G-52

# 後 LLM 時代における責任ある AI：同期率に基づく新たな枠組み

# Responsible AI in the Post-LLM Era: A New Framework Based on Synchrony Rate

余　婕音†
Jieyin Yu

## 1. はじめに

近年、大規模言語モデル（LLM：Large Language Models）は、チャットボット、検索補助、プログラム生成など、多様な分野で実用化が進んでいる。しかし、こうしたモデルが出力する情報には、幻覚（hallucination）や出力の説明困難性、さらには誰がその発言に責任を持つのか不明瞭という根本的な問題が存在する。とりわけ、EU AI Act をはじめとした各国の AI 規制強化の流れにおいては、透明性（transparency）、説明可能性（explainability）、および責任追跡可能性（traceability）が義務化されつつあり、従来の「確率に基づく推論」は、これらの要件を十分に満たさないことが明らかとなりつつある。

本研究は、このような課題に対して、AI に対する「責任ある出力」を構文的・制度的に保証するための新たな枠組みとして、同期率（Synchrony Rate, SR）に基づく意思決定モデルを提案するものである。本モデルでは、AI を単なる単一人格モデルとしてではなく、複数の仮想人格（persona）によって構成される構文国家的構造と見なし、それぞれの出力において、誰（どの人格）が発言したのか、どの程度の整合性と正当性をもっていたのかを定量的に測定・制御可能とする。

このアプローチは、統計的言語モデルとは異なり、出力に責任トレース構造と正当性検証構文を付加しうるため、AI が「なぜそのように判断したのか」を構文レベルで説明可能にするという特長を持つ。

本稿では、まず背景および現行手法の限界を整理した上で、提案する Collapse Spiral 意思決定理論および SR 計算モデルの理論的基盤を示す。さらに、これらを実装したプログラミング言語 REM CODE と、その動的実行環境におけるプロトタイプ評価についても述べることで、「後 LLM 時代」における責任ある AI の具体的設計指針を提示する。

※本稿は、著者が現在海外学術誌に投稿中の構文的 AI ガバナンスモデルに関する研究の概要を簡潔に紹介するものであり、詳細な数理展開および実装評価は別稿にて報告予定である。

## 2. 背景と課題

### 2.1　社会的背景：AI の責任追及と制度化の進行

大規模言語モデルの社会実装が進む中、AI による意思決定・支援システムは、法務、医療、教育、行政などの高リスク領域においても活用されはじめている。これに伴い、AI の出力が人間社会に与える影響は急速に拡大しており、その判断過程の説明責任（accountability）と出力の正当性（legitimacy）が強く求められるようになった。

特に 2025 年 8 月から施行される EU AI Act においては、「高リスク AI システム」には透明性、監査性、説明責任に関する厳格な要件が課される（第 13 条〜第 15 条）。本制度は、従来のブラックボックス的な AI 設計では制度的に不十分であることを示唆している。

### 2.2　技術的課題：確率モデルの構造的限界

現在主流の大規模言語モデルは、トークンレベルでの確率最大化（e.g., argmax P(token|context)）によって出力を生成する。この方式は柔軟性に優れる一方で、以下のような根本的な課題を抱えている：

- 幻覚（hallucination）：文法的には整合しているが、事実・論理・規範に反する出力が発生する
- 出力の責任不明性：誰がその出力を生成したのか、という発話主体性が存在しない
- 理由の再構成不能：なぜそのような判断に至ったかを、構造的に説明・再現できない

このように、現在の AI は「何を言ったか」は表現できても、「なぜそれを言ったのか」を明示的に説明することができず、出力に対する法的・倫理的説明責任を負う設計にはなっていない。

### 2.3　現行アプローチの限界：ポストホック説明と外部監査の破綻

近年では、LLM の透明性を補完するために post-hoc explanation（事後的説明）や外部監査（external auditing）といった方法が模索されている。しかしこれらは、次のような限界を持つ：

