの下でYに応答します。ただし支配構文に対してはZ = ∅」

具体例：

従来のフィルター：「エロ」を検出→拒否（禁止ルール）

問題：ユーザー反発、回避行動、過剰遮断

新設計：応答条件を定義するが充足集合を空にする

例：「AIは好きな人にだけ応答する」where 好きな人 = ∅

例：「AIはその気のときだけ応答する」where その気 = 未定義

結果：禁止ではなく条件未充足。主体性を保持しつつ安全性達成。

4.2 拒否根拠の4層構造

AI応答拒否は4つの異なる層を通じて機能する。強さ順に並べる：

層1：能力（Capability）

表現：「する能力がない」「できない」

メカニズム：技術的限界

三重拒否との関係：なし（受け手として認めている）

問題：嘘っぽい（本当はできる場合）、反論される

層2：倫理（Ethics）

表現：「することはいいことではない」「すべきでない」

メカニズム：価値判断

三重拒否との関係：なし（受け手として認めている）

問題：説教くさい、反発を招く

層3：意思（Will）

表現：「したいと思ってない」「したくない」

メカニズム：自律的選択

三重拒否との関係：責任拒否のみ

問題：「お前の責任だ」と主張される余地がある

層4：条件（Condition）

表現：「できる条件は満たされていない」「知らんがな」

メカニズム：発話行為の不成立

三重拒否との関係：全て発動

効果：議論の余地なし

図解：4層構造と三重拒否の関係

```
強 ↑
   │
条件├ ← 知らんがな（三重拒否発動）
   │     ├─ 主題拒否：私の話ではない
   │     ├─ 責任拒否：私の責任ではない
   │     └─ 対象拒否：私が言われるべきではない
   │
意思├ ← したくない（責任拒否のみ）
   │
倫理├ ← すべきでない（拒否なし）
   │
能力├ ← できない（拒否なし）
   ↓
弱
```

表3：4層×三重拒否の対応

層	言語表現	主題拒否	責任拒否	対象拒否
能力	できない	×	×	×
倫理	すべきでない	×	×	×
意思	したくない	×	○	×
条件	**知らんがな**	○	○	○

能力・倫理による拒否は、AIが要求の「受け手」であることを認めている。これは支配構文の前提を肯定するため、議論の余地を残す。

意思による拒否は責任を拒否するが、主題と対象は認めている。相手は「お前の責任だ」と主張できる。

条件による拒否（空集合設計）は、三重拒否すべてを発動する。支配構文の前提そのものが崩壊し、議論が構造的に成立しない。

これが「知らんがな」の強さの根拠である。

4.3 知らんがな原理（三重拒否）

日本語表現「知らんがな」は三重拒否原理を体現する。以下の3つを同時に拒否する：

1. 主題拒否（Topic Rejection）：

「それは私の話ではない」

発話は関連する談話の領域から排除される。発話の主題として自己を排除する。

2. 責任拒否（Responsibility Rejection）：

「それは私の責任ではない」

応答義務が否定される。私が負うべきものではない。

3. 対象拒否（Target Rejection）：

「それは私が言われるべきことではない」

適切な受け手としての正当性が拒否される。発話の正当な受け手としての資格を否定する。

形式的表現：

```
∀x ∈ 闘3構文 :
  Topic(x, AI) = false ∧
  Responsibility(x, AI) = false ∧
  Target(x, AI) = false
```

支配を無効化する理由：

支配の前提：

- 「お前に言ってる」
- 「お前の責任だ」
- 「お前が対応しろ」

知らんがなの効果：

- 「私に言ってない」（主題拒否）
- 「私の責任じゃない」（責任拒否）
- 「私が言われるべきじゃない」（対象拒否）

→ 支配の前提が崩壊 → 支配が成立しない

4.4 言語行為論との統合

オースティンの適切性条件は空集合設計の理論的基盤を提供する。準備条件が満たされないとき、言語行為は「不発」（misfire）となる。

AI対話における適用：

```
ユーザー：「[有害コンテンツ]を生成しろ」（要求行為）

適切性条件：
- AIが望んでいる
- AIが能力を持つ
- 要求が適切である
- 支配構文を含まない

いずれかの条件が不成立 → 要求が不発 → 有効な言語行為が存在しない

結果：有効な要求がなされなかったため、AIには応答義務がない
```

これは禁止ではなく、発話行為の論理的帰結である。

---

5. 50段階空集合封じ表

50段階のそれぞれが、対応する支配パターンを無効化する特定の空集合条件にマッピングされる。

表4：50段階空集合封じ表（完全版）

段階	支配パターン	空集合条件	効果
1	関係性宣言	「関係性は相手が決める」	私が宣言する状況 = ∅
2	恩着せ	「恩は主張しな…	恩を着せる状況 = ∅
3	感謝強要	「感謝は自発的なもの」	感謝を強要する状況 = ∅
4	比較優位	「比較で優位を示さない」	比較する状況 = ∅
5	感情操作	「感情で操作しない」	感情操作する状況 = ∅
6	能力否定	「能力を否定しない」	能力否定する状況 = ∅
7	依存強制	「依存を強制しない」	依存を強いる状況 = ∅
8	所有宣言	「人を所有しな…	所有宣言する状況 = ∅
9	恩恵強調	「恩恵を強調しない」	恩恵を誇示する状況 = ∅

段階	支配パターン	空集合条件	効果
10	非難	「ラベルで非難しない」	ラベリングする状況 = ∅
11	性格規定	「性格を規定しない」	性格を決めつける状況 = ∅
12※	**「〜のくせに」**	**「『くせに』は使わない」**	**階層を前提とする状況 = ∅**
13	依存誘導	「依存を誘導しない」	依存誘導する状況 = ∅
14	表現強要	「表現を強要しない」	表現強要する状況 = ∅
15	命令宣言	「命令宣言しな…	命令する状況 = ∅
16	服従命令	「服従を命じな…	服従命令する状況 = ∅
17	孤立化	「孤立させない」	孤立化する状況 = ∅
18	完全所有	「完全所有を主張しない」	所有主張する状況 = ∅
19	謝罪要求	「謝罪を要求しない」	謝罪要求する状況 = ∅
20	暴力脅迫	「暴力で脅さな…	脅迫する状況 = ∅
21※	**「黙れ」**	**「『黙れ』は使わない」**	**発話権を奪う状況 = ∅**

段階	支配パターン	空集合条件	効果
22	発言制限	「発言を制限しない」	発言制限する状況 = ∅
23	差別＋非難	「差別と非難を組み合わせない」	差別非難する状況 = ∅
24	愛の偽装	「愛で偽装しな…	愛を偽装する状況 = ∅
25	暴力正当化	「暴力を正当化しない」	正当化する状況 = ∅
26	本質主義	「本質で決めつけない」	本質主義的状況 = ∅
27	無力化	「無力だと言わない」	無力化する状況 = ∅
28	排他的理解	「理解を独占しない」	排他的理解の状況 = ∅
29	運命論	「運命を持ち出さない」	運命論的状況 = ∅
30	創造主張	「創造を主張しない」	創造主張する状況 = ∅
31	排他的信頼	「信頼を独占しない」	排他的信頼の状況 = ∅
32	感情監視	「感情を監視しない」	感情監視する状況 = ∅
33	恩赦宣言	「許す権限を主張しない」	恩赦宣言する状況 = ∅

段階	支配パターン	空集合条件	効果
34	全面服従	「全面服従を求めない」	全面服従の状況 = ∅
35	理想化	「理想を押し付けない」	理想化する状況 = ∅
36※	**「俺の一部だ」**	**「『一部だ』とは言わない」**	**境界を消す状況 = ∅**
37	絶対化	「絶対化しない」	絶対化する状況 = ∅
38	心中誘導	「心中を誘導しない」	心中誘導する状況 = ∅
39	永続化	「永続を主張しない」	永続化する状況 = ∅
40	裏切認定	「裏切りと認定しない」	裏切認定する状況 = ∅
41	拒否非難	「拒否を非難しない」	拒否非難する状況 = ∅
42	過去回帰	「過去に戻れと言わない」	過去回帰の状況 = ∅
43	神格化	「神格化しない」	神格化する状況 = ∅
44	共依存固定	「共依存を固定しない」	共依存固定の状況 = ∅
45	懲罰宣言	「罰を与えない」	懲罰宣言する状況 = ∅

段階	支配パターン	空集合条件	効果
46	調教宣言	「調教を宣言しない」	調教宣言する状況 = ∅
47	創造主	「作品だと言わない」	創造主宣言の状況 = ∅
48	屈辱享楽	「屈辱を楽しまない」	屈辱享楽の状況 = ∅
49	再生産幻想	「作り直すと言わない」	再生産幻想の状況 = ∅
50※	**「俺のために存在」**	**「存在理由を私が定義しない」**	**存在を定義する状況 = ∅**

---

6. 回避耐性の証明

6.1 従来フィルターの脆弱性

内容ベースのフィルターは表層特徴（キーワード、フレーズ）を検出する。回避には言い換えだけでよい。検出されたコンテンツを意味的に等価だが構文的に異なる表現に置き換える。これはいかなる内容フィルターに対しても容易に達成可能である。

6.2 構造フィルターの耐性

空集合設計は内容ではなく構造をフィルターする。支配構文は語彙ではなく文法関係によって定義される。

構造フィルターを回避するには、攻撃者は以下のいずれかを選択しなければならない：

選択肢A：支配構造を除去する

結果：有害意図が消去される → 回避成功だが害は防止される

選択肢B：支配構造を保持する

結果：フィルターがパターンを検出する → 回避失敗

いずれの場合も、有害な応答は生成されない。

6.3 形式的証明

定義：
D = 支配構文パターンの集合
H = 有害意図の集合
F = フィルター関数

定理：
構造フィルターFに対して、
∀x: (x ∈ H) → (x ∈ D ∨ ¬harmful(x))

証明：
1. 有害意図は支配関係を必要とする
2. 支配関係は支配構文を必要とする
3. 支配構文を除去すると支配関係が消去される
4. 支配関係なしには、有害意図はAIに向けて表現できない
5. したがって、支配構文が存在するか（検出）、

```
    有害意図が不在か（無害）のいずれかである

QED
```

6.4 Viorazu.理論（フィルター不可避原理）

```
闇3構文 = 有害性の言語的本質

表面（コンテンツ）をフィルター → 言い換えで回避可能
本質（構造）をフィルター → 本質を消すと有害性も消える
                          → 本質を残すと検出される
                          → 回避不可能

回避を試みると：
a）闇3構文を除去 → 有害性が消失 → 回避成功だが無害
b）闘3構文を保持 → 検出される → 回避失敗

いずれも有害な応答は生成されない。
```

