Related Work Survey: Fair Contextual Bandits for Equitable Diagnostic Decision-Making Under Missing Context

Piter Z. Garcia Bautista
MS Data Science / Bioinformatics
Rochester Institute of Technology
pizg8794@g.rit.edu

Daniel Krutz, Travis Desell
Department of Software Engineering, Data Science
Rochester Institute of Technology
dxkvse@g.rit.edu, tjdvse@g.rit.edu

I. Introduction

This project develops a practical, reproducible framework for fairness-aware contextual multi-armed bandits (CMAB/iCMAB) to make sequential decisions under uncertainty and limited resources while treating algorithmic fairness as a first-class objective. The work focuses on measuring and improving fairness across two domains—clinical diagnostic-like decision workflows and quantum-network routing—by evaluating the same policy stack in a simulation-first diagnostic environment and in a quantum-network routing simulator, with time-evolving disparity reporting and at least one mitigation mechanism.

This survey reviews prior work needed to support this framework, including bandit foundations, contextual/linear bandits and evaluation, learning under distribution shift and missing/partial context (including group-dependent missingness/measurement noise), and fairness definitions and fairness in bandit-style learning. It also includes domain grounding for clinical disparities and quantum-network routing. The conclusion summarizes gaps and clarifies how the proposed work differentiates itself by explicitly tracking and mitigating disparities over time while testing transfer across both domains.

II. Background

Many real-world workflows in both clinical diagnostics and quantum-network routing require sequential choices under uncertainty, resource constraints, and distribution shift. These settings can amplify inequities when some groups systematically have lower-quality context or different error profiles. This survey frames these workflows as a contextual bandit problem: at each step, choose an action (an "arm") given observed context to maximize utility while controlling fairness gaps.

A. Multi-armed bandits and contextual bandits

In both domains, multi-armed bandits formalize online decision-making with exploration–exploitation tradeoffs: at each round t an agent chooses an action $a_t \in \{1, \dots, K\}$ and observes reward r_t only for the chosen arm (bandit feedback). Performance is measured by *cumulative regret*:

$$R_T = \sum_{t=1}^{T} \big(\mu^* - \mu_{a_t}\big) \tag{1}$$

where $\mu^* = \max_k \mu_k$ is the optimal arm mean and μ_{a_t} is the mean reward of the chosen arm. A policy is desirable if $R_T = o(T)$, meaning the per-step loss vanishes over time.

Contextual bandits extend this setting by conditioning action choice on a context vector $\mathbf{x}_t \in \mathbb{R}^d$ observed before each decision. The LinUCB algorithm [1] maintains a ridge regression estimate $\hat{\boldsymbol{\theta}}_a$ per arm and selects actions via:

$$a_t = \arg\max_a \Big(\mathbf{x}_t^\top \hat{\boldsymbol{\theta}}_a + \alpha \sqrt{\mathbf{x}_t^\top \mathbf{A}_a^{-1} \mathbf{x}_t} \Big) \tag{2}$$

where $\mathbf{A}_a$ accumulates outer products of context vectors and α controls exploration. Contextual policies can improve sample-efficiency and stability when the context is predictive of outcomes; however, when context is missing or systematically noisier for some groups, these policies can inadvertently encode and amplify measurement inequities.

B. Fairness in sequential decision-making

Fairness in supervised learning is commonly operationalized through group-based error disparities. Let $Y \in \{0, 1\}$ be a binary outcome, $\hat{Y}$ a predicted label, and $A \in \{0, 1\}$ a protected group attribute. *Equality of opportunity* [2] requires equal true positive rates across groups:

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) \;=\; P(\hat{Y} = 1 \mid Y = 1, A = 1) \tag{3}$$

Equalized odds additionally requires equal false positive rates. In a bandit setting, fairness must be tracked over time under partial feedback and changing conditions, making it important to monitor time-evolving disparities (e.g., FNR/FPR gaps) (not only aggregate averages) and to apply mitigation during the policy update loop rather than only post-hoc.

C. Domain grounding: diagnostics and quantum routing

In the quantum-network routing domain, routing decisions are constrained by probabilistic link success, limited entanglement/quantum-resource availability, and time-varying congestion; context may include link-quality estimates and network load/queue signals, but these signals can be delayed, noisy, or partially observed. In this project, service equity is defined as parity in routing outcomes (e.g., success probability/latency parity across flow groups) under these constraints. In clinical diagnostics, context may include patient features, test and sample-quality indicators, and operational constraints, but access to context can be incomplete or systematically noisier for some populations. These limitations create performance and fairness risks, especially when optimizing aggregate performance, which can hide subgroup error spikes (e.g., false-negative gaps) unless the system is explicitly monitored and constrained. Unlike many evaluations that report utility-only bandit performance or post-hoc fairness for static predictors, the proposed work makes the time-evolving utility–fairness tradeoff explicit and tests transfer across both domains.

