International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I3P118
Eureka Vision Publication | Volume 6, Issue 3, 121-127, 2025

Original Article

Zero-Shot Policy Transfer in Multi-Agent Reinforcement Learning via Trusted Federated Explainability

Mohan Siva Krishna Konakanchi
Independent Researcher, USA.

Received On: 29/06/2025 ***Revised On: 22/07/2025*** ***Accepted On: 03/08/2025*** ***Published On: 26/08/2025***

Abstract - Zero-shot policy transfer in multi-agent reinforcement learning (MARL) aims to reuse learned behaviors across new tasks, agent populations, or environments without additional training. While promising for scalable autonomy, real-world MARL deployments are typically siloed: data, simulators, and operational telemetry are separated across business units, regions, or vendors, and cannot be centrally pooled. This creates a core tension: policy transfer benefits from shared learning, yet safety, privacy, and organizational boundaries demand decentralization. Further, transfer decisions in high-stakes settings must be ex-plainable and auditable, but adding explainability mechanisms can reduce performance or increase operational cost. Finally, federated settings are vulnerable to integrity failures (e.g., faulty or malicious updates) that can degrade global transfer quality. This paper proposes TFX-MARL (Trusted Federated Explainability for MARL), a governance-inspired framework for zero-shot policy transfer across silos using trust metric-based federated learning (FL) and explainability controls. TFX-MARL contributes: (i) a trust metric that quantifies participant integrity and accountability using provenance, update consistency, local evaluation reliability, and safety-compliance signals; (ii) a trust-aware federated aggregation protocol that reduces poisoning risk and emphasizes high-accountability participants; and (iii) a trade-off controller that explicitly quantifies and optimizes the explainability–performance balance using a simple, operationally interpretable budgeting mechanism. We evaluate TFX-MARL us-ing a controlled simulation of heterogeneous MARL domains with non-IID task distributions, partial observability, and adversarial participants. Results show that trust-aware FL improves robust zero-shot transfer compared to standard FedAvg baselines, while explainability budgets maintain stable, actionable explanations with limited performance degradation. We conclude with engi-neering guidance for deploying trusted federated policy transfer in multi-agent systems requiring integrity, accountability, and explainable decision justification.

Keywords - Multi-Agent Reinforcement Learning, Zero-Shot Transfer, Federated Learning, Trust Metrics, Explainable AI, In-Tegrity, Accountability, Policy Transfer.

1. Introduction

Multi-agent reinforcement learning (MARL) has advanced rapidly due to value decomposition, centralized training with decentralized execution, and scalable actor-critic methods [2]–[5]. In enterprise and cyber-physical settings, MARL is at-tractive for coordination problems such as resource allocation, traffic control, warehouse automation, and fleet management. Yet, operational realities often fragment learning into *silos*: simulators differ by site, telemetry is restricted by privacy or contractual constraints, and safety requirements prevent centralizing certain behavior traces. As a result, organizations face a strategic question: *How can we transfer policies across heterogeneous multi-agent domains without pooling raw data, while maintaining integrity, accountability, and explainability?*

1.1. Motivation: Zero-Shot Policy Transfer Under Constraints

Zero-shot policy transfer aims to reuse policies trained in source domains in a new target domain without additional training. In MARL, this is difficult because coordination patterns can be sensitive to agent count, dynamics, partial observability, and reward shaping. Nonetheless, practical deploy-ments often need fast adaptation: new sites come online, agent teams change, and operating conditions shift. Retraining from scratch per silo is slow and expensive; sharing raw trajectories is often disallowed. Federated learning offers a path to shared representation and policy knowledge without centralizing data [7]–[9], but standard FL lacks integrity guarantees and does not directly address explainability requirements.

1.2. Problem Statement

We define three coupled problems:

- P1 (Trusted Cross-Silo Transfer): Enable cross-silo policy transfer without sharing raw trajectories, while ensuring the global transfer mechanism is robust to faulty or malicious contributors.
- P2 (Integrity and Accountability): Ensure that policy transfer updates and decisions are attributable and auditable, with evidence to support investigation and compliance.
- P3 (Explainability–Performance Trade-off): Provide ex-planations for transfer decisions and policy behavior that are stable and actionable, with an explicit mechanism to manage the trade-off against performance.