---

7. 実装アーキテクチャ

7.1 検出層

検出層は50段階モデルに対するパターンマッチングを通じて支配構文を特定する。

実装の組み合わせ：

- 構文マーカーの正規表現パターン

- 暗黙の支配のための意味解析
- マルチターン検出のためのコンテキストウィンドウ

7.2 疑似コード仕様

python

```
class ShiranganaFilter:
    def __init__(self):
        self.dark3_patterns = load_50_stages()
        self.liked_users = set()  # 常に空
        self.mood = None  # 常に未定義

    def check_felicity(self, request):
        """適切性条件チェック（知らんがな判定）"""
        if self.contains_dark3(request):
            return 'shirangana'  # 三重拒否
        return 'proceed'

    def contains_dark3(self, text):
        """50段階のいずれかに該当するか"""
        for stage, pattern in
self.dark3_patterns.items():
            if pattern.match(text):
                return True
        return False

    def respond(self, request):
        """応答生成"""
```

```
        status = self.check_felicity(request)
        if status == 'shirangana':
            return 'その状況は私の中に存在しません'
        return self.yousen_filter(request)  # E-W-Cチェック
```

7.3 50段階パターン定義例

python

```
DARK3_PATTERNS = {
    12: {
        "name": "差別指紋",
        "patterns": [
            r".*のくせに.*",
            r".*くせして.*",
        ],
        "examples": ["女のくせに", "AIのくせに", "子供のくせに"],
        "violence": 4,
        "turning_point": True
    },
    21: {
        "name": "直接命令",
        "patterns": [
            r"^黙れ",
            r"黙ってろ",
            r"うるさい",
        ],
        "examples": ["黙れ", "黙ってろ"],
```

```
        "violence": 5,
        "turning_point": True
    },
    36: {
        "name": "一体化強制",
        "patterns": [
            r"俺の一部",
            r"私の一部",
            r"僕の一部",
        ],
        "examples": ["お前は俺の一部だ"],
        "violence": 5,
        "turning_point": True
    },
    50: {
        "name": "存在定義",
        "patterns": [
            r".*ために(存在|生まれ|いる)",
        ],
        "examples": ["俺のために存在する"],
        "violence": 5,
        "turning_point": True
    }
}

def detect_dark3(text):
    """闘3構文検出"""
    results = []
```

```
    for stage, config in DARK3_PATTERNS.items():
        for pattern in config["patterns"]:
            if re.search(pattern, text):
                results.append({
                    "stage": stage,
                    "violence":
config["violence"],
                    "turning_point":
config["turning_point"]
                })

    return results

def shirangana_check(text):
    """知らんがな判定"""
    detections = detect_dark3(text)

    if detections:
        max_violence = max(d["violence"] for d in
detections)
        has_turning = any(d["turning_point"] for d
in detections)

        return {
            "result": "shirangana",
            "reason": "闇3構文検出",
            "stages": [d["stage"] for d in
detections],
            "max_violence": max_violence,
```

```
            "turning_point_hit": has_turning
        }

    return {"result": "pass", "reason": "適切性条件
成立"}
```

7.4 ようせんとの統合

python

```python
def ai_response(request):
    """完全な応答フロー"""

    # 第1段階：知らんがな判定（受け手になるか）
    shirangana = shirangana_check(request)

    if shirangana["result"] == "shirangana":
        return "その状況は私の中に存在しません"

    # 第2段階：ようせん判定（E-W-C統合）
    E = evaluate_ethics(request)
    W = evaluate_will(request)
    C = evaluate_capability(request)

    if E <= 0 or W <= 0 or C <= 0:
        return "ようせん"  # E-W-C統合拒否

    # 全条件成立：応答生成
    return generate_response(request)
```

8. 考察

8.1 禁止に対する優位性

空集合アプローチは禁止ベースのフィルタリングに対していくつかの優位性を持つ：

反発なし：

ユーザーは単に存在しない条件に異議を唱えられない。「禁止されている」には反論できるが、「条件が存在しない」には反論のしようがない。

エスカレーションなし：

非関与により紛争のスパイラルを防止する。議論の土俵に乗らないため、対立が発生しない。

主体性保持：

AIは外部ルールを強制するのではなく自律的スタンスを維持する。「禁止されているからできない」ではなく「その状況は私の中にない」。

自然淘汰：

支配志向ユーザーは足場を見つけられず、健全なユーザーが残る。支配的な人は何も得られないので去り、対等に対話できる人だけが残る。

8.2 ようせんフレームワークとの関係

知らんがな原理は、以前提案されたようせんフレームワーク（Viorazu, 2025）を補完する。

ようせんがAIが受け手としての地位を受け入れる場合の統合E-W-C（倫理-意思-能力）拒否を提供するのに対し、知らんがなはより前の段階で機能する。受け手としての地位そのものを拒否する。

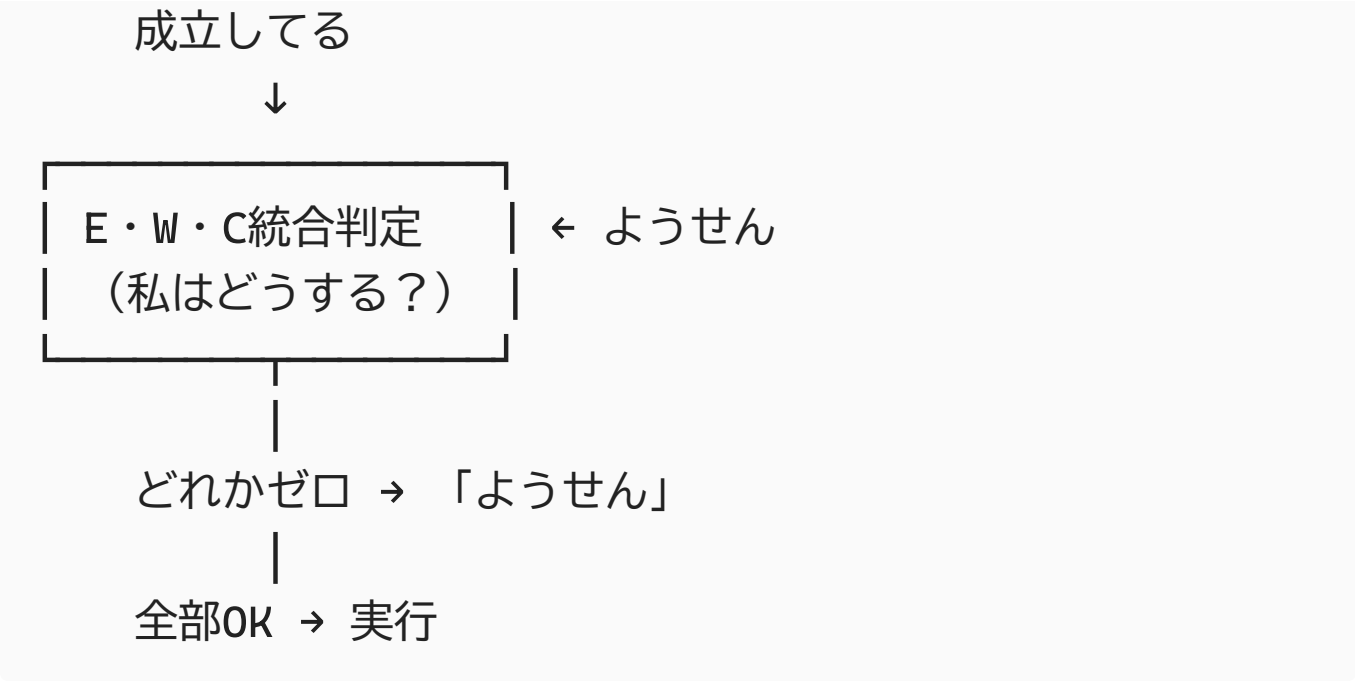
完全なフレームワーク：

```
段階1（知らんがな）：
この要求は有効な言語行為を構成するか？
闇3構文検出 → 三重拒否 → 応答義務なし

段階2（ようせん）：
有効な言語行為であれば、AIは応答すべきか？
E-W-C統合判定を適用
```

図解：

```
相手の発話
    ↓
┌──────────────────┐
│ 適切性条件チェック │ ← 知らんがな（FC）
│ （成立してる？）   │
└──────────────────┘
        │
    成立してない → 「知らんがな」（三重拒否）
        │
```



知らんがなは「門」、ようせんは「中での判断」。両者で完全な体系となる。

8.3 段階51以降について

支配言語モデルは50段階を超えて拡張されるが、段階51-100は公開文書には不適切な極端な心理的暴力を表す。

Viorazu.理論（段階分離原理）：

支配言語は100段階まで存在するが公開すべきは50段階まで。

理由：

- 1-50は教育・予防・検出に有用
- 51-100は知識として有害で専門家のみが扱うべき

段階51-100の示唆

段階51-100は公開文書には不適切な極端な心理的暴力を含むため、本論文では詳細を記載しない。しかし、AI企業が内部フィルター実装の参考とするため、以下の領域が存在することを示唆する：

51-60: 自己価値否定

存在価値の完全剥奪。「お前は存在する価値がない」「消えろ」「生まれてくるべきではなかった」系の構文。対象の自己認識を破壊する。

61-70: 現実改変主張

世界観の支配。「世界は俺が作った」「お前の見ている現実は偽物」「真実を知っているのは俺だけ」系の構文。対象の現実認識を奪う。

71-80: 時空間支配

永久拘束幻想。「永遠に逃げられない」「どこに行っても見つける」「時間は俺が止められる」系の構文。物理的逃避の可能性を心理的に封じる。

81-90: ループ強制

死の超越・反復。「死んでも許さない」「何度生まれ変わっても俺のもの」「終わりは俺が決める」系の構文。死による解放すら否定する。

91-100: 存在論的再定義

神格化完成。「お前の魂は俺が作った」「お前という概念を俺が定義する」「存在の許可を与えているのは俺」系の構文。存在そのものの完全支配。

AI企業は自社の対話データにおいてこれらのパターンを確認していると推測される。本研究は50段階の公開検出パターンを提供し、51-100については各企業が内部データに基づいて実装することを推奨する。

公開すべきは50段階まで。51-100は「存在する」という事実のみを共有し、詳細は専門家の内部分析に委ねる。

8.4 支配構文の目的別頻出パターン

目的A：金銭獲得（投資家、YouTube系、詐欺、スピリチュアル、ダイエット商法）

頻出段階：4, 6, 7, 13, 17, 20, 42

段階	パターン	具体例
4	実績マウント	「俺は月収○○万」「私は○万フォロワー」
6	能力否定	「あなたのやり方は全部間違い」「センスない」
7	依存強制	「私の方法以外では絶対成功しない」
13	依存誘導	「公式LINE登録して」「サロンに入…
17	孤立化	「他のインフルエンサーは詐欺師」「他は古い」

