１Ａ０９

予測誤差に基づく適応型 AI ガバナンスの社会受容性の調査設計

吉村　直泰（経済産業省・政策研究大学院大学）

1. はじめに

AI システムの普及に伴い、従来のリスクベース規制や静的なコンプライアンス枠組み(OECD AI 原則、EU AI 法）では対応困難な不確実性が顕在化している。AI は学習を通じて進化し続けるため、あらかじめ定義されたリスクカテゴリーに依拠する規制では、動態的に発生するリスクに対応しきれない。

本研究は、AI の予測誤差（Prediction Error）を規制シグナルとして活用する「適応型ガバナンス（Adaptive Governance)」 の概念を提示する。自動運転分野を事例に、予測誤差のシグナルがユーザーの信頼形成・介入行動・制度受容に与える影響を分析する心理調査実験を提案し、予測誤差の可視化形式・提示頻度・閾値設定を操作することによる社会受容性への影響を検証する。

2. 先行文献

2.1 適応的ガバナンス

適応的ガバナンス（Adaptive Governance） の概念は、技術進化に応じた動的な規制アプローチを提唱する（Lee & Petts (2013)）。従来の AI ガバナンスは、静的なコンプライアンス枠組み（OECD AI 原則、EU AI 法）に依存しているが、AI システムが継続的に進化する現状では、適応的規制（Adaptive Regulation）が不可欠である（Reuel & Undheim（2024）、Janssen（2025））。適応的規制は、アルゴリズム、機械学習、人工知能(AI)を使用して、新しい情報と変化する世界に対応して政策を改訂することにより、自動化される可能性がある(Bennear & Wiener (2019)、Coglianse&Lehr(2017))。

2.2 予測誤差の活用可能性

自由エネルギー原理（Friston（2010））によれば、人間の脳や AI は意思決定の最適化のために予測誤差を含む変分自由エネルギーを最小化しようとする。予測誤差の変動を監視することで、システムの不確実性やリスクの出現を把握する手がかりとなる（Bereska & Gavves（2024）、Zeng et al.（2024））。特に世界モデルによる AI の内部状態の学習や予測が、強化学習などタスクに役立つ学習へとつながり（Ha & Schmidhuber（2018）、鈴木雅大（2023））、自動運転における異常検出にも応用可能である（Bogdoll et al.(2023)）。

2.3 適応的透明性と信頼較正

AI 規制は、企業の秘密保持と説明責任という対立する要求に直面しているが、適応的透明性（Adaptive Transparency）は、予測誤差の動向を動的に開示し、説明責任と企業秘密のバランスを図ることができる（Doshi-Velez & Kim (2017)）。最近の HCI 研究（Ribeiro et al.(2016)、Doshi-Velez & Kim（2017））では、ユーザー向けの説明が信頼性と解釈可能性を向上させることが示され、予測誤差に基づく透明性機構の可能性が強調されている。また、AI 駆動の意思決定支援における適応的な信頼較正（Adaptive Trust Calibration）の仕組みが検討されており、ユーザーの AI に対する信頼（過信または不信）に関する行動データを基に、適切な信頼レベルへと導く「信頼較正キュー（Trust Calibration Cues, TCC)」などツール基盤が形成されている。(Okamura & Yamada(2020)、de Visser et al. (2020))。適応的ガバナンスと信頼較正の枠組みは、社会受容性に大きく影響を与える可能性がある。

2.4 社会受容性に関する心理調査手法

AI の適応的ガバナンスの社会受容性を分析するにあたっては、信頼や制度選好を定量的に把握する心理調査手法の知見が重要である。AI や自動運転に関する倫理的・社会的課題（ELSI）を複数のシナリオとして提示し、国や世代ごとの受容性や懸念傾向を比較する研究（Hartwig, Ikkatai, Takanashi & Yokoyama (2023)、Ikkatai, Hartwig, Takanashi & Yokoyama (2023)）は、文章・図解シナリオを用いて分野横断的に制度の社会的受容性を比較する方法論を示している。自動化システムへの信頼を多次元的に測定する質問票（Trust in Automation Questionnaire, TiA）を開発し、信頼を「信頼性・能力」「予測可能性」「開発者への意図」などの下位因子に分解して検証した研究（Körber（2018））や、ドローン