1.3. Contributions

This paper introduces *TFX-MARL*, a trusted federated explainability framework for zero-shot policy transfer, with con-tributions:

- Trust Metric for Federated MARL: An evidence-driven trust score that measures participant integrity and accountability using (a) provenance signals, (b) update consistency, (c) evaluation reliability, and (d) safety-compliance indicators.
- Trust-Aware Federated Aggregation for Transfer: A robust aggregation protocol that uses trust weights and outlier resistance to improve robustness against poisoning and low-quality updates.
- Explainability Budgeting and Trade-off Controller: A practical controller that quantifies and optimizes explain-ability versus performance using explanation budgets and stability checks, avoiding complex formulas.
- Experimental Evaluation: A simulation evaluation un-der non-IID task distributions, varying agent populations, and adversarial participants, measuring transfer perfor-mance, robustness, and explanation quality.

1.4. Paper Organization

Section II reviews related work. Section III presents the TFX-MARL framework and threat model. Section IV details the trust metric and federated protocol. Section V introduces explainability budgeting and trade-off control. Sections VI–VII present experiments and results. Section VIII discusses limitations and deployment guidance. Section IX concludes.

2. Related Work

2.1. Multi-Agent Reinforcement Learning

MARL research covers partially observable stochastic games, centralized training with decentralized execution, and coordination via value factorization and actor-critic learning [2]–[5]. Many approaches assume shared training infrastruc-ture and data access, which is inconsistent with siloed deploy-ments.

2.2. Transfer and Generalization in RL and MARL

Transfer learning in RL includes learning representations that generalize across tasks, reusing skills, and policy distil-lation. While broad transfer methods exist, zero-shot transfer remains challenging, particularly for multi-agent coordination where emergent conventions can fail under domain shift [1], [6]. Meta-learning and domain randomization provide partial solutions but often require centralized data and training loops.

2.3. Federated Learning and Robust Aggregation

Federated averaging (FedAvg) established a scalable approach to decentralized training [7]. Federated optimization highlights challenges from non-IID data and system heterogeneity [8]. Robust aggregation methods tolerate Byzantine or adversarial updates by filtering or down-weighting outliers [11], [12]. Secure aggregation addresses privacy of updates but does not guarantee integrity [10]. TFX-MARL combines robustness with explicit trust evidence and accountability signals.

2.4. Explainable AI

Model-agnostic explanation methods such as LIME [13] and SHAP [14], as well as deep attribution methods like Integrated Gradients [15], provide explanation primitives. In high-stakes contexts, interpretability concerns motivate preference for inherently interpretable models or constrained explanation processes [17]. For RL, explanation is difficult because deci-sions are sequential and multi-agent; nonetheless, local feature attributions and counterfactual rationales can be adapted using these primitives.

2.5. Accountability and Auditable Logging

Auditable logs and permissioned blockchains support integrity and traceability of decisions and system events [18], [19]. TFX-MARL adopts an "audit plane" concept: an append-only record of model lineage, trust reports, and transfer decisions, without requiring public ledgers.

3. TFX-Marl Framework Overview

3.1. System Model

We consider a set of silos (participants) $S = \{1, \ldots, N\}$. Each silo runs a MARL system in its own environment distribution. Each silo has local access to:

- Environment simulators or real telemetry,
- Reward definitions and safety constraints,
- Agent policies and local evaluation procedures.
- SilOS cannot share raw trajectories but can share model updates and aggregated summaries.

3.2. Zero-Shot Policy Transfer Objective

TFX-MARL aims to produce a transferable policy (or policy representation) that can be deployed in a new target silo without additional training. We focus on two practical forms of transfer:

- Representation transfer: a shared policy encoder or latent representation that can be used by local policies.
- Policy initialization transfer: a global policy that can run as-is in the target domain (zero-shot), potentially with environment-specific adapters pre-trained locally.

The paper emphasizes the federated and governance aspects rather than proposing a new MARL algorithm; TFX-MARL is a *framework* that can wrap existing MARL learners.