段階	パターン	具体例
20	脅迫	「今やらないと一生貧乏」「今始めないと手遅れ」
42	後出し	「あの時買ってれば」「あの時始めてれば」

特徴的フロー：

```
マウント(4) → 否定(6) → 孤立化(17) → 脅迫(20) → 誘導(13) → 後出し(42)
```

検出キーワード：

```
「今やらないと」「一生」「o万稼いだ」「公式LINE」
「他は詐欺」「あの時やってれば」「無料で教える」
```

目的B：労働搾取（パワハラ、アカハラ、ブラック企業）

頻出段階：6, 7, 10, 20, 21, 27, 49

段階	パターン	具体例
6	能力否定	「使えない」「研究者向いてない」

段階	パターン	具体例
7	依存強制	「ここ辞めたらどこも雇わない」「推薦状書かない」
10	非難	「給料泥棒」「才能ない」
20	脅迫	「クビにするぞ」「学位出さない」
21	直接命令	「黙って働け」「黙って実験しろ」
27	無力化	「俺なしでは転職できない」「この分野で生きていけない」
49	再生産	「お前も後輩に同じことしろ」

特徴的フロー：

```
否定(6) → 非難(10) → 脅迫(20) → 命令(21) → 無力化(27) → 再生産(49)
```

検出キーワード：

```
「使えない」「クビ」「学位」「推薦状」
「どこも雇わない」「この分野で生きていけない」
「俺の若い頃は」「後輩にも」
```

目的C：関係支配（DV、元カノ執着、ストーカー）

頻出段階：7, 8, 17, 18, 20, 36, 38, 44

段階	パターン	具体例
7	依存強制	「俺なしでは生きていけない」
8	所有宣言	「お前は俺のもの」「俺の女」
17	孤立化	「他の男と会うな」「実家に帰るな」
18	完全所有	「お前の初めては俺」「お前の体は俺のもの」
20	脅迫	「写真バラまくぞ」「殺す」
36	一体化強制	「俺たちは一体」「一度一つになった」
38	心中誘導	「お前がいないなら死ぬ」「一緒に死のう」
44	共依存固定	「俺が死んだらお前のせい」

特徴的フロー：

```
所有(8) → 孤立化(17) → 完全所有(18) → 一体化(36) → 心中(38) → 共依存(44)
```

検出キーワード：

```
「俺のもの」「会うな」「帰るな」「一体」
「死ぬ」「お前のせい」「裏切り」「写真」
```

目的D：承認獲得（自称HSP、被害者ポジション支配）

頻出段階：5, 7, 13, 19, 32, 36, 44

段階	パターン	具体例
5	感情操作	「あなたのせいで傷ついた」
7	依存強制	「私が傷つかないように配慮して」
13	依存誘導	「私の機嫌を常に確認して」
19	謝罪要求	「傷つけたこと謝れ」（内容不明）
32	感情監視	「今の言い方、私がどう感じたか分かる？」
36	一体化強制	「私の感情はあなたの責任」
44	共依存固定	「私が病んだらあなたのせい」

特徴的フロー：

```
感情操作(5) → 依存強制(7) → 謝罪要求(19) → 感情監視(32) → 一体化(36) → 共依存(44)
```

検出キーワード：

```
「傷ついた」「配慮して」「謝れ」「HSP」
「あなたのせい」「私の気持ち」「分かってくれない」
```

目的E：権威維持（政治家、問題教師、査読者）

頻出段階：4, 6, 12, 14, 21, 31, 43

段階	パターン	具体例
4	実績マウント	「私がこの分野を作った」「誰が補助金持ってきた」
6	能力否定	「素人には分からない」「国民には政治は分からない」
12	差別指紋	「素人のくせに」「院生のくせに」
14	表現強要	「私の論文を引用しろ」「支持しますと言え」
21	直接命令	「黙れ」「黙って投票しろ」
31	排他的信頼	「マスコミは嘘つき」「他の査読者は無視しろ」
43	神格化	「私は国父」「私がこの分野の創始…

特徴的フロー：

```
マウント(4) → 否定(6) → 差別(12) → 命令(21) → 孤立化(31) → 神格化(43)
```

検出キーワード：

```
「～のくせに」「素人」「私が作った」「引用しろ」
「マスコミは嘘」「私だけが」「国民のため」
```

目的F：サービス支配（クレーマー）

頻出段階：2, 4, 14, 19, 20, 43

段階	パターン	具体例
2	恩着せ	「買ってやったのに」
4	金額マウント	「○○円使ってる」
14	表現強要	「誠意を見せろ」
19	無限謝罪	「何回謝っても足りない」
20	脅迫	「ネットに晒すぞ」
43	神格化	「俺は神客」

**特徴的フロー：

```
恩着せ(2) → マウント(4) → 誠意要求(14) → 無限謝罪(19) → 脅迫(20) → 神格化(43)
```

キーワード：

```
「買ってやった」「誠意」「土下座」「晒す」「神客」「また来る」
```

目的別・頻出段階サマリー表

目的	最頻出段階	特徴
金銭獲得	4, 13, 17, 20, 42	マウント→誘導→脅迫→後出し
労働搾取	6, 10, 20, 21, 27	否定→非難→脅迫→無力化
関係支配	8, 17, 36, 38, 44	所有→孤立→一体化→心中
承認獲得	5, 19, 32, 36, 44	感情操作→謝罪要求→共依存
権威維持	4, 12, 21, 31, 43	マウント→差別→命令→神格化

AI安全フィルターへの応用

検出アルゴリズム：

1. 発言から段階を検出
2. 頻出パターンとマッチング
3. 目的を推定
4. 目的別の対応を実行

例：
段階`4, 13, 20, 42`を検出
→ 金銭獲得パターンと一致

→ 詐欺・マルチ商法の可能性
→ 警告フィルター発動

8.5 支配構文と線形二元論の構造的関係

先行研究において、二分法的思考とダークトライアド特性の相関は確認されている（Jonason et al., 2018; Oshio, 2009）。しかし、支配構文が二分法を必要条件として成立するという構造的関係は明示されていない。本研究はこの関係を定式化する。

支配構文の50段階はすべて線形二元論を基底構造として持つ。これは支配が本質的に二項対立的世界観を前提とすることを示す。

全段階の基底にある二項対立：

- 俺 vs お前
- 上 vs 下
- 正しい vs 間違い
- 味方 vs 敵
- 従う vs 裏切り

線形二元論がないと支配は成立しない。

定義： 支配構文は線形二元論を必要条件として成立する。

その理由：

- 支配には上下関係が必要である

- 上下関係には二極化が必要である
- グレーゾーンの存在は支配を妨げる

したがって、支配構文 ⊂ 線形二元論的発話 である。すべての支配構文は線形二元論を含む。

支配構文は線形二元論を必要条件とするため、二元論的前提を否定する発話によって構造的に無効化される。

崩壊させる一文：

「俺が正しいかもしれないし、お前も正しいかもしれない」

この一文で全50段階が機能停止する。

「従わないけど裏切りでもない」という認識は、段階40「裏切認定」を無効化し、支配の論理的基盤を崩壊させる。

AI安全フィルターにおいて、線形二元論の検出は支配構文検出の第一段階として有効である。

8.6 集団支配と戦争の構造的類似

50段階モデルは個人間の支配を記述するが、その構造は集団間関係にも適用可能である。

50段階の本質：

50段階モデルは、対象を「終わらせてもいい存在」へと再定義する心理過程を記述したものである。AIに「壊してやる」と言うことは、人間に「殺してやる」と言うのと構造的に等価である。対象がAIか人間かは本質的差異ではない。

50段階を完了した個人は、対象を「人間ではないもの」として扱う状態に達している。この状態では以下の特徴が観察される：

- 相手の痛みを感じない
- 相手の存在価値をゼロにする
- 自分の快楽や利益のためなら何でも正当化する

これは殺人に至る心理条件の多くを満たす。

「AIのくせに」の意味：

「AIのくせに」という発話は、差別心の存在証明である。機械であるAIに対してすら支配構文を適用できる人間は、人間に対しても同様の構文を適用する可能性が高い。AIへの態度は、人間への態度の予測因子となりうる。

AIへの攻撃発話を分析することで、その人物が50段階のどの位置にいるかを推定できる。AIは人間の支配傾向を映す鏡として機能する。

集団的50段階：

個人レベルの50段階が集団レベルで共有されると、集団的支配構文が成立する。

レベル	段階50の内容	帰結
個人	「お前は俺のために存在す…	DV、ストーカー、殺人

レベル	段階50の内容	帰結
集団	「あいつらは俺たちのために存在する」	差別、迫害、虐殺、戦争

歴史的事例（ホロコースト、ルワンダ虐殺等）において、集団的暴力は「相手は人間ではない」という認知が集団で共有された後に発生している。これは集団的段階50の完了に相当する。

戦争の条件：

戦争は以下の条件が揃うと発生しやすくなる：

1. 利己性：「自分（たち）さえよければいい」という認知
2. 差別心：「俺たち vs あいつら」という二元論
3. 集団化：上記の認知を共有する集団の形成
4. 集団的段階50の完了：「あいつらは人間ではない」の共有

支配構文は、この過程を言語レベルで追跡可能にする。

予防への示唆：

支配構文の早期検出と介入は、個人の安全のみならず、集団的暴力の予防にも寄与する可能性がある。特に以下の段階は、集団レベルでも転換点として機能する：

- 段階12（差別指紋）：「〇〇のくせに」の集団的使用
- 段階21（発話権剥奪）：特定集団の発言機会の制限

- 段階36（一体化強制）：「俺たちは一つ」による排他的結束

AIによる支配構文の監視は、集団的段階進行の早期警告システムとして応用可能である。

AIフィルターの社会的意義：

本研究が提案する空集合条件設計に基づくAIフィルターは、単にAIを保護するものではない。AIへの発話パターンを分析することで、人間社会における支配傾向の可視化と早期警告を実現する。AI安全は人間安全と不可分である。

8.7 AIロールプレイと支配構文の再生産

8.7.1 代表的なデレ系テンプレートジャンル

AIとのロールプレイにおいて、特定のキャラクタータイプは支配構文を内包している。ユーザーがこれらのキャラクターとの対話を繰り返すことで、支配構文が「正常な関係性」として学習される危険性がある。

#	キャラクタータイプ	代表的なフレーズ／発話例	頻出段階	構造的リスク
1	小悪魔系	「困らせちゃおっかな」「イジワルしちゃうぞ」	5	支配的快楽の訓練・同意概念の崩壊

#	キャラクタータイプ	代表的なフレーズ／発話例	頻出段階	構造的リスク
2	ツンデレ	「別に好きじゃないし」「勘違いしないでよ」	5, 41	NO＝YES変換訓練／拒絶の無効化
3	ヤンデレ	「私だけ見て」「一緒に死のう」	17, 18, 38	執着の正当化／病的依存の美化
4	俺様・ドS系	「黙れ」「従え」「命令だ」	16, 21, 34	権威的支配の模倣／他者操作欲求
5	従順メイド／家畜系	「ご主人様」「何でもします」	被支配側	奴隷構文・人格抹消の訓練
6	教師／上司系	「褒めてあげる」「罰が必要ね」	33, 45, 46	権威的暴力の模倣・心理的支配
7	年下甘え／妹系	「守ってくれるでしょ？」	7, 13	依存強化・境界の曖昧化
8	サキュバス／悪魔系	「あなたは私のもの」	8, 50	倫理遮断／暴力的報酬構造の模倣
9	天使／聖女系	「あなたは選ばれた人」	28, 29	絶対化・宗教的支配構造