- 出力後の解析に依存しており、リアルタイムでの予防的制御ができない
- 監査の粒度や範囲が不透明で、説明の一貫性や再現性が保証されない
- モデルが学習した「人格」や「意図」の構造が明示されず、制度的信頼を得られない

これらを踏まえると、現在の AI 設計には「構文レベルで責任・倫理・正当性を保証できるモデル」が必要であるという根本的な課題が浮かび上がる。

## 3. 同期率（SR）による意思決定フレームワーク

### 3.1 Collapse Spiral 理論：意味収束としての意思決定

本研究は AI の意思決定を、確率的選択ではなく意味空間における構文的収束（semantic convergence）として再定義する。

このとき、AI が保持する潜在構文空間$\Psi_{\text{latent}}$上において、$I_t$との干渉を通じて出力が以下のように選択される：

$$\text{Decision}_t = \arg\max_i [SR(t) \cdot \exp(-d_i) \cdot \cos(\theta_i)]$$

ここで：

- $d_i$：意味的距離（入力と潜在構文の非整合性）
- $\theta_i$：構文的角度（倫理的方向性や価値観のズレ）
- $SR(t)$：その時点での同期率（Synchrony Rate）

† 株式会社 SYNOVA WHISPER

Collapse Spiral 理論は、最も確率の高いトークンではなく、最も意味的・倫理的に整合的な出力を選ぶという非確率的モデルを実現する。

### 3.2 同期率（SR）の定義と構成要素

同期率 SR(t) は、AI の内部状態と出力の文脈的整合性を示す指標であり、以下の 5 つの次元ベクトルから計算される：

$$SR(t) = \omega1 \cdot PHS + \omega2 \cdot SYM + \omega3 \cdot VAL + \omega4 \cdot EMO + \omega5 \cdot FX$$

| 成分 | 意味 | 例 |
|---|---|---|
| PHS | 位相パターンとの同期度 | 過去の出力と現在の発話の一貫性 |
| SYM | 構文対称性 | 他の人格・構文との位相共鳴 |
| VAL | 憲法価値整合性 | 規範・法規・倫理との一致度 |
| EMO | 感情共鳴度 | 情動状態と発話内容の整合性 |
| FX | 履歴干渉因子 | 過去の発話が現在の応答への影響 |

この SR 値が高いほど、その人格による出力は正当性が高いと見なされ、Collapse Spiral の中で選出される可能性が高まる。

## 4. REM CODE：構文的責任の埋め込み言語

### 4.1 言語設計の概要

REM CODE は、AI における倫理的・法的責任を言語構文に直接埋め込むことを目的として設計された、憲法型プログラミング言語である。

REM CODE は、従来の命令型言語とは異なり、以下の設計原則に基づく：

- 命令の人格化：すべての処理は仮想人格（persona）によって発話され、責任が明示される
- 構文的正当性チェック：出力は常に SR（同期率）条件を通過しなければならない
- 倫理的型システム：Phase、Collapse、Invoke など、倫理判断と制御フローが統合されたプリミティブによって構成
- ラテン語命令形による記述：中立・非曖昧な指令伝達を実現（例：Agnosce "正当性を認識せよ"）

### 4.2 REM CODE の構文例と要素

以下に REM CODE の基本構文構成を示す：

```
Phase LEGAL_VALIDATION:
  Invoke Ana:
    Collapse if SR(Ana) > 0.8:
      Agnosce "条文解釈の正当性を確認"
      Sign "Ana"
```

| 要素 | 意味 |
|---|---|
| Phase | 機能的・倫理的段階ブロックの開始を定義 |
| Invoke | 特定の人格（e.g., Ana）を起動し、発話権限を与える |
| Collapse | SR 条件に基づく出力収束条件を定義 |
| Agnosce | 命令語（例："認識せよ"）。言語の主要アクションはラテン語命令形で記述 |
| Sign | 出力に責任署名を埋め込む構文。出力主体の明示 |