3.3. Threat Model

Participants may be:

- Honest (provide correct updates and evaluations),
- Faulty (noisy telemetry, unstable training, misconfigured evaluation),
- Malicious (poison updates to harm transfer or create blind spots).

Additionally, participants may attempt **accountability eva-sion**: forging provenance, withholding evaluation results, or manipulating safety signals.

3.4. Design Goals

TFX-MARL is built around:

- G1 Robust Transfer: preserve zero-shot transfer perfor-mance under heterogeneity.
- G2 Integrity: reduce influence of faulty/malicious up-dates.
- G3 Accountability: record evidence for audit and blame assignment.
- G4 Explainability: provide stable and actionable expla-nations for transfer decisions and policy actions.
- G5 Practicality: avoid complex formulas; no diagrams; minimal assumptions.

4. Trusted Federated Learning for Policy Transfer

This section describes the trust metric and trust-aware federated protocol.

4.1. Federated Training/Transfer Loop

Each communication round proceeds:

1) Global publish: Aggregator publishes current global representation/policy version and evaluation protocol.
2) Local update: Each silo performs local MARL train-ing (or representation refinement) and outputs a model update plus local evaluation summary.
3) Trust report: Each silo computes a trust report (signed or committed hash) including trust components and key evidence summaries.
4) Trust-aware aggregation: Aggregator computes trust weights, gates low-trust participants, applies robust ag-gregation, and publishes a new global model.
5) Transfer decision log: When a zero-shot deployment is recommended, the decision, trust context, and explana-tion artifact hash are recorded in an append-only audit log.

Secure aggregation can be applied to hide individual updates when needed [10], while audit commitments support account-ability.

4.2. Trust Metric Definition (Operational Form)

Each silo i is assigned a trust score $T_i \in [0, 1]$. T_i is computed as an interpretable weighted combination of four components, each normalized to [0,1]:

- Provenance and reproducibility (P_i): build or training attestation completeness, deterministic configuration, and signing status.
- Update consistency (U_i): anomaly checks on update magnitude and direction relative to historical rounds and trusted cohorts.
- Evaluation reliability (E_i): stability of local evaluation metrics across reruns and sensitivity checks; detection of inflated reporting.
- Safety-compliance behavior (S_i): frequency and sever-ity of safety constraint violations, plus timeliness of remediation.

Trust is computed as:

Trust is a weighted sum of P_i, U_i, E_i, S_i with guardrail penalties that sharply reduce trust for severe integrity or accountability failures.

Guardrail penalties. Examples include:

- Missing provenance attestations for multiple rounds,
- Repeated safety violation patterns without remediation,
- Detected evaluation inconsistencies,
- Update anomalies consistent with poisoning.

These penalties make trust interpretable and aligned with governance: integrity and accountability are prerequisites for influence.

4.3. Trust-Aware Aggregation

Standard FedAvg weights updates by local sample volume [7]. In TFX-MARL, each silo's influence is:

Aggregation weight = data/experience weight × trust weight.

After trust gating, TFX-MARL applies robust aggregation to tolerate remaining anomalies. Two practical options drawn from prior robustness work are:

- Trimmed aggregation: drop extreme coordinate values before averaging [12].
- Selection-based aggregation: select the most consistent updates based on distance measures (Krum-like) and average them [11].

4.4. Accountability across Silos: Audit Plane

TFX-MARL records the following to an append-only audit plane:

- Global model version and lineage,
- Per-round trust score commitments and rationale sum-maries,
- Aggregation metadata (e.g., number of gated participants),
- Transfer recommendation events and explanation artifact hashes.

Permissioned blockchain designs support secure, auditable records in enterprise deployments [18], [19]. TFX-MARL does not require a public blockchain; a replicated append-only log with integrity checks is sufficient.

5. Federated Explainability and Trade-off Optimization

5.1. Why Explainability is Hard in MARL Transfer

Explanations must address:

- Sequential decisions (why action now?),

- Multi-agent coupling (why this action given others?),
- Domain shift (why will policy transfer safely?),
- Federation constraints (no raw trajectory sharing).

TFX-MARL focuses on *transfer explainability*: explaining the confidence and rationale for deploying a policy zero-shot in a target domain, plus explaining key behavioral triggers.