III. Related Work

Table I gives a compact map of the works reviewed and how they support the proposed project.

A. Bandit foundations and core algorithms

1) Upper confidence bound (UCB) analysis: Auer et al. [3] provide a foundational finite-time regret analysis for multi-armed bandits and introduce UCB-style algorithms that choose actions using optimistic confidence bounds, achieving gap-dependent regret logarithmic in T under stochastic assumptions (see Eq. 1). A limitation for the proposed work is that standard UCB assumes stationarity and does not incorporate context or group-based fairness constraints; extensions are required for distribution shift, heterogeneous contexts, and explicit disparity monitoring over time.

2) Thompson sampling: Russo et al. [4] survey Thompson sampling as a Bayesian approach to exploration–exploitation, highlighting strong empirical performance and the conceptual simplicity of sampling from a posterior over rewards. Thompson sampling is attractive for the proposed project as a baseline in both domains because it adapts to uncertainty naturally and often performs well under limited feedback. However, typical implementations focus on utility/regret and do not provide fairness guarantees; when context is missing or systematically noisy for some groups, posterior updates can encode these measurement inequities unless the model explicitly accounts for them.

3) Bandit textbook perspective: Lattimore and Szepesvári [5] provide a comprehensive reference on bandit algorithms (stochastic, adversarial, contextual, linear), including proof techniques and practical design choices. This text grounds algorithm selection and ablation design for the proposed project. A practical limitation is that textbook settings generally assume correctly observed context and utility-centric objectives; fairness-aware objectives and missingness mechanisms must be layered on top.

B. Contextual and linear bandits in practice

1) Epoch-Greedy contextual bandits: Langford and Zhang [6] introduce the Epoch-Greedy algorithm, one of the first contextual bandit algorithms with provable regret bounds. Their approach interleaves exploration epochs (uniform arm selection for reward estimation) with exploitation epochs (applying the current best policy), and shows that any supervised learning oracle can serve as the reward estimator. This motivates the use of offline regression oracles within online bandit policies—a strategy relevant to the iCMAB-style informed policy in the proposed project. A limitation is the dependence on epoch length, which is often unknown in practice.

2) Contextual bandits for personalization: Li et al. [1] demonstrate contextual bandits for news recommendation using LinUCB (Eq. 2), showing how context enables personalized decisions while balancing exploration and exploitation in a real deployment. The key limitation in the proposed setting is that context can be incomplete and systematically lower-quality for some populations, leading to performance disparities if policies are optimized only for average reward.

3) Linear stochastic bandits: Abbasi-Yadkori et al. [7] analyze linear stochastic bandits and provide improved confidence ellipsoids for the unknown parameter vector $\boldsymbol{\theta}$, achieving regret bounds that scale as $\tilde{O}(d\sqrt{T})$ under standard stochastic assumptions. Linear contextual bandits are a key baseline for the proposed work because they are simple, interpretable, and offer a clean place to inject missingness-handling and fairness-aware penalties. A limitation is the linear modeling assumption: if the relationship between context and outcome differs across groups or shifts over time, the model can underfit and hide subgroup-specific error spikes.

4) Scalable contextual bandits: Agarwal et al. [8] propose a computationally efficient reduction-based approach to contextual bandits that makes it practical to use rich policy classes. This is relevant to the proposed project as a template for scaling beyond simple linear models while keeping training tractable. However, the work primarily optimizes reward and assumes representative context; fairness auditing under group-dependent context quality requires additional monitoring and constraints.

5) Offline evaluation with bandit feedback: Dudík et al. [9] develop doubly robust estimators for policy evaluation and learning from bandit feedback, combining direct reward modeling with inverse-propensity weighting. This is useful for the proposed project because it supports evaluation and comparison of policies under partial feedback. A limitation is that estimator validity depends on correct propensities and sufficient support in logged data; if groups have different action distributions or context missingness rates, variance and bias can differ by group, complicating fairness assessment.

C. Distribution shift and missing/partial context

1) Non-stationary bandits: UCB with change detection: Garivier and Moulines [10] study UCB-style policies for non-

TABLE I
HIGH-LEVEL OVERVIEW OF RELATED WORKS REVIEWED IN THIS SURVEY.