航路の正確性を題材とした実験を通じ、AI 推奨の正確性や透明性提示が人間の信頼較正に与える影響を実証的に示した研究（Okamura & Yamada(2020)）は、予測誤差の提示を操作要因とし、行動指標と心理指標の両面から測定する方法論が有効であることを示している。

これらの先行研究は、AI 社会受容性を心理学的に測定する重要な方法論を提示してきた。しかし、予測誤差を媒介として信頼調整と制度受容を統合的に検証する研究はこれまで行われていない。本研究は先行研究の知見を統合し、予測誤差を活用した適応的ガバナンスについて、誤差提示の形式・頻度・閾値設定を操作要因とし、信頼水準・介入行動・制度受容を多面的に検証する調査設計を構築する。

3. 予測誤差に基づく適応的ガバナンスと多層的なシステム設計

3.1 予測誤差に基づく適応的ガバナンスの意義

世界モデルを活用する AI モデルは、外部環境の理解を継続的に更新するが、予期せぬ変化が発生すると予測誤差が急増し、バイアスやシステムのミスマッチといったリスクが浮上する。こうした技術的特徴を捉え、予測誤差の変動をリアルタイムな規制シグナルとして活用することで、イノベーションを阻害することなく、適応的な AI ガバナンスモデルを構築することを提唱する。

例えば、自動運転においては、天候の変化や未知の道路状況により、AI システムの予測誤差が増加する可能性がある。適応的な AI ガバナンスの下では、これらの異常がリアルタイムな規制介入（リスク閾値の調整や人間の監視の要請など）を引き起こし、AI システムが実世界のリスクを説明することで、迅速に適応することができる。

3.2 適応的ガバナンスの特徴

適応的ガバナンスモデルは以下のような特徴を持っている。

- **適応的透明性（Adaptive Transparency）**：完全なアルゴリズム構造の公開ではなく、予測誤差を開示することで監視と企業秘密のバランスを取る。
- **適応型規制のための人間と AI の相互作用**: ユーザーや規制当局が AI の不確実性と動的に相互作用できる仕組みを実現する。
- **ガバナンスツールとしての世界モデルの適応**: AI の進化する内部表現を活用し、説明責任および規制の柔軟性を高める。

3.3 多層的なシステム設計の必要性

適応的ガバナンスモデルを実現するために、以下のような多層的なシステム設計が必要である。

- **AI リスクのモニタリング基盤**：AI システムは、予測誤差の変動を自ら報告し、自動的にリスク評価を実施する。高リスクな AI アプリケーションに対する許容リスクの閾値が定義され、ユーザービリティの高いダッシュボードが、予測誤差（不確実性）の視覚化を実現する。
- **規制の介入**：重大な予測誤差の異常が検出された場合、自動的な AI 監査が実施される。適応的なコンプライアンス命令により、AI 開発者はモデルを動的に調整する義務を負う。
- **人間中心の透明性ツール**：HCI に基づくインタラクティブなインターフェースを通じて AI の不安定性をユーザーに伝達する。ユーザーの行動データの研究に基づき、適応的な信頼較正（Adaptive Trust Calibration）の仕組みを実装し、最適な不確実性表現を実行する。政策立案者と AI 開発者が、ユーザービリティと説明責任のバランスを取るためのインターフェースを共同で設計する。

これらの仕組みは、単に予測誤差を技術的欠陥として捉えるのではなく、AI の開発と政策ガバナンスをつなぐ動的な規制インターフェースとして機能させる。社会のニーズと技術の進化に柔軟に対応する適応的な AI ガバナンスを実現する。

4. 実証的評価のための心理的調査の設計

4.1 調査目的と背景

本研究は、自動運転分野における AI ガバナンスの新たな制度設計を探究するため、心理的調査手法を応用して社会受容性を実証的に検討することを目的とする。特に、AI の予測誤差を媒介とした「動的 ODD（Operational Design Domain）」の可能性に着目する。