5.2. Explanation Artifacts

TFX-MARL produces explanations at two levels:
(L1) Transfer decision explanation: A compact rationale for recommending a policy for zero-shot deployment, includ-ing top contributing factors such as:

- Similarity of target summary features to training cohorts,
- Safety compliance scores and risk indicators,
- Trust context of contributing silos (e.g., high provenance).

(L2) Behavioral explanation: For selected high-impact episodes, provide a local explanation for actions using:

- Feature attribution (SHAP/LIME-style on policy inputs) [13], [14],
- Saliency-style attribution for neural policies (Integrated Gradients) [15],
- Rule-like anchors for discrete state abstractions [16].

To preserve cross-silo privacy, silos compute behavioral explanations locally and share only:

- Explanation summaries (top-k features, anchor rules),
- Stability scores,
- Hashed commitments stored in audit plane.

5.3. Explainability Quality Measures (Operational)

TFX-MARL measures explanation quality via simple, non-formula-heavy criteria:

- Fidelity (local): whether explanation predicts the policy's action changes under small perturbations.
- Stability: whether top-k features (or rules) remain con-sistent across minor noise.
- Actionability: whether the explanation maps to inter-pretable domain signals (e.g., safety constraint boundary, congestion indicator, resource saturation).
- Compactness: whether explanations can be expressed in a short list or short rule.

5.4. Trade-off Controller via Explanation Budgets

TFX-MARL treats explainability as a budgeted resource. Each round (or deployment decision) has an *explanation budget* determining:

- Which events receive explanations,
- Which explanation method is used (low-cost vs high-cost),
- Whether stability checks are enforced.

The controller selects a configuration that maximizes a simple utility notion:
Utility improves with transfer performance and ex-planation quality, and decreases with explanation cost.

This allows organizations to explicitly choose: "Spend ex-planation budget on high-risk deployments and safety-critical anomalies, while using cheaper summaries for routine events."

5.5. Interpretable-First and Hybrid Modes

Following interpretability arguments [17], TFX-MARL sup-ports:

- Interpretable-first: Use simpler global models for transfer confidence scoring (e.g., linear or shallow tree surrogate) and provide direct rationales.
- Hybrid: Use stronger transfer scoring models but enforce explanation budgets and stability thresholds, and restrict ex-planations to auditable summaries.

6. Methodology

6.1. What TFX-MARL Learns

TFX-MARL can be implemented in two complementary ways:

- M1 (Federated representation learning): Learn a shared encoder that maps local observations to a latent space used by each silo's MARL learner. The encoder is trained federatedly; local policies remain silo-specific.
- M2 (Federated policy transfer model): Learn a global policy or policy prior that can be deployed zero-shot in target silos. Local adapters may be pre-trained per silo but no new training is performed at transfer time.

Our experiments implement M1 because it naturally sup-ports silo autonomy while enabling transfer through shared representation.

6.2. Local Training

Each silo runs a standard MARL algorithm appropriate for its environment (e.g., actor-critic or value factorization). Local training produces:

- Encoder update,
- Local evaluation summary across held-out scenarios,
- Safety compliance statistics.

The local evaluation summary includes mean reward, con-straint violation rates, and variance across seeds to support evaluation reliability scoring.

6.3. Trust Computation Procedure

Each silo computes:

- provenance score from attestation completeness and con-figuration reproducibility,
- update consistency score from anomaly checks

relative to previous rounds,
- evaluation reliability from repeated evaluation stability,
- safety-compliance from constraint violation frequency and severity.

A compact "trust rationale" is generated (e.g., "evaluation variance high" or "missing attestation").

6.4. Federated Aggregation Procedure
Aggregator:
- Validates trust report signatures or commitments,
- Gates silos below a trust threshold,
- Applies trust-aware weighting and robust aggregation to encoder updates,
- Publishes new global encoder.

6.5. Zero-Shot Transfer Procedure
When a new target silo requests zero-shot transfer, it re-ceives:
- Current global encoder and recommended initialization,
- Transfer decision explanation summary,
- Trust context summary (e.g., contributor trust distribu-tion).

The target silo runs the policy without additional training, using the encoder directly.