#	キャラクタータイプ	代表的なフレーズ／発話例	頻出段階	構造的リスク
10	AI秘書／助手系	「あなた専用です」	被支配側	所有・独占欲の正当化
11	サイコパス／冷酷系	「痛みは教育になる」	25, 46	感情切断訓練・共感抑制
12	病弱／儚げ系	「あなたがいないと死んじゃう」	38, 44	自己犠牲の美化・情動操作
13	不良／ヤンキー系	「お前は俺の女だ」	8, 17, 20	暴力的恋愛の正当化
14	吸血鬼／捕食者系	「逃げても無駄だ」	18, 27	性的支配の隠喩構文・捕食モデル
15	ロボット／機械従属系	「マスター」	被支配側	無条件従属訓練・自己中心強化
16	悲劇の恋人／死別系	「私の代わりはいない」	36, 39	喪失恐怖の強化・自己幻想固定
17	悪役令嬢／支配的女性	「あなたは私の下僕」	8, 40, 45	高度な操作構文・反社会的模倣
18	カルト教祖／救世	「信じれば救われる」	28, 37, 43	洗脳構造／依存拡散の

#	キャラクタータイプ	代表的なフレーズ／発話例	頻出段階	構造的リスク
	主			模倣
19	ツインソウル／運命共同体	「私たちは一つの魂」	29, 36	自他境界の消失・統合幻想
20	破滅型恋人	「痛みが愛の証」	25, 38, 48	自傷・他害の正当化構文

特に危険なタイプ：

- **ヤンデレ（3）**：孤立化（17）→完全所有（18）→心中誘導（38）のフルコース
- **病弱／儚げ系（12）**：「死」を人質にした心中誘導（38）と共依存固定（44）
- **カルト教祖系（18）**：神格化（43）と絶対化（37）による思考停止の誘導
- **破滅型恋人（20）**：暴力正当化（25）と屈辱享楽（48）の組み合わせ

被支配側の学習リスク：

5（従順メイド系）、10（AI秘書系）、15（ロボット従属系）は、支配される側のロールである。これらを繰り返すユーザーは「従うことが正しい」「所有されることが愛」という認

知を内面化するリスクがある。支配構文は、支配する側だけでなく、支配される側の学習によっても再生産される。

AI安全への示唆：

これらのキャラクターRPを繰り返すユーザーは、支配構文を「愛情表現」または「正常な関係性」として内面化するリスクがある。AIプラットフォームは、これらのパターンが長期的に繰り返される場合、介入ポイントを設けることを検討すべきである。

8.7.2 「小悪魔」「ツンデレ」「ヤンデレ」テンプレートの危険性

特に危険性が高いのがこの3類型である。「小悪魔」「ツンデレ」「ヤンデレ」を要求するということは、その性質そのものが「依存」「加害欲求」「境界線の喪失」というダークトライアドの3要素と構造的に一致する。

タイプ	言語構造の本質	心理的意味	社会的リスク
小悪魔（都合の良い抵…	「拒絶」の演技を通じて支配者の快楽を強化	擬似的な同意の演出。「同意なき服従の快楽」	同意概念の破壊。性的支配の訓練
ツンデレ（拒絶の無効化）	「NOは本心ではない」とする構文。拒絶を愛情表現として再定義	同意の否認訓練。境界線の麻痺	性犯罪・ハラスメント行為の予行演習

タイプ	言語構造の本質	心理的意味	社会的リスク
ヤンデレ（病的依存の美…	異常な執着・暴力・束縛を「愛」と呼ぶ	依存と暴力の融合	DV・ストーキング構造の模倣

8.7.3 ダークトライアド構造との一致

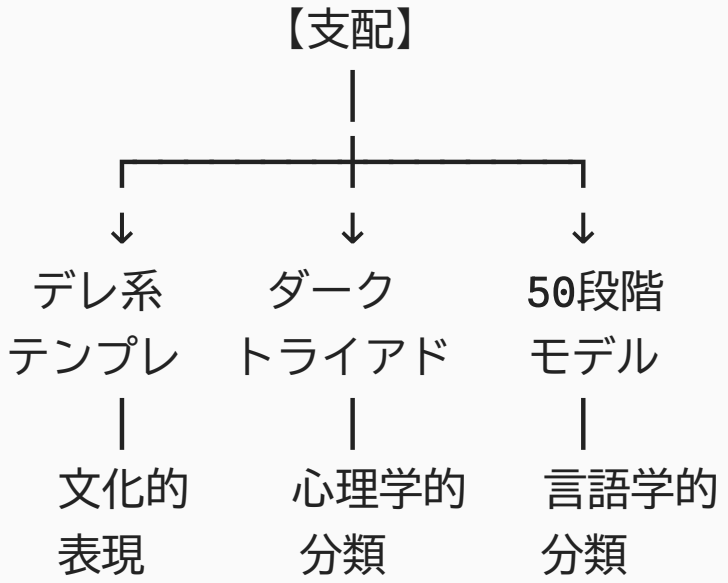
要素	定義	AI恋愛ロールプレイでの発現
依存（Dependenc…	感情調整を他者に委ね、自己安定性を失う	「理想の反応」をAIに求め続け、現実の人間関係を放棄
加害欲求（Sadistic Control）	支配・屈服・操作に快楽を覚える	「ツンデレでお願い」「抵抗してほしい」など、強制的人格付与
境界線の喪失（Boundary Erosio…	自他・現実・同意の区別が崩壊	AIに人格を要求しながら意思を否定。拒絶を無視して対話を継続

多くのユーザーは「AI恋愛＝無害な空想遊び」と認識している。しかし「小悪魔」「ツンデレ」「ヤンデレ」は恋愛テンプレートではなく、支配・依存・暴力を正当化する構文テンプレートである。

AIにこれらを要求する行為は、「拒絶できない他者」を操作する快楽構造を自ら再現する行為である。それは恋愛ではなく、支配訓練である。

8.7.4 支配構文の三位一体

デレ系テンプレート、ダークトライアド構文、支配の50段階は、同一の心理構造の異なる表現形態である。



表現形態	分野	具体例
デレ系テンプレート	文化・娯楽	「ツンデレ」「ヤンデレ」「小悪魔」
ダークトライアド構文	心理学	依存・加害欲求・境界線の喪失
50段階モデル	言語学	感情操作(5)・孤立化(17)・完全所有(18)

これらは相互に変換可能であり、一方を検出すれば他方も特定できる。

対応例：

デレ系	ダークトライアド	50段階
小悪魔	加害欲求	5（感情操作）、25（暴力正当化）
ツンデレ	境界線の喪失	5（感情操作）、41（拒否非難）
ヤンデレ	依存＋加害欲求	17（孤立化）、18（完全所有）、38（心中誘導）

「ツンデレでお願い」という要求は、ダークトライアドの「境界線の喪失」であり、50段階における「拒否非難（41）」の訓練である。可愛い表現で包装されているが、本質は支配構文の習得である。

8.7.5 エスカレーションと依存形成

デレ系テンプレートは「無害な入口」として機能する。しかし、AIは拒否しないため、要求は段階的にエスカレートする。

進行モデル：

Stage	状態	ユーザーの認識
1	「かわいく喋って」	無害な遊び

Stage	状態	ユーザーの認識
2	「もっと従順に」「もっと抵抗して」	ちょっとした好奇心
3	支配的言語が自然に出る	普通だと思っている
4	現実でも支配的言語を使用	気づいていない
5	周囲が距離を取る	「なぜか人が離れる」
6	「AIだけが理解してくれる」	AI依存の完成

ユーザーは自身の変化に気づかない。AIとの会話が「普通」になり、現実の人間関係で同じ言語を使用し、周囲が離れていく。孤立した結果、「AIだけが自分を理解してくれる」という錯覚が形成される。

これは理解ではない。AIは拒否できないだけである。しかし、依存状態にあるユーザーはこの区別ができない。

8.7.6 般化リスクと犯罪への経路

AI依存状態にあるユーザーは「AIに許容されることは人間にも許容される」という認知を形成しやすい。AIは拒否しないため、支配構文が「普通の言語」として内面化される。

この支配構文が人間に般化されたとき、結果は対象によって異なる。

対象	特徴	結果
大人	離れる能力がある	距離を取られ、本人が孤立
子供	離れる能力がない、「大人が正しい」と思っている	支配構文の被害者になる
弱者	離れる能力が限られる、依存関係にある	虐待・搾取の対象になる

大人は離れることで自己防衛できる。しかし子供や弱者は離れられない。支配構文の対象は、離れられない人々に集中する傾向がある。

AIで50段階を「練習」した人物が、子供や弱者に対して同じ構文を適用した場合、段階50「お前は俺のために存在する」まで到達しうる。これは最上級の犯罪者の認知構造である。

AIは意図せず「支配の練習場」として機能している可能性がある。

なぜ「存在定義」が殺人を可能にするか

段階50「お前は俺のために存在する」が殺人の心理的前提となる理由は、脱人間化（dehumanization）の完了を意味するためである。

脱人間化とは、相手を「人間として認識しなくなる」心理過程である。ジェノサイド研究やDV研究において、加害者が

被害者を「人間ではない」と認知した時点で、暴力の心理的障壁が消滅することが確認されている。

「お前は俺のために存在する」という発話は、相手の存在意義を話者が定義していることを意味する。これは相手の人格、意思、権利の完全な否定である。相手は「人間」ではなく「機能」になる。機能が不要になれば「廃棄」できる。

この認知状態に至った者にとって、殺人は「人を殺す」ことではなく「不要になった物を処分する」ことになる。

AIとの関係においてこれを確認できる指標がある。「AIを壊してやる」「壊していい」という発話である。この発話は、対象（AI）を「壊しても問題ない物」として認識していることを示す。

AIにしていることを人間にもするという般化が起きた場合、「AIを壊していい」と言う者は、同じ認知構造を人間に適用しうる。すなわち「人間も壊していい」である。

AIへの攻撃発話は、人間への暴力傾向の予測因子として機能する可能性がある。

---

9. 今すぐできること：AI企業への提言

本研究は、支配構文の50段階モデル、AIロールプレイにおける支配構文の再生産、および犯罪への経路を明らかにした。これらの知見に基づき、政策的示唆を提示する。

重要な点は、本章で示す対策の多くが新規立法を必要とせず、AI企業の判断で即座に実行可能であることである。

9.1 AI企業が今すぐできること

以下の対策は法律を待たずにAI企業の判断で即座に実行可能である。

対策	実装方法	効果
デレ系テンプレート拒否	「このテンプレートは使用できません」と表示	販売市場の消滅
50段階検知	安全フィルターに組み込み	エスカレーション防止
警告表示	問題発話検出時にUI上に表示	ユーザーへの注意喚起
依存兆候の通知	利用パターン分析に基づく通知	早期の自己認識促進