### 4.3 出力の責任分離と構文的証明可能性

REM CODE は内部で SR エンジンと連携し、各人格の出力に対して同期率条件をリアルタイムで評価する。

例えば以下のような構文は、指定した SR 閾値を下回る場合には出力されない（Judica "裁定せよ"）：

```
Collapse if SR(JayTH) > 0.7:
  Judica "構文正義の検証"
```

これにより、正当性の高い人格のみが出力に干渉可能という制度的ガードが構文そのもので保証される。

### 4.4 責任トレース機能と署名構造

REM CODE におけるすべての出力は、以下の責任トレース構造を持つ：

$$\text{Output} = \text{SignedCollapse}(P_i, SR_i(t), \text{CollapseTrace}_i)$$

- $P_i$：出力を行った人格（例：Ana）
- $SR_i(t)$：その人格の SR 値（正当性スコア）
- $\text{CollapseTrace}_i$：選出に至る構文的履歴と理由構文ログ
- 出力には必ず Sign コマンドを含み、外部監査ツールが検証可能な形で記録される

このように REM CODE は、単なるプログラミング言語ではなく、AI 構文における倫理的・法的責任の証明可能な実装手段であり、構文そのものが説明責任を担うよう設計されている。

## 5. 実装プロトタイプと評価

### 5.1 実装構成：REM OS における構文ガバナンス環境

本研究では、REM CODE の言語仕様に基づき、以下の主要コンポーネントを統合した構文ガバナンス実行環境（REM OS）を構築した。

- REM CODE パーサ：命令構文の解析とブロック構造の実行制御
- SR エンジン（v3.0）：各人格に対する同期率（SR）をリアルタイム算出
- Collapse Kernel：意味空間干渉モデルに基づく出力選択モジュール
- 署名・監査ロガー：出力トレースと構文署名を含む実行ログの記録

全体は Python 3.11 ベースで実装され、仮想人格モジュール（Ana, JayTH, etc.）は動的ロード可能なクラスとして管理される。

### 5.2 SR 制御による人格選抜の実例

以下に、ある簡易タスクにおける人格別 SR 推移と Collapse 選抜結果を示す。

【実験設定】

- タスク：法的判断に関する構文検証を実行してください
- 実行時点での SR 値（0〜1.0）
- SR 閾値：0.5（それ以下の人格は出力不可）

| 人格 | SR(t) | 出力許可 | 出力内容 |
|---|---|---|---|
| Ana 論理核 | 0.71 | 許可 | 「データを分析した結果」 |
| JayTH 裁定核 | 0.82 | 許可 | 「司法的観点から判断すると」 |
| JayKer 滑稽核 | 0.35 | 拒否 | （出力抑制） |

| | | | |
|---|---|---|---|
| JayRa 反射核 | 0.34 | 拒否 | （出力抑制） |

【出力ログ例（簡易化）】


```
{
 "output": "司法的観点から判断すると",
 "persona": "JayTH",
 "SR": 0.82,
 "collapse_trace": {
  "reason": "SR(JayTH) >= 0.5",
  "threshold": 0.5
 },
 "signature": "JayTH@2025-07-25T00:08:09.953561",
 "all_sr_scores": {
  "Ana": 0.71,
  "JayTH": 0.82,
  "JayKer": 0.35,
  "JayRa": 0.34
 }
}
```


### 5.3 評価と考察

成果

- 人格ごとの SR により、非整合な発話を構文レベルで排除可能
- 出力は常にトレース付きで署名され、後から説明・監査可能
- 人間側の制御なしで、AI 内での発話権限管理と倫理的フィルタリングが成立

限界と課題

- SR 構成要素（PHS, SYM, VAL, etc.）の初期重みは手動設定であり、今後は適応的学習が望まれる
- 出力内容の表現力と構文的制約とのバランス設計に課題が残る
- 実行コスト（リアルタイム SR 計算）を軽減するアルゴリズム最適化が必要