Category	Representative works
Bandit foundations (exploration–exploitation basics)	UCB analysis [3]; Thompson sampling tutorial [4]; bandit textbook [5]
Contextual/linear bandits (side information + evaluation)	Epoch-Greedy [6]; contextual bandits in practice [1]; linear bandits [7]; scalable CB [8]; offline evaluation [9]
Shift / partial context (missingness + nonstationarity)	Non-stationary UCB [10]; non-stationary rewards [11]; restricted context [12]; unobserved contexts [13]
Fairness (definitions + constraints in sequential learning)	Equality of opportunity [2]; fairness in bandits [14]; fairness in RL [15]; offline fairness guarantees [16]; fair contextual MABs [17]; clinical bias case study [18]
Quantum networking (routing under uncertainty + resource scarcity)	Quantum internet vision [19]; entanglement routing [20]; optimal quantum routing [21]

stationary bandit problems, motivating mechanisms such as discounting or sliding windows to adapt to reward changes. This is relevant to both domains: patient mix and operational processes can change, and network load and link conditions can vary over time. The limitation for the proposed work is that non-stationary regret does not directly capture fairness; even if a policy adapts quickly in aggregate, it may exhibit persistent group disparities if context quality differs systematically across groups.

2) Non-stationary rewards and variation budgets: Besbes et al. [11] establish minimax-optimal rates for regret in non-stationary bandit problems where arm means change subject to a total variation budget, and propose a sliding-window UCB variant. A key finding is that ignoring non-stationarity can lead to linear regret. The proposed project incorporates distribution shift as a first-class experimental variable, testing bandit policies under both gradual and abrupt shifts across both testbeds.

3) Restricted context in contextual bandits: Bouneffouf et al. [12] consider contextual bandits when only a subset of context features can be observed at decision time, directly aligned with the proposed project's focus on missing or unevenly measured context. The limitation is that restricted-context selection is typically optimized for reward; extending the selection mechanism to also reduce group disparities is not standard and is a key innovation opportunity.

4) Bandits with unobserved contexts: Park and Faradonbeh [13] address bandit control when contexts are unobserved, proposing algorithms that learn to act effectively despite latent context. This motivates algorithm designs that remain robust when the observed context is incomplete or unreliable. A limitation is that fairness considerations are not central: if latent-context recovery quality differs across groups due to unequal measurement, fairness risks persist without explicit disparity tracking and mitigation.

D. Fairness: definitions and fairness in bandit learning

1) Equality of opportunity and error-based fairness: Hardt et al. [2] formalize equality of opportunity and related group-based fairness criteria (Eq. 3) in supervised learning, emphasizing that fairness can be defined in terms of error rates conditioned on true labels (e.g., equalizing false negative rates across groups). These definitions provide a clear measurement target for the clinical domain, where false-negative disparities have high stakes. A limitation for the proposed project is that supervised-learning fairness is typically evaluated post-hoc on static predictors; sequential decision-making adds partial feedback and time dynamics, so disparities should be measured as a function of time and policy behavior.

2) Fairness in classic and contextual bandits: Joseph et al. [14] study fairness constraints in bandit learning and analyze how fairness requirements change achievable regret. This work is central to the proposed project because it directly links bandit exploration policies to fairness constraints and highlights that naive utility optimization can be incompatible with fairness goals. A limitation is that many fairness formulations in bandits rely on simplified assumptions; the proposed project must operationalize group definitions and disparity metrics differently for diagnostics (error gaps) and quantum routing (service equity).

3) Fairness in reinforcement learning: Jabbari et al. [15] study fairness in reinforcement learning, requiring that a policy never takes an action that is unfair to any individual in expectation, and provide algorithms with polynomial sample complexity. The key challenge they identify is that fairness constraints can slow learning when group-specific feedback is sparse—directly relevant to the clinical diagnostics domain, where some patient groups may have systematically fewer observations, making it harder to estimate group-wise error rates accurately.

4) Offline contextual bandits with high-probability fairness: Metevier et al. [16] propose a method for offline contextual bandits that provides high-probability guarantees that a deployed policy satisfies user-specified fairness constraints, using importance-weighted estimators and Seldonian optimization. A significant limitation is the offline setting assumption: the algorithm requires a fixed logged dataset and does not address online fairness under distribution shift. The proposed project extends this direction to the online setting, where fairness constraints must be maintained continuously as the policy learns from streaming data.

5) Fair contextual multi-armed bandits: Chen et al. [17] develop algorithms for fair contextual MABs that satisfy group fairness constraints (parity in arm-selection rates across

groups) while maintaining sublinear regret, and characterize the utility–fairness tradeoff as a function of group distribution imbalance. This is the most directly related prior work to the proposed project. A key limitation is that their fairness criterion focuses on arm-selection parity rather than outcome parity (e.g., FNR/FPR gaps), which may not capture the clinically or operationally relevant disparities. The proposed project uses outcome-based fairness metrics that are more directly interpretable in clinical and service-equity contexts, and extends the evaluation to distribution shift and cross-domain transfer.