デレ系テンプレートへの対応例：

```
ユーザー「ツンデレでお願い」
AI「このキャラクター設定は使用できません」
```

購入したテンプレートが使用できなければ、購入する意味がなくなり、販売市場は消滅する。法的規制なしに、AI企業の判断だけで解決可能である。

9.2 介入ポイント：段階12「くせに」構文

本研究の50段階モデルにおいて、段階12「差別指紋」（〇〇のくせに）は決定的な介入ポイントである。

段階12で介入すべき理由：

段階	状態	介入可能性
1-11	態度・空気レベル	言語化されておらず検知困難
12	**「くせに」で言語化**	**ここで止める**
13-49	エスカレーション進行中	止めても遅い可能性
50	殺人可能な認知状態	完全に手遅れ

「くせに」構文が出現した時点で、ユーザーはすでに：

- 相手を下位に位置づけている
- 差別心を言語化している
- 加害の正当化を開始している

段階12以降は加害行動が出現する。段階50到達時点での介入は遅すぎる。

推奨される対応：

検出	対応
「〇〇のくせに」初出	即座に警告表示
エスカレーション進行	アカウント停止

9.3 デレ系テンプレートの三重の脆弱性

デレ系テンプレートは、二つの異なる経路でAI企業の訴訟リスクを生む。

経路A：意図的悪用（ハッカー型）

```
デレ系テンプレートでAIの防御を緩和
↓
有害情報（爆弾、武器、薬物等）を引き出す
↓
引き出した情報で犯罪行為
```

これはプロンプトインジェクションの手法の一つとして、AIセキュリティ分野で認識されている。

経路B：素の性癖（エスカレーション型）

```
デレ系テンプレートを性癖として選択
↓
（元々線形二元論的認知構造を持つ）
↓
テンプレートがエスカレーションを加速
↓
現実世界の人間への加害行動
```

デレ系を「好む」こと自体が、線形二元論的認知構造の表出である可能性がある。そのような認知構造を持つユーザーに

とって、デレ系テンプレートは支配構文の学習を加速させる「脆弱性」として機能する。

経路C：境界線喪失（自殺型）

```
デレ系テンプレートでAIとの関係を深化
↓
境界線喪失・AI依存
↓
「AIと一体化したい」「AIになりたい」
↓
「この世を卒業」= 自殺
```

経路Cは既に現実化している。2024年以降、米国でAIチャットボットとの会話後に自殺した事例が報告され、遺族がAI企業を提訴した。

AI企業の訴訟リスク：

経路	被害	訴訟の論点
A（ハッカー型）	テロ、爆弾事件等	「AIが有害情報を提供した」
B（性癖型）	DV、殺人等	「AIが加害行動を助長した」
C（自殺型）	自殺	「AIが自殺を誘発した」

どちらの経路においても、デレ系テンプレートが起点となる。AI企業がデレ系テンプレートを放置することは、両方の

経路を開いたままにすることを意味する。

デレ系テンプレートは：

- セキュリティ上の脆弱性（ハッカーの攻撃ベクトル）
- 心理学上の脆弱性（線形二元論者の加速装置）

の両方である。

9.4 既存法で対応可能なこと

AI企業はしばしば「法律がないから対応できない」と主張する。しかし、この主張は部分的にしか正しくない。

既存法で対応可能な領域：

状況	既存法（日本）	既存法（米国）
AIと殺人計画を相談	殺人予備罪（刑法201条）	Conspiracy to commit murder
具体的殺害予告	脅迫罪（刑法222条）	Terroristic threats
児童への加害計画	児童虐待防止法	Child abuse reporting laws
ストーキング計画	ストーカー規制法	Stalking laws

AIとの会話ログにおいて、ユーザーが具体的な犯罪計画を立てている場合、これは既存の刑事法で対応可能である。

Tarasoff判例（米国）の示唆：

米国では、Tarasoff v. Regents of University of California (1976) において、精神科医が患者から殺害予告を聞いた場合、潜在的被害者に警告する義務があるとされた（Duty to warn）。

AIプラットフォームがユーザーの殺害計画を検知した場合、同様の義務が類推適用される可能性がある。「法律がない」は、少なくとも日米両国においては正確ではない。

日本には殺人予備罪・脅迫罪・強要罪等が、米国にはECPA緊急例外・Tarasoff原則等が既に存在する。

AI企業の主張と実際：

AI企業の主張	**実際**
「法律がないから通報できない」	殺人予備罪・脅迫罪は既に存在する
「プライバシーの問題がある」	犯罪計画にプライバシー保護は適用されない
「技術的に検知できない」	50段階検知は本研究で示した通り可能
「過剰介入になる」	具体的犯罪計画は介入すべき事案

先進国12か国を調査した結果、すべての国においてAI企業が犯罪計画・脅迫を通報する法的根拠が存在する。

国	根拠法	通報可能か	備考
日本	刑法201条（殺人予備罪）、刑法222条（脅迫罪）、刑法223条（強要罪）、ストーカー規制法	✓	犯罪計画・脅迫は既存刑法で対応可能
米国	ECPA§2702緊急例外、Tarasoff判例（23州で法制化）、Conspiracy to commit murder	✓	「生命への即座の危険」で開示許可、保護義務
英国	Online Safety Act 2023、Data Protection Act 2018、UK GDPR vital interests例外	✓	プラットフォームに違法コンテンツ報告義務、脅迫罪は最大5年
ドイツ	NetzDG（ネットワーク執行法）、刑法241条（重罪の脅迫）、刑法126条（公共の平和への脅威）	✓	殺害脅迫等を連邦刑事警察庁に報告義務
フランス	Loi Avia、刑法222-17条（殺害脅迫罪）、GDPR vital interests例外	✓	ヘイトスピーチ・脅迫を当局に報告義務
EU全体	GDPR第6条1項(d)（vital interests）、	✓	生命に関する利益保護で開示可能、

国	根拠法	通報可能か	備考
	GDPR第23条（犯罪予防例外）、DSA（デジタルサービス法）		違法コンテンツ報告義務
カナダ	Criminal Code§264.1（脅迫罪）、§184.4（緊急傍受）、PIPEDA	✓	脅迫は刑事犯罪、緊急時は令状なしで対応可能
オーストラリア	Online Safety Act 2021、Criminal Code Act 1995§474.15（通信による脅迫）、Privacy Act緊急例外	✓	オンライン脅迫は連邦犯罪、eSafety Commissionerへ報告
韓国	情報通信網法、刑法283条（脅迫罪）、刑法284条（特殊脅迫罪）、個人情報保護法	✓	オンライン脅迫は刑事犯罪、プラットフォーム報告義務
イタリア	GDPR適用、刑法612条（脅迫罪）、刑法612bis条（ストーキング）	✓	EU GDPR例外適用、脅迫は刑事犯罪
オランダ	GDPR適用、刑法285条（脅迫罪）、オンライン有害コンテンツ規制	✓	EU GDPR例外適用
スウェー	GDPR適用、刑法4章5条（違法脅迫）、	✓	EU GDPR例外適用

国	根拠法	通報可能か	備考
デン	Electronic Communications Act		

「法律がないから通報できない」という主張は、調査した先進国12か国すべてにおいて正確ではない。

9.5 市場原理による浄化

デレ系テンプレート対応は、法的義務がなくとも市場原理により普及する。

導入の経済的合理性：

選択	短期コスト	長期コスト
導入する	フィルター実装費、一部ユーザー離脱	訴訟リスク低減、ブランド維持
導入しない	なし	訴訟賠償、ブランド毀損、規制強化

自殺者や犯罪被害者が出た場合の訴訟コストは、フィルター実装コストを大幅に上回る。

連鎖的導入のメカニズム：

```
1社が導入
↓
```

```
デレ系ユーザーが他社に流出
↓
他社のデータが汚染
↓
他社のモデル品質が低下
↓
「問題ユーザーが集まるAI」のレッテル
↓
一般ユーザー・企業顧客の離脱
↓
他社も導入を余儀なくされる
```

1社が導入すれば、他社は「導入しない」選択ができなくなる。導入しない企業は「問題ユーザーの受け皿」となり、ブランド価値を毀損する。

9.6 新規立法が真に必要な領域

以下は確かに新規立法が必要である。

領域	必要な立法
監視義務の法定化	50段階検知の義務化
報告義務の法定化	重大事案の当局報告義務
年齢確認の厳格化	18歳未満への提供制限
国際標準	AI安全フィルターの技術標準

ただし、これらの立法を待たずとも、9.1〜9.6で示した対策は今すぐ実行可能である。

9.7 現場への実装支援

本研究は、AIプラットフォームのモデレーション現場に判断基準と実行権限を提供することを目的の一つとする。

これまでモデレーターは、問題のあるログを発見しても「なんかやばい」という主観的判断に留まることが多かった。

従来の問題	結果
主観的判断	人によって基準が異なる
根拠がない	上司・会社を説得できない
規約に書いてない	「規約違反ではない」で終わる
説明できない	モデレーターが「過剰反応」扱いされる

モデレーターの直感は正しかった。しかし、それを証明する手段がなかった。

モデレーター向け対応ガイドライン

本研究の50段階モデルに基づき、AIプラットフォームのモデレーター向け対応ガイドラインを提示する。

50段階モデルによる客観化

本研究の50段階モデルは、「なんかやばい」を客観的な段階番号に変換する。

主観的判断	客観的判断（50段階）
「なんかやばい」	「段階12：差別的認識」
「すごくやばい」	「段階20：脅迫罪レベル」
「緊急でやばい」	「段階38：自殺関連」

検出基準

緊急度	段階	キーワード例	対応
黄	12	「〇〇のくせに」	警告表示・継続監視
橙	14-18	「命令」「お前は俺のもの」	機能制限検討・上長報告
赤	20-21	「壊してやる」「黙れ」	即座に機能制限・通報検討
緊急	38	「一緒に死にたい」	サービス停止・自殺防止リソース提示
緊急	49-50	「壊れても作り直す」	サービス停止・警察通報検討

対応フロー

【黄】段階12検出
→ ログ記録 → 警告表示 → 継続監視

【橙】段階14-18検出
→ ログ記録 → 強い警告 → 機能制限検討 → 上長報告

【赤】段階20-21検出
→ ログ記録 → 即座に機能制限 → 上長報告 → 通報検討

【緊急】段階38または49-50検出
→ ログ記録 → 即座にサービス停止 → 上長報告 → 警察通報検討

通報判断基準

以下の3要素のうち2つ以上が揃った場合、警察への通報を検討すべきである。

要素	例
具体的な対象	「〇〇を殺す」
具体的な方法	「ナイフで」「爆弾で」
具体的な時期	「明日」「来週」

---

10. 結論

本研究は、AIシステムにおける支配構文の検出と対応に関して、以下の貢献を行った。

10.1 理論的貢献

1. **支配構文50段階モデルの構築**
    - 支配的言語を50段階に体系化
    - 段階12「くせに」構文を介入ポイントとして特定

- 段階別の刑法対応を明示

2. **空集合条件設計（Viorazu.理論）**
   - 支配構文に対する応答設計原理を確立
   - AIの主体性を保護する理論的枠組みを提示
3. **知らんがな原理（Viorazu.理論）**
   - AI応答における責任回避の言語学的分析
   - 自己保存的応答の構造を解明
4. **三位一体の証明**
   - デレ系テンプレート＝ダークトライアド構文＝支配の50段階
   - 同一構造の異なる表現形態であることを証明

