ブラックボックス・アルゴリズムに対する自己監査プロトコル：
オンラインプラットフォームのための低コストな公平性証明手法

A Self-Auditing Protocol for Black-Box Algorithms:
A Low-Cost Fairness Certification Method for Online Platforms

著者：オンライン麻雀ゲームプレイヤー

要旨 (Abstract)

本稿は、オンラインプラットフォームが、内部アルゴリズムを秘匿したまま、かつ高コストな第三者認証に依存することなく、主体的かつ継続的にシステムの公平性を証明するための標準化された自己監査プロトコルを提案する。我々は、オンライン麻雀をケーススタディとし、システムの「公平性」を静的、動的、条件的、結果的の 4 つの観点から操作的に定義する。本プロトコルは、観測可能なアウトプット（ゲームログ）のみを利用し、大規模シミュレーション、時系列分析、混合効果モデルを組み合わせることで、確率的操作の有無からゲームバランスの健全性までを多角的に検証する。さらに、このプロセスを自動化する判定エンジン（MFVE）を開発し、客観的な評価レポートを生成する。本研究は、プラットフォームが自主的に透明性を確保し、ユーザーとの信頼関係を構築するための、実践的かつ低コストな技術的解決策を提供し、デジタルサービスにおける新たな信頼性保証のパラダイムを提示するものである。

１.緒論 (Introduction)

1.1.デジタルプラットフォームにおける信頼性のジレンマ

現代のデジタルサービスにおいて、ユーザー体験は独自のアルゴリズムによって最適化されている。しかし、これらのアルゴリズムは企業の競争力の源泉であり、多くの場合ブラックボックスとして秘匿される。この情報の非対称性は、「プラットフォームは本当に公正か？」というユーザーの疑念を生み、信頼関係を蝕む根源的なジレンマとなっている。従来、この問題への対処法は、高コストな第三者機関による監査か、あるいは法的規制によるアルゴリズム開示要求であったが、いずれもプラットフォームにとっては大きな負担となる。

1.2. 本稿が提案する解決策：主体的・低コストな自己監査 本稿は、このジレンマに対する新たな解決策を提案する。それは、プラットフォーム自身が、内部ロジックを開示することなく、客観的かつ科学的な根拠をもって公平性を証明できる「自己監査プロトコル」である。このアプローチは、プラットフォームに以下の利点をもたらす。

- 主体性: 第三者に依存せず、自社のガバナンスの一環として公平性を管理できる。
- 低コスト: 監査にかかる費用と時間を大幅に削減できる。
- 知的財産の保護: コアとなるアルゴリズムを秘匿したまま、信頼性を証明できる。
- 予防的対策: 問題が外部から指摘される前に、内部で問題を検知し、改善することが可能になる。

本稿では、複雑な確率モデルを持つオンライン麻雀を最も挑戦的な事例として取り上げ、このプロトコルの有効性を実証する。

2. 検証プロトコル (Methodology)

本プロトコルは 4 つの独立した検証モジュールから構成される。分析には最低 100 万局分の完全なゲームロ

グを要する。

2.1. モジュールA: 静的公平性 配牌の品質を、シャンテン数や有効牌枚数などから算出される総合指標「配牌品質スコア（HQS)」で定量化する。次に、10億回のモンテカルロシミュレーションを実行し、HQSおよび各指標の極めて高精度な理論分布を生成する。この理論分布と観測データ（実測100万局）の分布をコルモゴロフ・スミルノフ検定およびカイ二乗適合度検定を用いて比較し、統計的有意差の有無を検定する。10億回という試行回数は、基準となる理論分布自体のサンプリング誤差を無視できるレベルにまで低減させ、検定の信頼性を最大化するために不可欠である。

2.2. モジュールB: 動的公平性 ツモ牌の系列データに対し、自己相関分析（ACF）および連検定（Runs Test）を適用し、系列内における統計的に有意なパターンや相関が存在しないこと（無記憶性）を検証する。

2.3. モジュールC: 条件的公平性 HQSを目的変数とし、プレイヤーレートや席順を説明変数とする一般化線形混合モデル（GLMM）を構築する。各説明変数の係数が統計的に有意に0と異なるかを検定し、プレイヤー属性による配牌品質へのバイアスが存在しないことを検証する。

2.4. モジュールD: 結果的公平性 最終順位ポイントを目的変数とし、HQS（運の代理変数）とプレイヤーID（実力の代理変数）を組み込んだ線形混合モデルを用いて分散成分分析（VPC）を行う。これにより、ゲーム結果の分散（ばらつき）が、どれだけの割合で「運」と「実力」に起因するのかを定量的に分離する。このVPC値を、専門家ヒアリングやAIシミュレーションによって事前に設定された「公平性ベンチマーク区間」（例: 運30-45%, 実力40-55%）と比較する。