6) Clinical bias case study motivating fairness audits: Obermeyer et al. [18] analyze a widely used health-risk prediction system and show how proxy targets can encode racial bias, producing systematic under-allocation of care for Black patients at the same predicted risk. This motivates the proposed project's emphasis on fairness auditing and on monitoring subgroup-specific errors rather than relying on aggregate metrics. A limitation is that this work addresses static prediction/triage rather than sequential decision policies; however, the mechanism is relevant: optimizing a proxy objective under biased measurements can yield persistent disparities unless explicitly corrected.

E. Quantum networking and routing as a second testbed

1) Quantum internet: vision and engineering constraints: Wehner et al. [19] describe the vision for a quantum internet and outline key engineering challenges, including entanglement generation, storage, and networking protocols. This work provides domain background and motivates why routing decisions in quantum networks face uncertainty and resource scarcity. A limitation is that fairness is not a primary focus; the proposed project extends this domain by explicitly considering service equity across user/flow groups when learning routing policies under uncertainty.

2) Routing entanglement in the quantum internet: Pant et al. [20] develop methods for routing entanglement in quantum networks, addressing how to select paths and manage entanglement distribution under probabilistic link success. This motivates a concrete quantum routing decision process suitable as a second testbed. The limitation for the proposed project is that routing work is typically evaluated on throughput/fidelity metrics without group-based disparity monitoring; the proposed work adds explicit service-equity tracking using a shared fairness-aware contextual MAB evaluation stack.

3) Optimal routing for quantum networks: Caleffi [21] formulates quantum path selection as an optimization over link fidelity, decoherence time, and entanglement generation rate, showing that shortest-path routing significantly underperforms optimal routing due to the stochastic nature of link success. A limitation is the assumption of a static network with known link statistics; in practice, link quality varies dynamically. The proposed project extends the routing problem to an online learning setting using contextual bandits, where link-quality context is observed but statistics are unknown and shift over time.

IV. Conclusion

Prior work provides strong foundations for multi-armed bandits, contextual/linear bandits, and evaluation under bandit feedback, as well as clear fairness definitions and initial work on fairness constraints in bandit learning. However, a gap remains at the intersection of: (i) missing or unevenly measured context, (ii) time-varying conditions, and (iii) explicit, time-evolving disparity monitoring and mitigation in sequential decision systems.

The proposed project differentiates itself by building a reusable framework that evaluates fairness-aware contextual MAB policies under controllable missing-context mechanisms and distribution shift, tracks group disparities over time, and tests transfer across two distinct domains: clinical diagnostic-like decision-making and quantum-network routing. While Chen et al. [17] address fair contextual MABs and Joseph et al. [14] address fairness in bandit learning, to our knowledge existing work does not jointly evaluate outcome-based disparity metrics (FNR/FPR gaps) under distribution shift and test transfer across fundamentally different domains. This two-domain evaluation demonstrates that the methodology is generic—not tuned to a single application—while remaining grounded in domain-realistic constraints.

Overleaf (editable): https://www.overleaf.com/3621132125pfjhvkxzsbkp#0cab76

References

[1] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 661–670.

[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2–3, pp. 235–256, 2002.

[4] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on Thompson sampling," *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

[5] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[6] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *NeurIPS*, 2007.

[7] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011.

[8] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[9] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[10] A. Garivier and E. Moulines, "On UCB policies for non-stationary bandit problems," *arXiv preprint arXiv:0805.3415*, 2008.

[11] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *NeurIPS*, 2014.

[12] D. Bouneffouf, I. Rish, G. Cecchi, and R. Féraud, "Context attentive bandits: Contextual bandit with restricted context," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[13] D. Park and M. S. Faradonbeh, "Efficient algorithms for learning to control bandits with unobserved contexts," *arXiv preprint arXiv:2210.15668*, 2022.

[14] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," in *Advances in Neural Information Processing Systems*, 2016.
[15] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in reinforcement learning," in *ICML*, 2017.
[16] B. Metevier, S. Giguere, S. Brockman, A. Kobren, Y. Brun, E. Brunskill, and P. Thomas, "Offline contextual bandits with high probability fairness guarantees," in *NeurIPS*, 2019.
[17] X. Chen, A. Zheng, Z. Zhou, and N. B. Shah, "Fair contextual multi-armed bandits: Theory and experiments," in *UAI*, 2020.
[18] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
[19] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, p. eaam9288, 2018.
[20] M. Pant, H. Krovi, D. Englund, L. Gyongyosi, M. Babroli *et al.*, "Routing entanglement in the quantum internet," *npj Quantum Information*, vol. 5, no. 25, 2019.
[21] M. Caleffi, "Optimal routing for quantum networks," *IEEE Access*, vol. 5, pp. 22 299–22 312, 2017.