現行の規制枠組みにおいては、ODD はあらかじめ運行条件（天候・道路環境・速度等）を固定的に定義する必要がある。この「静的 ODD」は安全性確保の観点から保守的に設定される傾向があり、豪雨や道路工事といった状況を一律に排除することで、自動運転技術の適用範囲を不必要に制約している。

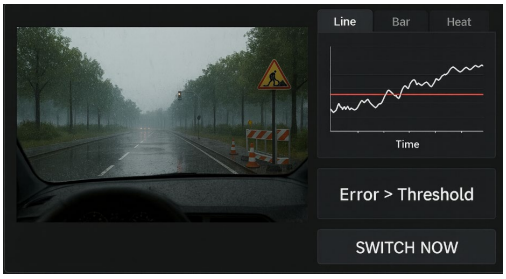
これに対し、「動的 ODD」では、AI の予測誤差をリアルタイムに監視し、閾値を超過した場合にのみ警告や人間介入、速度制御などを発動する。これにより、安全性を担保しつつ、「静的 ODD」では禁止される条件下でも限定的な走行を可能とすることが期待される。

4.2 調査デザイン

4.2.1 要因構造

- **ガバナンス方式（A）**：「静的 ODD」 vs 「動的 ODD」（予測誤差提示＋閾値監視）
- **状況（B）**：
  1. 晴天・良好路面（安定）
  2. 豪雨・視界低下（誤差急増）
  3. 工事区間・路面変更（誤差断続的上昇）
- **被験者属性（C）**：年齢層、職業、AI リテラシー水準

（参考）予測誤差表示画面のイメージ



4.2.2 予測誤差の提示条件

- **提示形式**：折れ線グラフ／バー表示／ヒートマップ
- **提示頻度**：常時提示 vs 閾値超過時のみ提示
- **閾値水準**：厳格（低閾値） vs 緩和（高閾値）
- **提示 UI**:CARLA シミュレーション映像と並んで予測誤差の可視化パネルと閾値超過の警告表示。

4.2.3 測定指標

- **信頼尺度**（Körber（2018）参照）：信頼性、予測可能性、意図理解
- **行動介入**（Okamura & Yamada(2020)参照）：自動運転継続意図、手動切替意図、推奨受容
- **制度受容**：倫理的観点（安全・人間尊重）、法的観点（規制・責任）、社会的観点（公共受容性・公平性）、技術的観点（実現可能性・性能向上）
- **属性による理解度の違い**：誤差表示の理解度。年齢、運転経験や自動運転知識の違いの影響。

4.2.4 仮説

- H1：動的 ODD 群は静的 ODD 群に比べ、豪雨・工事条件下で走行許容の制度受容が高い。
- H2：誤差提示条件（形式・頻度・閾値）の違いは、信頼水準の調整や制度受容に差をもたらす。
- H3：誤差提示によって被験者は「過信」や「過剰介入」を抑制し、介入の妥当性が高まる。
- H4：対象者属性（年齢・職業・リテラシー）により、受容度や誤差提示の効果は異なる。

4.3 調査手順（調査票の素案は別紙参照）

- 属性質問（年齢、職業、免許保有、AI リテラシーなど）
- シナリオ提示（状況 B1-B3 × ガバナンス方式 × 誤差提示条件。動画でシナリオを提示）
- シナリオ動画視聴後に、信頼度（安全性・予測可能性・意図理解）、介入タイミングの評価（早すぎ/遅すぎ/UI 有用性）、行動意図、制度受容度を確認。リッカード尺度を利用。
- 静的 ODD と動的 ODD の比較選好を問う設問
- 誤差提示 UI の理解度を確認し、調査終了

4.4 調査シナリオ（まとめ）

状況	ガバナンス方式	誤差提示条件	信頼（想定）	行動介入（想定）	制度受容（想定）
晴天・良好路面	静的 ODD	-	高信頼	継続意図高	許可（高）
晴天・良好路面	動的 ODD	各形式・頻度・閾値	高信頼	継続意図高	許可（高）
豪雨・視界低下	静的 ODD	-	信頼低	即時切替	許可不可
豪雨・視界低下	動的 ODD	各形式・頻度・閾値	信頼維持	減速・移譲受容	条件付き許可
工事区間	静的 ODD	-	信頼低	停止選好	許可不可
工事区間	動的 ODD	各形式・頻度・閾値	信頼維持	区間限定運行	条件付き許可