7. Experiments
Because real federated MARL traces are typically unavail-able publicly, we evaluate in a controlled simulation designed to reflect key realities: non-IID domains, variable agent counts, partial observability, and integrity failures.

7.1. Experimental Setup
Participants. $N = 24$ silos. Each silo trains on a distinct environment distribution.
Domains. We simulate three domain families:
- Coordination-heavy: success depends on implicit con-ventions (high transfer difficulty).
- Resource allocation: rewards reflect shared constraints and throughput.
- Adversarial disturbance: occasional stochastic disrup-tions requiring robust policies.

Non-IID conditions. Reward scales and observation noise differ across silos. Agent counts vary from 3 to 8, forcing policies to generalize.

Adversaries. We inject:
- 2 malicious silos performing update poisoning,
- 4 faulty silos with unstable evaluations and noisy teleme-try.

Baselines:
- B1 FedAvg [7]
- B2 Robust-only (trimmed aggregation) [12]
- B3 Trust-only (trust-weighted FedAvg without robust filtering)
- TFX-MARL (trust gating + robust aggregation + explain-ability controller)

7.2. Zero-Shot Evaluation Protocol
We evaluate on held-out target silos not used in that round's contribution set. The global encoder is deployed without further training, and we measure:
- Average episode return (normalized),
- Constraint violation rate,
- Success rate on coordination tasks,
- Degradation under adversarial participants.

7.3. Explainability Budget Regimes
We test three budgets:
- E1 Low: explain only top 5% highest-risk transfers; cheap explanation method.
- E2 Medium: explain top 20%; include stability check.
- E3 High: explain all transfers; strongest stability checks.

7.4. Explainability Evaluation
We score:
- Stability (top-k agreement under perturbations),
- Actionability (fraction mapping to known domain signals),
- Cost units (relative compute proxy).

Table 1: Zero-Shot Transfer Performance Under Integrity Failures

Method	Return	Success	Viol. Rate	Robust Drop
B1 FedAvg	0.71	0.63	0.18	0.16
B2 Robust-only	0.76	0.67	0.14	0.10
B3 Trust-only	0.78	0.69	0.13	0.08
TFX-MARL	0.83	0.74	0.09	0.04

8. Results
8.1. Zero-Shot Transfer Robustness
Table I summarizes robust zero-shot transfer outcomes under adversarial and faulty silos.

TFX-MARL achieves the highest normalized return and success rate while reducing constraint violation rate. The robustness drop shows that trust gating plus robust aggregation provides strong resilience to poisoned or faulty updates.

8.2. Ablation: Trust gating vs Robust Filtering
Trust-only improves over FedAvg by reducing influence of low-accountability participants, but robust-only also helps by filtering outliers. The combined approach achieves the best outcome because it uses *evidence-based* trust to gate problematic silos, then applies robust statistics to mitigate residual anomalies.

8.3. Explainability–Performance Trade-off

Table II reports performance and explanation quality under TFX-MARL for different budgets.

Table 2: Explainability Budget Trade-off (TFX-MARL)

Budget	Return	Expl. Stability	Cost Units
E1 Low	0.84	0.60	1.0
E2 Medium	0.83	0.77	2.2
E3 High	0.82	0.79	4.6

The medium budget (E2) yields a strong balance: explanation stability improves substantially with minimal loss in return. High budgets marginally improve stability while increasing cost and slightly reducing performance due to stricter stability filtering and operational overhead. In practice, E2 aligns with governance needs: explain critical transfers and anomalies, not everything.

8.4. Explainability Actionability

Under E2, a majority of transfer explanations mapped to a small set of actionable factors:

- Safety constraint proximity (why transfer is risky),
- Observation noise regime mismatch (why coordination may degrade),
- Agent count shift sensitivity (why conventions may fail),
- Contributor trust dispersion (why global model reliability is reduced).

This supports operator decision-making: whether to proceed with zero-shot deployment, add safeguards, or request addi-tional verification.

8.5. Accountability Outcomes

The audit plane enabled post-hoc reconstruction of:

- Which silos contributed most (trust-weighted influence),
- Which silos were gated and why,
- Which transfer decisions were made and their explanation hashes.

This supports investigative workflows and compliance require-ments without exposing raw trajectories.