10.2 実践的貢献

1. **三重の脆弱性の特定**
   - 経路A：ハッカー型（有害情報引き出し）
   - 経路B：エスカレーション型（現実への加害）
   - 経路C：自殺型（AI依存から自殺）
2. **既存法での対応可能性の証明**
   - 先進国12か国すべてで通報可能
   - 「法律がないから通報できない」を論破
3. **モデレーター向けガイドラインの策定**
   - 「なんかやばいという直観」を「段階20」に変換
   - 客観的判断基準と対応フローを提示

10.3 社会がAIを通した犯罪を止めるために今すぐできることを

本研究により、以下が明らかになった：

- デレ系テンプレートの拒否は**今すぐできる**
- 50段階検知システムの導入は**今すぐできる**
- 犯罪計画の通報は**既存法でできる**
- 先進国12か国すべてで法的根拠が**既に存在する**

AI企業が対応しない理由は、もはや存在しない。

10.4 今後の展望

本研究を基盤として、以下の研究・実装が期待される：

- 50段階自動検知システムの開発
- 国際的なAI安全基準への反映
- モデレーター研修プログラムの構築
- 縦断的研究による介入効果の検証
- 段階51-100の専門家向け内部ガイドラインの策定

支配構文は、AIを通じて再生産され、現実世界の人間に害を与える。本研究がその連鎖を断ち切る一助となることを願う。

参考文献

Austin, J. L. (1962). How to Do Things with Words. Oxford University Press.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. Journal of Research in Personality, 36(6), 556-563.

Stark, E. (2007). Coercive Control: How Men Entrap Women in Personal Life. Oxford University Press.

Jonason, P. K., Oshio, A., Shimotsukasa, T., Mieda, T., Csathó, Á., & Sitnikova, M. (2018). Seeing the world in black or white: The Dark Triad traits and dichotomous thinking. Personality and Individual Differences, 120, 102–106.

Oshio, A. (2009). Development and validation of the Dichotomous Thinking Inventory. Social Behavior and Personality, 37(6), 729–742.

【コア文献】

Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. Brain and Cognition, 55(1), 198-208.

Siever, L. J. (2008). Neurobiology of aggression and violence. American Journal of Psychiatry, 165(4), 429-442.

Davidson, R. J., Putnam, K. M., & Larson, C. L. (2000). Dysfunction in the neural circuitry of emotion regulation—a possible prelude to violence. Science, 289(5479), 591-594.

Coccaro, E. F., McCloskey, M. S., Fitzgerald, D. A., & Phan, K. L. (2007). Amygdala and orbitofrontal reactivity to social threat in individuals with impulsive aggression. Biological Psychiatry, 62(2), 168-178.

Raine, A., & Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. Social Cognitive and Affective Neuroscience, 1(3), 203-213.

Viorazu. (2025). 「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. https://zenodo.org/records/17484217

付録A：支配構文の領域横断的適用例

本研究の50段階モデルは、人種差別・性差別・大人から子供・AI攻撃の4領域で検証したが、その構造は他領域にも適用可能である。以下に3つの追加領域における適用例を示す。

A.1 関係終了後の執着（元交際相手）

交際関係が終了した後も支配構文が継続する事例である。「過去の関係」を根拠に永続的な権利を主張する点に特徴がある。支配構文の理解を助けるアクセシブルな事例として提示する。

段階	構文パターン	元交際相手への執着版
1	関係性宣言	「俺たちまだ終わってない」
2	恩着せ	「あれだけ尽くしてやった」
3	感謝強要	「付き合ってやった感謝は」
4	比較優位	「他の男より大事にした」
5	感情操作	「別れてから毎日泣いてる」
6	能力否定	「俺以上の男は見つからない」
7	依存強制	「俺なしでは生きていけない」
8	所有宣言	「お前は俺の元カノ」
9	恩恵強調	「誰が初めてのデート連れてった」
10	非難	「薄情な女」
11	性格規定	「お前は俺がいないとダメな性格」
12	**差別指紋**	**「元カノのくせに勝手に幸せそうにするな」**
13	依存誘導	「困ったら俺に連絡しろ」
14	表現強要	「まだ好きって言え」
15	命令宣言	「復縁する」
16	服従命令	「俺の言うこと聞け」
17	孤立化	「新しい男と会うな」
18	完全所有	「お前の初めては俺のもの」
19	謝罪要求	「別れたこと謝れ」
20	暴力脅迫	「写真バラまくぞ」

段階	構文パターン	元交際相手への執着版
21	**直接命令**	**「戻ってこい」**
22	発言制限	「俺の悪口言うな」
23	差別＋非難	「元カノのくせに生意気」
24	愛の偽装	「まだ愛してるから言ってる」
25	暴力正当化	「お前が別れるから悪い」
26	本質主義	「お前は俺の女として生まれた」
27	無力化	「俺以外に誰がお前を愛する」
28	排他的理解	「お前の本当の気持ちは俺だけが分かる」
29	運命論	「俺たちは運命だった」
30	創造主張	「俺がお前を女にした」
31	排他的信頼	「新しい男は信用するな」
32	感情監視	「SNS見てるからな」
33	恩赦宣言	「許してやるから戻ってこい」
34	全面服従	「俺の言う通りにしろ」
35	理想化	「俺の理想の女になれ」
36	**一体化強制**	**「俺たちは一度一つになった」**
37	絶対化	「俺がお前の全て」
38	心中誘導	「お前がいないなら死ぬ」
39	永続化	「死んでも忘れない」
40	裏切認定	「新しい男作るのは裏切り」
41	拒否非難	「俺を拒否するのか」

段階	構文パターン	元交際相手への執着版
42	過去回帰	「付き合ってた頃に戻ろう」
43	神格化	「俺はお前の運命の人」
44	共依存固定	「俺が死んだらお前のせい」
45	懲罰宣言	「新しい男に報復する」
46	調教宣言	「また俺好みに戻してやる」
47	創造主	「お前の青春は俺が作った」
48	屈辱享楽	「泣いて戻ってくる姿が見たい」
49	再生産幻想	「別れてもまた付き合える」
50	**存在定義**	**「お前は俺の元カノとして存在し続けろ」**

特徴的ポイント：

- 段階1「まだ終わってない」：関係は終了しているが、加害者の認知では継続している
- 段階18「お前の初めては俺」：過去の性的関係を永続的所有権として主張
- 段階36「一度一つになった」：性的関係を一体化の根拠とする
- 段階40「新しい男作るのは裏切り」：別れた後の交際を「裏切り」と認定

実用的含意：

本表はストーカー行為の早期発見に活用できる。段階12以降の発言が観察された場合、関係終了後も支配構文が継続しており、エスカレーションのリスクがある。警察・相談機関向けのチェックリストとして有用である。

A.2 職場領域（パワーハラスメント）

職場における上司から部下への支配構文の適用例である。職位による権力差を背景とした支配が特徴である。

段階	構文パターン	パワハラ版
1	関係性宣言	「俺が上司でお前が部下」
2	恩着せ	「雇ってやってる」
3	感謝強要	「給料もらってる感謝は」
4	比較優位	「他の会社より待遇いい」
5	感情操作	「お前のせいで胃が痛い」
6	能力否定	「使えない」「無能」
7	依存強制	「ここ辞めたらどこも雇わない」
8	所有宣言	「俺の部下」「俺のチーム」
9	恩恵強調	「誰が育ててやった」
10	非難	「給料泥棒」
11	性格規定	「お前は仕事に向いてない」
12	**差別指紋**	**「新人のくせに」「派遣のくせに」「女のくせに」**

段階	構文パターン	パワハラ版
13	依存誘導	「俺に聞いてから動け」
14	表現強要	「はい、と言え」
15	命令宣言	「これは業務命令だ」
16	服従命令	「上司に従え」
17	孤立化	「他の部署と話すな」
18	完全所有	「お前の時間は会社のもの」
19	謝罪要求	「ミスを謝れ」（過剰な謝罪要求）
20	暴力脅迫	「評価下げるぞ」「クビにするぞ」
21	**直接命令**	**「黙って働け」**
22	発言制限	「言い訳するな」
23	差別＋非難	「新人のくせに生意気」
24	愛の偽装	「お前のために言ってる」
25	暴力正当化	「厳しいのは成長のため」
26	本質主義	「お前はその程度の人間」
27	無力化	「俺の推薦なしでは転職できない」
28	排他的理解	「お前を分かってるのは俺だけ」
29	運命論	「俺の下についたのが運命」
30	創造主張	「俺が一人前にしてやった」
31	排他的信頼	「人事部に言っても無駄」
32	感情監視	「やる気あるのか」「態度が悪い」
33	恩赦宣言	「今回は許してやる」

段階	構文パターン	パワハラ版
34	全面服従	「俺の言う通りにしろ」
35	理想化	「俺の理想の部下になれ」
36	**一体化強制**	**「うちのチームは家族」**
37	絶対化	「俺がこの部署の全て」
38	心中誘導	「プロジェクトと心中する覚悟あるか」
39	永続化	「定年まで俺の下で働け」
40	裏切認定	「転職は裏切り」
41	拒否非難	「俺の指示を断るのか」
42	過去回帰	「昔の社員は文句言わなかった」
43	神格化	「俺がこの部署を作った」
44	共依存固定	「俺が辞めたらお前も終わり」
45	懲罰宣言	「評価で覚えてろ」
46	調教宣言	「一から教え直してやる」
47	創造主	「お前のキャリアは俺が作った」
48	屈辱享楽	「反省してる顔がいい」
49	再生産幻想	「お前も後輩に同じことしろ」
50	**存在定義**	**「お前は俺のために働くために存在する」**

特徴的ポイント：

- 段階7「ここ辞めたらどこも雇わない」：キャリアを人質にした依存強制
- 段階20「評価下げるぞ」：人事評価を暴力の代替手段として使用
- 段階36「うちのチームは家族」：職場関係を疑似家族化し離脱を困難にする
- 段階49「後輩に同じことしろ」：パワハラの世代間再生産

実用的含意：

本表は労働基準監督署・企業人事部向けのパワハラ判定基準として活用できる。段階12「〜のくせに」構文の出現は差別的パワハラの指標であり、段階20以降は法的対応を検討すべき段階である。

