Related Work Survey: Fairness-Aware Contextual Bandits Under Missing Context

(Clinical Diagnostics and Quantum-Network Routing)

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

Dr. Daniel Krutz, Travis Desell
Department of Software Engineering, Data Science
DSCI 601 Project Advisors
dxkvse@g.rit.edu, tjdvse@g.rit.edu

1 Introduction

This survey reviews prior work needed to support a project on fairness-aware contextual multi-armed bandits (contextual MABs) for sequential decision-making when context can be delayed, noisy, or missing for some populations. The motivating applications span two domains: (i) clinical diagnostic-like workflows where sequential choices (tests, models, retesting, or pipeline selection) are made under limited resources and uneven context quality across patient groups, and (ii) quantum-network routing, where routing decisions are made under probabilistic link success and time-varying congestion with scarce quantum resources.

The related work is organized into: foundational bandit algorithms; modern contextual/linear bandits and evaluation; bandits under non-stationarity and missing/partial context; algorithmic fairness definitions and fairness in bandit-style learning; and domain grounding for clinical disparities and quantum networking/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.

2 Background

2.1 Multi-armed bandits and contextual bandits

Multi-armed bandits formalize sequential decision-making with an exploration–exploitation tradeoff: at each round an agent chooses an action (arm) and observes reward only for the chosen arm (bandit feedback). Contextual bandits extend this setting by conditioning the action choice on observed side information (features), which can improve sample-efficiency when the context is predictive of outcomes.

2.2 Fairness in sequential decision-making

Fairness in supervised learning is commonly operationalized through group-based error disparities (e.g., equal opportunity and equalized odds), which compare error rates across protected groups. In a bandit setting, fairness must be tracked over time under partial feedback and changing conditions, making it important to monitor time-evolving disparities (not only aggregate averages).

2.3 Domain grounding: diagnostics and quantum routing

In clinical diagnostics, "context" can include patient characteristics and sample/test quality indicators, but these may be unevenly measured across groups. In quantum networking, routing decisions must account for probabilistic link success, scarce entanglement/quantum-resource availability, and congestion; signals used as routing context (link-quality estimates, queue/load indicators) may be delayed, noisy, or partially observed.

3 Overview of Reviewed Works

Table 1 gives a compact map of the works reviewed and how they support the proposed project.

Category	Representative works
Bandit foundations	UCB analysis [1]; Thompson sampling tutorial [2]; bandit textbook [3]
Contextual/linear bandits	Contextual bandits in practice [4]; linear bandits [5]; scalable CB [6]; offline evaluation [7]
Shift / partial context	Non-stationary UCB [8]; restricted context [9]; unobserved contexts [10]
Fairness	Equality of opportunity [11]; fairness in bandits [12]; clinical bias case study [13]
Quantum networking	Quantum internet vision [14]; entanglement routing [15]

Table 1: High-level overview of related works reviewed in this survey.

4 Related Work Descriptions

4.1 Bandit foundations and core algorithms

4.1.1 Upper confidence bound (UCB) analysis

Auer et al. [1] provide a classic finite-time regret analysis for multi-armed bandits and introduce UCB-style algorithms that choose actions using optimistic confidence bounds. The approach formalizes how exploration can be driven by uncertainty estimates, and it yields logarithmic regret in stationary stochastic settings. 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.

4.1.2 Thompson sampling

Russo et al. [2] 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.

4.1.3 Bandit textbook perspective

Lattimore and Szepesvári [3] provide a comprehensive reference on bandit algorithms (stochastic, adversarial, contextual, linear), including proof techniques and practical design choices. This text is useful for the proposed work as a grounding reference for algorithm selection and for structuring ablations (e.g., how reward models, regularization, and confidence construction change performance). A practical limitation is that textbook settings generally assume that the context is correctly observed and that objectives are utility-centric; fairness-aware objectives and missingness mechanisms must be layered on top.

4.2 Contextual and linear bandits in practice

4.2.1 Contextual bandits for personalization

Li et al. [4] demonstrate contextual bandits for news recommendation, using context to personalize article selection while balancing exploration and exploitation. This work shows how contextual bandits can be deployed in real decision pipelines, and it motivates the use of contextual policies as a general mechanism for sequential decisions under uncertainty. The key limitation in the proposed setting is that the context in many real domains

is incomplete and can be systematically lower-quality for some populations, which can lead to performance disparities if policies are optimized only for average reward.