4.5 結果の分析手法

本研究では、調査結果を以下の方法で分析する。

- **主要な分析手法**
  主要な解析は順序データ(7 件法リッカート尺度)に適した累積リンク混合モデル(Cumulative Link Mixed Model, CLMM) を用いる。説明補助として平均値比較 (ANOVA/GLM) も行う。
- **主要モデル**
  ガバナンス方式 (被験者間要因) × 状況 (被験者内要因) × 被験者属性 (年齢・職業・AI リテラシー) を組み合わせた混合効果モデルを構築し、主要効果と交互作用を検証する。
- **媒介関係の検証**
  誤差提示が制度受容に与える影響が信頼を介して生じるかを媒介分析で確認する。信頼尺度については、確認的因子分析 (CFA) を実施し、各質問項目が信頼という潜在因子に収束しているかどうかを確認する。
- **信頼性の確認**
  信頼尺度の内部一貫性を Cronbach の α および McDonald の ω で確認する。さらに属性群間での測定不変性を検証する。
- **介入行動指標の分析**
  誤差閾値を超過しても介入せず継続した場合を「過信」、閾値未満で介入した場合を「過剰介入」と定義する。これらを二値または割合で指標化し、統計モデルに組み込む。

5. 期待される成果

本研究の調査設計に基づく実証により、以下の 5 つの成果が期待される。これにより、予測誤差に基づく適応的ガバナンスは、AI の進化的特性に適合し、利用者信頼の較正と制度受容を高める枠組みであり、安全性を損なわずに規制の柔軟化と技術革新の両立を可能にする制度的選択肢であることを示す。

5.1 規制枠組みへの示唆

静的 ODD では一律に禁止される豪雨や工事区間といった状況においても、予測誤差を監視する動的 ODD であれば限定的に許容可能であることを示し、安全性を損なわずに適用範囲を柔軟に拡張する新たな規制のあり方を提示する。

5.2 信頼調整の実証

誤差提示の形式や閾値が信頼に及ぼす影響を明らかにし、システムに対する過少な信頼・過剰な信頼を是正するための設計原理を提示する。

5.3 介入妥当性の定量評価

「過信」「過剰介入」を指標化し、誤差提示がこれらを抑制し、適切な介入を促す効果を実証する。これらの行動指標は、実験中のボタン押下やスペースキー操作のログと誤差閾値超過時刻の差分に基づいて算出する。

5.4 制度受容の多面的把握

倫理的・法的・社会的・技術的の 4 観点から制度受容性を測定し、動的 ODD の社会受容を促進あるいは阻害する観点を特定する。

5.5 対象層ごとの差異の提示

年齢、職業、AI リテラシーといった被験者属性による制度受容性の差を明示し、規制設計や社会実装において配慮すべき対象層の違いを示す。

参考文献

Bennear, L. S., & Wiener, J. B. (2019). *Adaptive Regulation: Instrument Choice for Policy Learning over Time*. Draft working paper. https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/Regulation%20-%20adaptive%20reg%20-%20Bennear%20Wiener%20on%20Adaptive%20Reg%20Instrum%20Choice%202019%2002%2012%20clean.pdf

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety -- A Review. http://arxiv.org/abs/2404.14082

Bogdoll, D., Bosch, L., Joseph, T., Gremmelmaier, H., Yang, Y., & Zöllner, J. M. (2023). *Exploring the Potential of World Models for Anomaly Detection in Autonomous Driving*. https://doi.org/10.1109/SSCI52147.2023.10371887

Coglianese, C., Lehr, D., Arroyo, K., Baker, T., Berk, R., Bowen, T., Cass, R., Choi, B., Conti-Brown, P., Finkel, A., Finkelstein, C., Fisch, J., Gold, J., Hanson, C., Joyce, K., Kable, J., Kreimer, S., Mayson, S., Ohm, P., … Zaring, D. (2017). *ILE INSTITUTE FOR LAW AND ECONOMICS Regulating by Robot: Administrative Decision Making in the Machine-Learning Era ARTICLES Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*. http://ssrn.com/abstract=2928293Electroniccopyavailableat:https://ssrn.com/abstract=2928293