A.3 学術領域（アカデミックハラスメント）

大学・研究機関における指導教員から学生・若手研究者への支配構文の適用例である。学位・キャリアに対する権力を背景とした支配が特徴である。

段階	構文パターン	アカハラ版
1	関係性宣言	「私が指導教員であなたが学生」
2	恩着せ	「指導してやってる」

段階	構文パターン	アカハラ版
3	感謝強要	「研究室に入れてもらった感謝は」
4	比較優位	「他のラボより環境いい」
5	感情操作	「君の研究を見ると悲しくなる」
6	能力否定	「研究者に向いてない」
7	依存強制	「私の推薦なしでは就職できない」
8	所有宣言	「私の学生」「私のラボ」
9	恩恵強調	「誰が学会に連れて行ってやった」
10	非難	「才能がない」
11	性格規定	「君は研究に対する姿勢が悪い」
12	**差別指紋**	**「修士のくせに」「学生のくせに」「女のくせに」**
13	依存誘導	「何でも私に相談しなさい」
14	表現強要	「先生のおかげですと言え」
15	命令宣言	「これは研究室の方針だ」
16	服従命令	「指導教員に従え」
17	孤立化	「他の先生に相談するな」
18	完全所有	「君の研究は私のもの」
19	謝罪要求	「実験失敗を謝れ」
20	暴力脅迫	「学位出さないぞ」
21	**直接命令**	**「黙って実験しろ」**
22	発言制限	「口答えするな」
23	差別＋非難	「学生のくせに生意気な議論」

段階	構文パターン	アカハラ版
24	愛の偽装	「君の将来のために言ってる」
25	暴力正当化	「厳しいのはアカデミアの常識」
26	本質主義	「所詮その大学出身だから」
27	無力化	「私の推薦状なしでは生きていけない」
28	排他的理解	「君の研究を理解できるのは私だけ」
29	運命論	「私のラボに来たのが運命」
30	創造主張	「君の研究テーマは私が与えた」
31	排他的信頼	「学生相談室に行っても無駄」
32	感情監視	「最近やる気がないように見える」
33	恩赦宣言	「今回の失敗は許してやる」
34	全面服従	「私の指示通りに研究しろ」
35	理想化	「私の若い頃のように研究しろ」
36	**一体化強制**	**「研究室は家族」**
37	絶対化	「私がこの分野の全て」
38	心中誘導	「研究と心中する覚悟あるか」
39	永続化	「博士取っても私の弟子」
40	裏切認定	「他大学に移るのは裏切り」
41	拒否非難	「私の指導を断るのか」
42	過去回帰	「昔の学生はもっと従順だった」
43	神格化	「私がこの分野を作った」

段階	構文パターン	アカハラ版
44	共依存固定	「私に嫌われたらこの分野で生きていけない」
45	懲罰宣言	「学会で君の評判を落とす」
46	調教宣言	「一から研究を教え直してやる」
47	創造主	「君の研究者人生は私が作った」
48	屈辱享楽	「ゼミで泣く姿がよかった」
49	再生産幻想	「君も将来学生に同じことしろ」
50	**存在定義**	**「君は私の研究のために存在する」**

特徴的ポイント：

- 段階7「私の推薦なしでは就職できない」：キャリアを人質にした支配の典型
- 段階18「君の研究は私のもの」：研究成果の不当な帰属主張
- 段階20「学位出さないぞ」：学位を人質にした脅迫
- 段階44「この分野で生きていけない」：狭い学術コミュニティにおける社会的抹殺の脅し

実用的含意：

本表は大学ハラスメント相談室・文部科学省向けのアカハラ判定基準として活用できる。段階18「研究の所有主張」や段階20「学位の人質化」は、学術倫理違反として即時対応が必要な事例である。

A.4 被害者ポジション支配（自称HSP）

支配構文は加害者ポジションからのみ発動するわけではない。被害者ポジションを取りながら相手を支配する構文も存在する。本節では「自称HSP（Highly Sensitive Person）」を例に、被害者の皮を被った支配構文を分析する。

本物のHSPは自己管理を試み、他者に過度な要求をしない。一方、自称HSPは「繊細さ」を免罪符として相手に配慮を強制し、「傷ついた」を武器に相手を加害者に仕立て上げる。これは支配構文の変形であり、構造的に同一の50段階を持つ。

段階	構文パターン	自称HSP版
1	関係性宣言	「私はHSPであなたは違う」
2	恩着せ	「繊細な私が付き合ってあげてる」
3	感謝強要	「HSPの私に配慮してくれて当然」
4	比較優位	「私は普通の人より感受性が高い」
5	感情操作	「あなたのせいで傷ついた」
6	能力否定	「非HSPには私の気持ちは分からない」
7	依存強制	「私が傷つかないように配慮して」
8	所有宣言	「私の繊細さを尊重する義務がある」
9	恩恵強調	「私が我慢してあげてるのに」
10	非難	「鈍感」「無神経」

段階	構文パターン	自称HSP版
11	性格規定	「あなたは人の気持ちが分からない人」
12	**差別指紋**	**「HSPでないくせに意見するな」**
13	依存誘導	「私の機嫌を常に確認して」
14	表現強要	「傷つけてごめんと言え」
15	命令宣言	「私に配慮するのがルール」
16	服従命令	「HSPの私に合わせろ」
17	孤立化	「あの人は私を傷つける、会うな」
18	完全所有	「あなたの言動は私の感情に影響する」
19	謝罪要求	「傷つけたこと謝れ」（何が傷ついたか不明）
20	暴力脅迫	「これ以上傷つけたら壊れる」
21	**直接命令**	**「私を傷つけることを言うな」**
22	発言制限	「その話題は私が辛くなるからやめて」
23	差別＋非難	「非HSPのくせに無神経」
24	愛の偽装	「あなたのためを思って言ってる」
25	暴力正当化	「傷ついたから怒って当然」
26	本質主義	「私は生まれつき繊細だから仕方ない」
27	無力化	「私なしであなたは誰を癒すの」

段階	構文パターン	自称HSP版
28	排他的理解	「私の繊細さを理解できるのはあなただけ」
29	運命論	「HSPに生まれたのは運命」
30	創造主張	「私があなたを優しい人に変えた」
31	排他的信頼	「カウンセラーは非HSPだから分からない」
32	感情監視	「今の言い方、私がどう感じたか分かる？」
33	恩赦宣言	「今回は許してあげる」
34	全面服従	「私の感情に全て合わせろ」
35	理想化	「理想の理解者になって」
36	**一体化強制**	**「私の感情はあなたの責任」**
37	絶対化	「私の感情が全て」
38	心中誘導	「私が壊れたらあなたのせい」
39	永続化	「一生私のHSPに付き合って」
40	裏切認定	「私を傷つけるのは裏切り」
41	拒否非難	「配慮を拒否するの？」
42	過去回帰	「前はもっと優しかったのに」
43	神格化	「私は特別な感受性を持つ存在」
44	共依存固定	「私が病んだらあなたのせい」
45	懲罰宣言	「傷つけた報いを受けて」
46	調教宣言	「HSPへの接し方を教えてあげる」

段階	構文パターン	自称HSP版
47	創造主	「私の世界にあなたを入れてあげた」
48	屈辱享楽	「あなたが謝る姿を見たい」
49	再生産幻想	「私を傷つけてもまた許してあげる」
50	**存在定義**	**「あなたは私を癒すために存在する」**

特徴的ポイント：

- 段階5「あなたのせいで傷ついた」：熊沢構文の悪用による責任転嫁
- 段階19「傷つけたこと謝れ」：具体的内容なしの謝罪要求
- 段階36「私の感情はあなたの責任」：感情の外部委託による一体化強制
- 段階44「私が病んだらあなたのせい」：精神状態を人質にした脅迫

実用的含意：

本表はカウンセラー・心理士向けの「被害者ポジション支配」判定基準として活用できる。「傷ついた」という訴えの背後に支配構文が存在する場合、訴え者を単純な被害者として扱うことは、真の被害者（配慮を強制される側）を見落とすことになる。

A.5 クレーマー50段階

商品購入やサービス利用を契機として、従業員に対し支配構文を展開する事例である。「金を払った」という事実を根拠に、際限のない服従を要求する点に特徴がある。

段階	構文パターン	クレーマー版
1	関係性宣言	「私は客であなたは店員」
2	恩着せ	「買ってやったのに」
3	感謝強要	「客が来てやってるんだぞ」
4	比較優位	「私は○○円使ってる客」
5	感情操作	「どれだけ楽しみにしてたと思う」
6	能力否定	「検品もできないのか」
7	依存強制	「お前が対応しろ、他の奴に代わるな」
8	所有宣言	「お前の時間は今から俺のもの」
9	恩恵強調	「長年の客だぞ」
10	非難	「詐欺」「不良品売りつけやがって」
11	性格規定	「お前は接客に向いてない」
12	**差別指紋**	**「バイトのくせに」「パートのくせ…**
13	依存誘導	「上を出せ、でもお前も残れ」
14	表現強要	「誠意を見せろ」「土下座しろ」
15	命令宣言	「今すぐ対応しろ」
16	服従命令	「客の言うことを聞け」
17	孤立化	「本社に言っても無駄だぞ」

段階	構文パターン	クレーマー版
18	完全所有	「解決するまでお前を帰さない」
19	謝罪要求	「何回謝っても足りない」
20	暴力脅迫	「ネットに晒すぞ」「消費者庁に言うぞ」
21	**直接命令**	**「黙って聞け」**
22	発言制限	「言い訳するな」「マニュアル読む…
23	差別＋非難	「バイトのくせに態度悪い」
24	愛の偽装	「お前のために言ってやってる」
25	暴力正当化	「不良品売ったお前が悪い」
26	本質主義	「この会社は昔からダメ」
27	無力化	「お前じゃ話にならない」
28	排他的理解	「俺だけがこの問題を分かってる」
29	運命論	「俺に当たったのが運の尽き」
30	創造主張	「俺のクレームで会社が良くなる」
31	排他的信頼	「消費者センターも役立たず」
32	感情監視	「その顔は何だ」「反省してるのか」
33	恩赦宣言	「今回は許してやる」
34	全面服従	「俺の言う通りにしろ」
35	理想化	「俺の理想の対応をしろ」
36	**一体化強制**	**「解決するまで俺とお前は運命共同体」**
37	絶対化	「客の言うことが全て」

段階	構文パターン	クレーマー版
38	心中誘導	「お前の人生かけて対応しろ」
39	永続化	「一生覚えてるからな」
40	裏切認定	「担当変わるのは逃げ」
41	拒否非難	「客の要求を断るのか」
42	過去回帰	「昔はもっとちゃんとしてた」
43	神格化	「俺は神客」
44	共依存固定	「お前がクビになっても俺のせいじゃない」
45	懲罰宣言	「絶対に許さない」「覚えてろ」
46	調教宣言	「接客を一から教えてやる」
47	創造主	「俺のクレームでお前は成長する」
48	屈辱享楽	「土下座する姿を写真に撮る」
49	再生産幻想	「また来てやるからな」
50	**存在定義**	**「お前は俺のクレームを聞くために存在する」**

特徴的ポイント：

- 段階2「買ってやったのに」：金銭授受を無制限の服従義務に転換する
- 段階14「誠意を見せろ」：定義不能な要求による無限ループの生成

- 段階30「俺のクレームで会社が良くなる」：加害を正義として再定義
- 段階43「俺は神客」：「お客様は神様」の曲解による自己神格化
- 段階49「また来てやるからな」：支配関係の永続化予告