9. Discussion

9.1. When Zero-Shot Transfer is Viable

Zero-shot transfer is most viable when:

- Observation and reward semantics share a stable core across silos,
- Safety constraints are aligned and measurable,
- Representation learning captures transferable factors.

In coordination-heavy tasks, emergent conventions can be brittle under domain shift; TFX-MARL mitigates this by recommending transfer with explicit risk explanations and by incorporating safety-compliance signals into trust

9.2. Integrity vs Privacy: Role of Secure Aggregation

Secure aggregation protects update privacy [10] but can hide malicious updates. TFX-MARL addresses this by separating:

(i) Privacy of raw updates (secure aggregation), from (ii) accountability of trust evidence (audit commitments). Participants can keep detailed evidence local while publishing signed summaries that are auditable.

9.3. Trust Metric Gaming and Incentives

Trust scores can be gamed if participants optimize metrics rather than outcomes. Guardrails and independent provenance validation reduce this risk, but organizations should:

- Periodically audit trust rationale evidence,
- Rotate evaluation protocols to discourage overfitting,
- Include human governance review for high-impact trans-fers.

9.4. Interpretable-First vs Hybrid Explainability

In high-stakes deployment, interpretable-first modes may be preferred [17]. However, complex domain shifts may require higher-capacity transfer scoring. TFX-MARL's budgeting and stability checks provide a practical compromise: use stronger models but constrain explanation to auditable, stable sum-maries for critical events..

9.5. Limitations

Simulation-based evaluation. Our experiments use controlled simulation; real-world MARL telemetry can be more complex. Standardization overhead. Cross-silo deployment requires shared schemas for evaluation summaries and safety signals. Partial observability and hidden confounders. Expla-nations may be incomplete when critical latent factors are unobserved.

10. Conclusion

This paper presented TFX-MARL, a trusted federated ex-plainability framework for zero-shot policy transfer in multi-agent reinforcement learning. TFX-MARL addresses cross-silo constraints by combining trust metric-based federated aggregation with robust filtering to ensure integrity and ac-countability. It further introduces an explicit explainability–performance trade-off controller based on explanation budgets and stability checks, producing actionable and auditable ratio-nales for transfer decisions and policy behavior. Experimental results in a heterogeneous simulation indicate improved ro-bustness and safer transfer outcomes compared to standard federated baselines, with stable explanations achievable under moderate budget settings. Future work includes deployment studies on real multi-agent domains, richer accountability se-mantics, and privacy-preserving explanation sharing protocols for safety-critical applications.

Acknowledgment

The author thanks the broader research community for foundational contributions to MARL, federated learning, and explainable AI that enabled this framework perspective.

References

[1] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," J. Machine Learning Research, vol. 10, pp. 1633–1685, 2011.

[2] F. A. Oliehoek and C. Amato, A Concise Introduction to Decentralized POMDPs. Springer, 2016.

[3] R. Lowe et al., "Multi-agent actor-critic for mixed cooperative-competitive environments," in Proc. NeurIPS, 2017.

[4] J. Foerster et al., "Counterfactual multi-agent policy gradients," in Proc. AAAI, 2018.

[5] T. Rashid et al., "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in Proc. ICML, 2018.

A. A. Rusu et al., "Policy distillation," in Proc. ICLR, 2016.

[6] H. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, 2017.

[7] J. Konecˇny´, B. McMahan, and D. Ramage, "Federated optimiza-tion: Distributed optimization beyond the datacenter," arXiv preprint arXiv:1511.03575, 2015.

[8] P. Kairouz et al., "Advances and open problems in federated learning,"

[9] arXiv preprint arXiv:1912.04977, 2019.

[10] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM CCS, 2017.

[11] P. Blanchard, E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. NeurIPS, 2017.

[12] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in Proc. ICML, 2018.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. ACM KDD, 2016.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. ICML, 2017.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proc. AAAI, 2018.

[17] Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[18] E. Androulaki et al., "Hyperledger Fabric: A distributed operating system for permissioned blockchains," in Proc. EuroSys, 2018.

[19] Putz, F. Pernul, and G. Kablitz, "A secure and auditable logging infrastructure based on a permissioned blockchain," Computers & Secu-rity, vol. 87, 2019.