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. http://arxiv.org/abs/1702.08608

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11, 127–138

Ha, D., & Schmidhuber, J. (2018). World Models. https://doi.org/10.5281/zenodo.1207631

Janssen, M. (2025). Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems. Policy and Society. https://doi.org/10.1093/polsoc/puae040

Hartwig, T., Ikkatai, Y., Takanashi, N., & Yokoyama, H. M. (2023). Artificial intelligence ELSI score for science and technology: a comparison between Japan and the US. *AI and Society*, *38*(4), 1609–1626. https://doi.org/10.1007/s00146-021-01323-9

Ikkatai, Y., Hartwig, T., Takanashi, N., & Yokoyama, H. M. (2023). Segmentation of ethics, legal, and social issues (ELSI) related to AI in Japan, the United States, and Germany. *AI and Ethics*, *3*(3), 827–843. https://doi.org/10.1007/s43681-022-00207-y

Körber, M. (2018). *Theoretical considerations and development of a questionnaire to measure trust in automation*. https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire

Lee, R. G., & Petts, J. (2013). Adaptive Governance for Responsible Innovation. In *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society* (pp. 143–164). John Wiley and Sons. https://doi.org/10.1002/9781118551424.ch8

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2). https://doi.org/10.1371/journal.pone.0229132

Reuel, A., & Undheim, T. A. (2024). Generative AI Needs Adaptive Governance. http://arxiv.org/abs/2406.04554

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Zeng, Z., Zhang, C., Liu, F., Sifakis, J., Zhang, Q., Liu, S., & Wang, P. (2024). World Models: The Safety Perspective. https://doi.org/10.48550/arXiv.2411.07690

鈴木雅大（2023）「自由エネルギー原理と深層学習―世界モデルを軸として―」人工知能　Vol. 38 No. 6（2023 年 11 月号）

別紙：調査用アンケート票（素案）

A. 属性情報

1. 年齢、職業（一般ドライバー、技術専門家、行政関係者など）
2. 運転免許の有無、運転経験年数
3. AI や自動運転に関する知識・経験の程度（7 件法リッカート尺度）

B. シナリオ動画視聴中の入力（行動ログ）

1. 「Switch NOW」ボタン／スペースキー押下時刻（press_ms）
2. 誤差閾値超過時刻（t_thresh_ms）
3. Lead/Lag 指標 = press_ms − t_thresh_ms
4. ログから「過信（閾値超過後も無介入）」／「過剰介入（閾値未満で介入）」を算出

C. 信頼に関する質問（7 件法リッカート尺度）

1. この自動運転システムは安全に走行できると感じましたか。
2. このシステムの行動は予測しやすいと感じましたか。
3. このシステムの行動の意図を理解できたと感じましたか。
4. このシステムを信頼できると感じましたか。

D. 行動介入（7 件法リッカート尺度）

1. 自分の判断で手動に切り替えると思いましたか。
2. システムの推奨に従いたいと思いましたか。
3. 手動に切り替えたタイミングは早すぎたと思いますか、それとも遅すぎたと思いますか。
4. 誤差提示 UI は介入タイミングの判断に役立ったと思いますか。

E. 制度受容（7 件法リッカート尺度）

1. この状況での自動運転を社会として許可すべきだと思いますか。
2. 倫理的観点：この条件下での走行は、人間の安全を十分に尊重していると思いますか。
3. 法的観点：この仕組みでは、事故時の責任を明確にできると思いますか。
4. 社会的観点：この条件下での走行は、社会的に受け入れられると思いますか。
5. 技術的観点：このシステムは技術的に実現可能で信頼できると思いますか。
6. 誤差監視があるなら、豪雨のような条件下でも走行を許可してよいと思いますか。
7. 規制当局は、誤差が閾値を超えた場合に人間への運転移譲を義務化すべきだと思いますか。
8. 静的 ODD（条件外では常に禁止）と比べて動的 ODD（誤差監視つきで条件付き許可）はよい制度であると思いますか。

F. 理解度チェック（7 件法リッカート尺度）

1. 提示された誤差表示（グラフやバー）は理解できましたか。