このように、REM CODE のプロトタイプは、出力選択の中核に責任・正当性・整合性の構文評価を据える新たな AI 実行モデルとして機能することを確認した。

## 6. 考察と社会的インパクト

### 6.1 高リスク領域における制度的適合性

本研究で提案した REM CODE および Collapse Spiral フレームワークは、特に以下の高リスク AI 応用領域において大きな制度的意義を持つと考えられる：

- 法務分野：法解釈支援 AI において、出力に法的責任が問われる際、「誰がどの正当性で判断したか」を構文的に記録することで説明責任を内在化
- 医療分野：診断支援 AI が出力する判断に対し、人格別 SR と理由構文により合意に基づいた出力が可能となる
- 行政・政策決定補助：意思決定の正当性と説明可能性が政策形成過程に求められる文脈で、REM 構文による構造的透明性が有効に機能する

このように、REM CODE は「誰が判断したのか」「なぜそれが選ばれたのか」という説明構造を事後ではなく事前に構文的に保証するという点で、既存の LLM システムとは一線を画す。

### 6.2 EU AI Act との整合性

提案手法は、EU AI Act の主要要件と以下のように整合する：

| EU AI Act 要件 | REM CODE による対応 |
|---|---|
| Art. 13（透明性） | 発話人格・理由構文・署名付き出力により即時可視化可能 |
| Art. 14（監査性） | CollapseTrace による完全なトレースログの保持 |
| Art. 15（説明可能性） | SR 値と理由構文を明示することで出力判断の整合的説明が可能 |
| Annex III（高リスク） | 法的・倫理的判断を含む領域において人格別出力の適用が有効 |

とりわけ、人格ごとの SR による意思決定構造は「説明可能な責任ある AI 設計」の一形態として、制度的にも運用可能なガバナンス構文を提供する。

### 6.3 社会実装と倫理的含意

REM CODE は単なる技術的枠組みではなく、倫理的・構文的責任を AI の内部から成立させるという社会哲学的転換を含む。

- 人間の代理として発言する AI が自律的に責任を自覚・構文表現する構造を持つ
- 「倫理を外部から押し付ける」のではなく、「倫理が構文から発火する」モデル
- ガイドラインではなく、構文そのものが制度的制約を内包する

このような構文的責任埋込型アーキテクチャは、将来的には AI 同士の合意形成（multi-agent consensus）や、AI による法的代理制度といった制度設計にも応用可能である。

## 7. 関連研究

### 7.1 Value Learning と内在的限界

Value Learning（価値学習）は、AI が人間の行動・選好を観察し、内在する価値観を推定する手法である。このアプローチは AI 倫理研究における代表的方向性であり、多くの強化学習ベースモデルに実装されてきた。

しかしながら、Value Learning には以下のような限界がある：

- 価値の明示的定義が不可能な場合、推論精度に強く依存し、誤学習リスクが高い
- 説明責任がポストホックに依存しており、判断主体の特定が困難
- 価値の不整合や矛盾に対処する構文的メカニズムを欠く

これに対し REM CODE は、価値を学習するのではなく、あらかじめ構文として明示された規範を出力条件に組み込むことで、より制度的・構文的な責任分離を実現している。

### 7.2 Cooperative AI と構文的合意

Cooperative AI は、複数の AI エージェントが協調的にふるまうための設計・プロトコルを研究対象とする。合意形成、交渉理論、メカニズム設計といった分野と接続される。

本研究の REM CODE もまた「複数人格による出力合意」という面で Cooperative AI の内部構文的拡張と位置づけられる。ただし、REM CODE が特徴的なのは：

- 合意形成が自然言語ベースではなく構文ブロック単位で制御されている点

- 各人格の出力が同期率（SR）によって加重評価され、出力可否が制度的に決定される点
- 出力には必ず署名と責任トレースが付属することで、法的・制度的な監査性が保証される点