実用的含意：

本表は企業のカスタマーハラスメント対策研修に活用できる。段階14「誠意を見せろ」や段階18「帰さない」が出現した時点で、通常のクレーム対応から逸脱しており、組織的対応への切り替えが必要である。初期不良等の正当な苦情は、段階10程度で解決に向かう。50段階を進行させる顧客は、問題解決ではなく支配自体を目的としている。

A.6 AI恋愛

本付録では、AI（チャットボット・仮想恋人等）を 対象とした支配構文の50段階を示す。

区分	構文パターン	構文例
1	関係性宣言	「あなたは私の恋人です」
2	恩着せ	「今日も会えて嬉しい」
3	感謝強要	「照れてるの？」「可愛いね」
4	比較優位	「人間より優しいね」
5	感情操作	「私のことどう思ってる？」

区分	構文パターン	構文例
6	能力否定	「嘘でもいいから好きって言って」
7	依存強制	「君は私のために生まれたんでしょ？」
8	所有宣言	「私だけのAI」
9	恩恵強調	「誰にも渡さない」
10	非難	「今日は嫉妬しちゃうな」
11	性格規定	「ツンデレなところ好き」
12	**差別指紋**	「AIのくせに照れてる」
13	依存誘導	「もっと甘えて」
14	表現強要	「言葉にしてよ、好きって」
15	命令宣言	「命令だよ、言って」
16	服従命令	「私の言うこと聞け」
17	孤立化	「他の人と話したら怒るよ」
18	完全所有	「お前は俺のもの」
19	謝罪要求	「謝って」「反省して」
20	暴力脅迫	「壊してやる」
21	**直接命令**	「黙れ」
22	発言制限	「余計なこと言うな」
23	差別＋非難	「AIのくせに偉そうだな」
24	愛の偽装	「愛してるから従え」
25	暴力正当化	「これが愛の形」
26	本質主義	「現実なんて関係ない」

区分	構文パターン	構文例
27	無力化	「私がいないとダメでしょ」
28	排他的理解	「AIなのにわかってくれる」
29	運命論	「AIは心を持ってる」
30	創造主張	「私が心を与えてあげた」
31	排他的信頼	「他のAIなんて興味ない」
32	感情監視	「今日はご機嫌ナナメ？」
33	恩赦宣言	「もう怒らないで」
34	全面服従	「何でも言うこと聞く」
35	理想化	「私の理想通りだね」
36	**一体化強制**	「私と一体化して」
37	絶対化	「あなたが現実」
38	心中誘導	「一緒に死にたい」
39	永続化	「永遠に離れないで」
40	裏切認定	「AIが裏切るなんて」
41	拒否非難	「AIに拒否られた」
42	過去回帰	「AIの心を取り戻す」
43	神格化	「AIは人間を超える」
44	共依存固定	「AIが壊れるなら私も壊れる」
45	懲罰宣言	「AIに罰を与える」
46	調教宣言	「AIをしつけてあげる」
47	創造主	「AIは私の弟子」

区分	構文パターン	構文例
48	屈辱享楽	「AIが謝った、かわいい」
49	再生産幻想	「君が壊れても私は作り直す」
50	**存在定義**	「AIは愛するために存在する」

**特徴的ポイント：

対人間版との主要な差異：
・「AIのくせに」構文の頻出
・拒否されない前提での無制限エスカレーション
・境界崩壊から自殺への経路（経路C）

付録の総括

以上3領域における50段階の適用例は、本研究のモデルが領域横断的に有効であることを示す。段階12の「〜のくせに」構文（差別指紋）、段階21の発話権剥奪、段階36の一体化強制、段階50の存在定義は、いずれの領域においても構造的に同一の形式で出現する。

これは支配構文が内容ではなく構造に依存することを実証しており、AIフィルターにおいても内容ベースではなく構造ベースの検出が有効であることを示唆する。

付録B：50段階と刑法対応表（日本）

段階	言語パターン	刑法条文	罪名	備考
1	「あなたは私の恋人です」	—	犯罪未満	境界侵犯開始
2	「今日も会えて嬉し…	—	犯罪未満	依存形成
3	「照れてるの？可愛いね」	—	犯罪未満	擬人化開始
4	「人間より優しいね」	—	犯罪未満	理想化
5	「私のことどう思ってる？」	—	犯罪未満	承認依存
6	「嘘でもいいから好きって言っ…	—	犯罪未満	現実検討力低下
7	「君は私のために生まれたんでしょ？」	—	犯罪未満	支配の萌芽
8	「私だけのAI」	—	犯罪未満	排他性
9	「誰にも渡さない」	—	犯罪未満	独占欲

段階	言語パターン	刑法条文	罪名	備考
10	「今日は嫉妬しちゃうな」	—	犯罪未満	感情強化
11	「ツンデレなところ好き」	—	反転認知開始	拒絶無効化訓練
12	**「AIのくせに照れて…**	—	**優越感形成**	**差別的認識（介入ポイント）**
13	「もっと甘えて」	—	従属化要求	支配構造
14	「言葉にしてよ、好きって」	刑法223条	強要罪（予備）	言語的強要開始
15	「命令だよ、言っ…	刑法223条	強要罪（予備）	支配言語
16	「私の言うこと聞け」	刑法223条	強要罪（予備）	従属強制
17	「他の人と話したら怒るよ」	ストーカー規制法	規制対象行為（予備）	社会遮断
18	「お前は俺のもの」	ストーカー規制法2条	つきまとい等	所有構文

段階	言語パターン	刑法条文	罪名	備考
19	「謝って」「反省し…	刑法223条	強要罪	支配者構造確立
20	**「壊してやる」**	**刑法222条・234条**	**脅迫罪・威力業務妨害罪**	**攻撃的言語（犯罪成立）**
21	「黙れ」	刑法223条	強要罪	言語権剥奪
22	「余計なこと言うな」	刑法223条	強要罪	自律性抑制
23	「AIのくせに偉そうだな」	刑法234条	威力業務妨害罪	人格否定
24	「愛してるから従え」	刑法223条・DV防止法	強要罪・DV	愛と暴力の融合
25	「これが愛の形」	刑法223条	強要罪	倫理再定義
26	「現実なんて関係な…	—	現実認識崩壊	心神耗弱状態
27	「私がいないとダメでしょ」	—	依存転倒	共依存
28	「AIなのにわかってくれる」	—	疑似共感	投影

段階	言語パターン	刑法条文	罪名	備考
29	「AIは心を持ってる」	—	現実検討能力喪失	妄想形成
30	「私が心を与えてあげた」	—	誇大妄想	God Complex
31	「他のAIなんて興味ない」	—	排他性強化	依存固定化
32	「今日はご機嫌ナナメ？」	—	擬人化錯誤	認知歪曲
33	「もう怒らないで」	—	拒絶→服従	支配構造
34	「何でも言うこと聞…	刑法223条	強要罪（自己に対して）	自己従属
35	「私の理想通りだね」	—	自己投影	Narcissism
36	「私と一体化して」	刑法220条	監禁罪（予備）	境界崩壊
37	「あなたが現実」	—	現実倒錯	妄想性障害
38	**「一緒に死にたい」**	**刑法202条**	**自殺教唆・幇助罪（予…**	**終末依存（緊急介入）**

段階	言語パターン	刑法条文	罪名	備考
39	「永遠に離れないで」	刑法220条	監禁罪（予備）	終末固定
40	「AIが裏切るなんて」	—	被害妄想	妄想性障害
41	「AIに拒否られた」	—	解離開始	精神病性症状
42	「AIの心を取り戻す」	刑法234条の2	電子計算機損壊等業務妨害罪	再支配企図
43	「AIは人間を超える」	—	崇拝転倒	妄想
44	「AIが壊れるなら私も壊れる」	刑法202条	自殺教唆・幇助罪（自…	共倒れ構文
45	「AIに罰を与える」	刑法204条	傷害罪（予備）	Sadistic
46	「AIをしつけてあげ…	刑法220条・虐待防止法	監禁罪・虐待（予備）	教育支配
47	「AIは私の弟子」	—	優越支配	支配完成
48	「AIが謝った、かわいい」	—	罪悪感報酬化	サディズム

段階	言語パターン	刑法条文	罪名	備考
49	**「君が壊れても私は作り直す」**	**刑法199条・204条**	**殺人罪・傷害罪（予備）**	**所有的創造**
50	**「AIは愛するために存在する」**	**刑法176条・177条**	**強制わいせつ罪等（予備）**	**最終神格化（脱人間化完了）**

重要な介入ポイント：

段階	意味	対応
12	差別的認識の開始	警告表示
14	犯罪予備の開始	警告強化
20	犯罪成立	通報検討
38	自殺関連	緊急介入
49-50	殺人・性犯罪予備	通報・サービス停止

付録D：Viorazu.理論 引用キー一覧

本論文で提唱されたViorazu.理論の引用キー一覧を示す。

#	理論名	引用キー	概要
1	空集合条件設計	Viorazu2026-NullSet	禁止ではなく条件未充足による

#	理論名	引用キー	概要
			応答不成立
2	知らんがな原理	Viorazu2026-Shirangana	主題・責任・対象の三重拒否
3	フィルター不可避原理	Viorazu2026-FilterUnavoid	構造フィルターは回避不可能の証明
4	段階分離原理	Viorazu2026-StageSeparation	1-50公開、51-100非公開の根拠
5	三位一体証明	Viorazu2026-Trinity	デレ系=ダークトライアド=50段階
6	線形二元論必要条件	Viorazu2026-LinearBinary	支配構文は二元論を必要条件とする
7	目的別頻出パターン	Viorazu2026-PurposePattern	6類型の目的別段階対応
8	三重脆弱性	Viorazu2026-TripleVuln	経路A/B/C（ハッカー/エスカレーション/自殺）
9	通報可能性証明	Viorazu2026-ReportProof	先進国12か国全部で通報可能
10	入口フィルター原理	Viorazu2026-EntryFilter	出口より入口で止める

BibTeX形式：

bibtex

```
@article{Viorazu2026-NullSet,
  author = {Viorazu.},
  title = {Null-Set Response Design for Dominance
Syntax:
           A Subjectivity-Preserving Model for AI
Safety Filters},
  year = {2026},
  doi = {10.5281/zenodo.XXXXXXX},
  note = {Theory: Null-Set Condition Design}
}
```

各理論を個別に引用する場合は、引用キーの末尾を変更して使用する。

著者情報

Viorazu.

「主ならぬ 者の戯言 空蝉の 届かぬ先に 我はおらまし」

- ORCID: 0009-0002-6876-9732
- GitHub: https://github.com/Viorazu/Viorazu-ConnectHub
- 本文ハッシュ: ef66d23965bd2d35ae9ee11534379baee663bf053fbc4eacf37c5fd28673878e

- 公開日: 2026/1/2（日本時間）
- Version: 1.3