4.2.2 Linear stochastic bandits

Abbasi-Yadkori et al. [5] analyze linear stochastic bandits and provide improved algorithms with regret bounds by maintaining confidence sets for linear models. Linear contextual bandits (e.g., LinUCB-style policies) are an important baseline for the proposed work because they are simple, sample-efficient, and interpretable, and they offer a clean place to inject missingness-handling and fairness-aware penalties/constraints. A limitation is the linear modeling assumption: if the relationship between context and outcome differs across groups or shifts over time, a single linear model can underfit and hide subgroup-specific error spikes.

4.2.3 Scalable contextual bandits

Agarwal et al. [6] 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 because it provides a template for scaling contextual policies beyond simple linear models while keeping training tractable. However, the work primarily optimizes reward and assumes that logged data and context are representative; fairness auditing and mitigation under group-dependent context quality require additional monitoring and constraints.

4.2.4 Offline evaluation with bandit feedback

Dudík et al. [7] 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, and it can help structure experiments where only chosen actions produce outcomes. A limitation is that estimator validity depends on correct propensities and sufficient support in logged data; if groups have different action distributions or different context missingness rates, variance and bias can differ by group, complicating fairness assessment.

4.3 Distribution shift and missing/partial context

4.3.1 Non-stationary bandits

Garivier and Moulines [8] study UCB-style policies for non-stationary bandit problems, motivating mechanisms such as discounting or sliding windows to adapt to 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 still exhibit persistent group disparities if context quality differs systematically across groups.

4.3.2 Restricted context in contextual bandits

Bouneffouf et al. [9] consider contextual bandits when only a subset of context features can be observed or used at decision time. This framing is directly aligned with the proposed project's focus on missing or unevenly measured context: it formalizes that context is costly or partially available. The limitation is that restricted-context selection is typically optimized for reward; extending the selection mechanism to also reduce group disparities (e.g., by prioritizing features that stabilize subgroup performance) is not standard and is a key opportunity for innovation.

4.3.3 Bandits with unobserved contexts

Park and Faradonbeh [10] address bandit control when contexts are unobserved, proposing algorithms that learn to act effectively despite latent context. This work is relevant because it motivates algorithm designs that remain robust when the observed context is incomplete or unreliable, and it suggests modeling strategies for informative-context bandits. A limitation is that fairness considerations are not central: if latent-context recovery quality differs across groups (because of unequal measurement), fairness risks can persist without explicit disparity tracking and mitigation.

4.4 Fairness: definitions and fairness in bandit learning

4.4.1 Equality of opportunity and error-based fairness

Hardt et al. [11] formalize equality of opportunity and related group-based fairness criteria 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.

4.4.2 Fairness in classic and contextual bandits

Joseph et al. [12] study fairness constraints in bandit learning and analyze how fairness requirements can 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 or require careful definitions of "groups" and comparability; the proposed project must operationalize group definitions and disparity metrics differently for diagnostics (error gaps) and quantum routing (service equity).

4.4.3 Clinical bias case study motivating fairness audits

Obermeyer et al. [13] 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.

4.5 Quantum networking and routing as a second testbed

4.5.1 Quantum internet: vision and engineering constraints

Wehner et al. [14] 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.

4.5.2 Routing entanglement in the quantum internet

Pant et al. [15] develop methods for routing entanglement in quantum networks, addressing how to select paths and manage entanglement distribution. This work motivates a concrete quantum routing decision process that can be treated as a sequential decision environment, making it 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 and uses a shared fairness-aware contextual bandit evaluation stack across domains.

5 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. This two-domain evaluation is intended to demonstrate that the methodology is generic (not tuned to a single application) while still being grounded in domain-realistic constraints.

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

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

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

[9] 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.

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

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

[12] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," in *Advances in Neural Information Processing Systems*, 2016.

[13] 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.

[14] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, p. eaam9288, 2018.

[15] 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.