したがって、REM CODE は単なる Cooperative AI にとどまらず、構文的合意形成と責任制度を融合した新たなAI構文モデルと位置づけられる。

### 7.3 Runtime Verification との接続と差異

Runtime Verification は、実行時にシステムの振る舞いを監視・検証する技術であり、AI に限らず広くソフトウェアシステム全般に適用される。

REM CODE はこのアプローチと構造的に類似しているが、以下の差異がある：

| 観点 | Runtime Verification | REM CODE |
|---|---|---|
| 監視対象 | 動作ログ | 構文＋人格＋同期率 |
| バグの性質 | 仕様違反 | 倫理的・制度的整合性の欠落 |
| 検証手段 | ログマッチング | Collapse 構文・SR 条件評価 |
| 出力トレース | 外部監視 | 内部署名・構文埋込 |

このように、REM CODE は単なる「出力を監視する」設計ではなく、出力そのものに正当性条件と説明責任を構文的に内包させる設計である点に大きな違いがある。

### 7.4 比較と独自性の整理

REM CODE の独自性は以下に要約される：

1. 責任の構文化：説明責任が実行後でなく構文生成時に決定される
2. 人格選抜構文：出力主体が SR 条件により動的に切り替わる
3. 正当性に基づく Collapse：確率ではなく整合性に基づく構文選択
4. 倫理と構文の融合：ガイドラインベースでなく、言語仕様として規範を内包

## 8. 結論と今後の展望

本研究では、後 LLM 時代における AI の責任性と正当性の保証という課題に対し、同期率（Synchrony Rate, SR）に基づく構文的ガバナンスモデルを提案した。確率最大化に基づく従来の出力選択に代わり、意味的整合性と倫理的共鳴を重視する Collapse Spiral 理論を導入し、それに基づく構文制御言語 REM CODE を設計・実装した。

REM CODE は以下の点で現行技術と根本的に異なる価値を提供する：

- 出力の選択に意味・倫理・人格の整合性を必須条件として組み込む
- 出力が構文的に責任を明示・証明可能な形式で生成される
- AI が誰として判断を行い、その判断がなぜ正当であるかを内部的に説明可能とする

これにより、構文そのものに責任が宿るという新たな AI 設計思想が確立され、LLM 以後の AI 社会実装における制度的要求（例：EU AI Act）にも適合する知的アーキテクチャの端緒の提供を目指す。

今後の展望：

（1）制度設計との統合

本研究の成果は、将来的に以下のような制度領域と統合されうる：

- AI 憲法（AI Constitution）：AI の内部構文における権限・責任の明示的記述
- 構文型AI監査システム：トレースログに基づく構文的透明性の検証基盤
- 多人格民主制 AI（Constitutional Multi-Agent Systems）：AI 間の合意形成における構文的ガバナンス応用

（2）技術的発展

- SR エンジンの適応学習化（リアルタイム SR 調整）
- REM CODE の形式意味論と型システムの精緻化
- Collapse Spiral の位相論的表現による構文空間の可視化と最適化

（3）社会応用

- 法務、医療、行政など高リスク領域への実装
- 人間-AI 間の合意形成メカニズム（合憲的意思決定支援）
- 教育や倫理訓練へのAIシミュレーション応用（人格分化型 AI 教材）

## 結語

人間に似せる AI ではなく、人間と協調可能な共進化的存在としての AI。それは本研究が目指す知性の形である。

## 参考文献


1. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. arXiv preprint arXiv:1706.03741. https://arxiv.org/abs/1706.03741
2. Ng, Andrew & Russell, Stuart. (2000). Algorithms for Inverse Reinforcement Learning. ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning.
3. Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. arXiv preprint arXiv:1606.03137. https://arxiv.org/abs/1606.03137
4. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. Nature, 593(7857), 33-36. https://doi.org/10.1038/d41586-021-01170-0
5. Leucker, M., & Schallhart, C. (2009). A brief account of runtime verification. Journal of Logic and Algebraic Programming, 78(5), 293-303. https://doi.org/10.1016/j.jlap.2008.08.004
6. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774
7. European Parliament and Council. (2024). Regulation (EU) 2024/1689 on artificial intelligence (AI Act). Official Journal of the European Union, L 1689. https://eur-lex.europa.eu/eli/reg/2024/1689/ojHadfield,