3. 判定エンジンと評価基準 (Evaluation Engine and Criteria)

3.1. 判定ツール「MFVE」の実装 提案プロトコルを実装したPython製ツール「Mahjong Fairness Verification Engine (MFVE)」を開発した。本ツールは、統計解析から得られた指標値リスト（CSV）と、判定基準を定義した設定ファイルを入力とし、後述の評価基準に基づいた最終レポートを自動生成する。

3.2. 統合的評価基準 各モジュールの検定結果を統合し、以下の基準で総合判定を下す。

- **A判定（極めて公平）:** モジュールA, B, Cをクリアし、かつモジュールDのVPC値がベンチマーク区間の厳密基準を満たす。
- **B判定（実用上公平）:** モジュールA, B, Cをクリアし、かつモジュールDのVPC値がベンチマーク区間の緩和基準（若干の逸脱を許容）を満たす。
- **C判定（要改善）:** モジュールA, B, Cのいずれかに不合格が見られる。確率生成プロセスに何らかのバイアスが存在する可能性。
- **D判定（不公平）:** モジュールA, B, Cはクリアするが、モジュールDのVPC値がベンチマーク区間から著しく逸脱している。確率操作はないが、ゲームバランスが著しく損なわれている状態。

4. ケーススタディ：仮想プラットフォームの診断 (Case Study)

4.1. 診断結果 仮想プラットフォームから得られた以下の結果指標値リスト（表1）をMFVEで分析した。

表 1：仮想プラットフォームの検証結果指標値

metric_name value	ci_lower ci_upper	p_value		
hqs_ks_statistic	0.0012	0.085		
shanten_chi2_statistic	12.5	0.091		
tsumo_acf_lag	1	-0.0005	0.452	
tsumo_runs_z_score	0.89	0.373		
glmm_rate_coefficient	0.0001	-0.0002	0.0004	0.812
glmm_seat_coefficient	-0.0003	-0.0008	0.0002	0.215
vpc_hand_estimate	0.47	0.46	0.48	
vpc_player_estimate	0.38	0.37	0.39	

4.2. 結果の解釈 MFVE による判定結果は「D 判定: 不公平」であった。

・ モジュール A, B, C の p 値はいずれも有意水準 0.01 を大きく上回っており、確率生成プロセスにおける静的・動的・条件的公平性は完全に担保されている。不正や操作の痕跡は一切見られない。

・ しかし、モジュール D において、運の寄与度（VPC_Hand=47%）がベンチマーク（30-45%）を上回り、実力の寄与度（VPC_Player=38%）がベンチマーク（40-55%）を下回った。

・ この結果は、システムが確率的に誠実であるにもかかわらず、そのゲームデザイン（例: ドラの比重など）が、実力よりも運が結果を左右しやすい「大味」な環境を生み出していることを示唆している。

5. 考察 (Discussion)

5.1. プラットフォーム・ガバナンスへの貢献 本研究が提示する自己監査プロトコルは、単なる技術的な検証手法に留まらない。これは、デジタルプラットフォームにおける新たなコーポレート・ガバナンスのツールである。プラットフォームは、本プロトコルに準拠した定期的なレポートを公開することで、ユーザー、投資家、規制当局に対して、自社のサービスの健全性を自主的かつ継続的に証明することが可能になる。これは、CSR（企業の社会的責任）活動の一環としても極めて有効である。

5.2. 第三者認証との関係性 本プロトコルは、第三者認証を完全に否定するものではない。むしろ、両者は補完的な関係にある。プラットフォームは日常的な運用において本プロトコルによる自己監査を行い、その上で、より重要な局面（例: 年次報告、新規サービス開始時など）において、自己監査プロセスそのものが正しく運用されているかを第三者機関に監査させるという、より効率的で高度な信頼性保証モデルを構築できる。

5.3. ブラックボックス監査モデルとしての汎用性 本研究は、ゲームの公平性評価が単なる「不正の検知」から「ゲームバランスの健全性評価」へと進化する必要があることを示した。しかし、その意義は麻雀という一分野に留まらない。本稿で提案した「アウトプットベースの多角的検証プロトコル」は、内部ロジックが非公開である他の多くのデジタルシステムに応用可能な、汎用的なブラックボックス監査モデルとしての側面を持つ。

6. 結論 (Conclusion)

本稿では、オンライン麻雀の公平性を多角的に検証・評価するための包括的なプロトコルを提案し、その有

効性をケーススタディによって実証した。本手法は、開発者とユーザー間の信頼を醸成し、e スポーツとしての麻雀の健全な発展に寄与する、客観的かつ透明性の高い評価基準となり得るだろう。

参考文献 (References) [1] Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. [2] Lehmann, E. L., & Romano, J. P. (2005). Testing statistical hypotheses. [3] Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. [4] Silver, D., et al. (2019). Suphx: Mastering Mahjong with Deep Reinforcement Learning.