

Designing an Interpretable Interface for Contextual Bandits

Andrew Maher*, Matia Gobbo, Lancelot Lachartre, Subash Prabanantham, Rowan Swiers and Puli Liyanagama

Metica, London

Abstract

Contextual bandits have become an increasingly popular solution for personalized recommender systems. Despite their growing use, the interpretability of these systems remains a significant challenge, particularly for the often non-expert operators tasked with ensuring their optimal performance. In this paper, we address this challenge by designing a new interface to explain to domain experts the underlying behaviour of a bandit. Central is a metric we term “value gain”, a measure derived from off-policy evaluation to quantify the real-world impact of sub-components within a bandit. We conduct a qualitative user study to evaluate the effectiveness of our interface. Our findings suggest that by carefully balancing technical rigour with accessible presentation, it is possible to empower non-experts to manage complex machine learning systems. We conclude by outlining guiding principles that other researchers should consider when building similar such interfaces in future.

Keywords

User interfaces for decision-making, Contextual bandits, Off-policy evaluation, Interpretable machine learning,

1. Introduction

Complex personalized recommender systems have become vital to building engaging, modern user experiences across a variety of domains [1–3]. Although powerful, these systems cannot properly function without a human operator in place who can deploy and manage their correct running. These people – typically non-experts in statistics and machine learning – are expected to make reasoned, higher-order decisions about the recommender system, to maximize its performance and ensure it adds holistic value to the broader environment in which it sits.

By default, however, modern recommender systems are complex and hard to interpret [4, 5]. They comprise multiple interlocking parts, each of which requires a strong mathematical background to understand. Take contextual bandits. They are an increasingly popular methodological approach that address known challenges such as the cold-start problem [6] and non-stationary environments [7, 8]. Despite their efficacy as a recommender system method, providing a robust interpretation to their decisions is an unsolved problem. Similar to traditional supervised learning systems, they depend on black-box models to estimate the expected performance of recommendable items given a context. This difficulty is compounded by several factors: observational data only becomes available for the items selected by the bandit; interpretation is required not only for a single output but for multiple items from which the bandit is choosing; the ongoing modulation between exploration and exploitation means a bandit system does not always select the arm it predicts as most valuable.

For the non-expert human operator, several higher-order considerations need to be made to ensure each bandit-based recommender is continually well-tuned. Is it performing well enough to keep it in production? Should new arms be added, or existing ones removed? Are the context fields considered by the bandit sufficiently discriminatory to yield interesting results? Answering these questions requires an understanding of the underlying system that is both deep and broad. Yet there is a gap between this need for interpretation and the tools and interfaces that exist to provide it.

IntRS'24: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, October 18, 2024, Bari (Italy)

*Corresponding author.

✉ andrew@metica.com (A. Maher); matia@metica.com (M. Gobbo); lancelot@metica.com (L. Lachartre); subash@metica.com (S. Prabanantham); rowan@metica.com (R. Swiers); puli@metica.com (P. Liyanagama)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

A comparison with a sister decision-making domain is apt. In A/B testing, there is a well-known set of metrics and visualizations that determine which of the options being tested is best [9, 10]. Statistical significance and MLE-based uplift charts predominate the field. Moreover, there exist dozens of commercially available platforms with interfaces designed for easy interpretation [11], and probably thousands of proprietary in-house equivalents [12, 13]. Few such equivalent interfaces exist for contextual bandits; none are publicly available.

To address these challenges, we have developed an intuitive interface designed to explain the behavior of a contextual bandit system. It is in production and being used in a commercial setting. By leveraging techniques from data visualization, off-policy evaluation and user-centric design, our interface aims to make the inner workings of a bandit system comprehensible to domain-expert operators. Central to the interface is a generic metric framework we term "value gain" – a measure derived from off-policy evaluation that provides a clear indication of the real-world value of different elements of the system.

It is important to clarify that our target audience is not the end-users who receive recommendations; significant research has already been conducted on boosting interpretability for these people [4, 14–17]. We focus instead on the people who choose the inputs to the recommender system, and who are responsible for its proper functioning. We assume nothing about their knowledge of statistics, only that they can read a quantitative dashboard. As well, it is worth noting that although the interface has been designed for contextual bandits, and works with any underlying bandit algorithm, the ideas apply equally well to any recommender methodology. The sole requirement is that the system comprise a (relatively) limited number of options from which to choose – and that there is value in understanding their respective performance. We do not try to solve for the problem of choosing from an ever-changing and very large library of options (as in, say, a video recommender system).

The rest of the paper is organised as follows. In Section 2 we discuss related work. We then present our developed interface in Section 3, followed by a user study in Section 4 to evaluate its effectiveness. Finally, in Section 5 we outline guiding principles for future practitioners looking to build similar dashboards, then discuss future directions in this space.

2. Related work

Contextual Bandits can be viewed as an extension of traditional experimentation in which the arm-assignment decision is both automated (hence bandit) and personalized (hence contextual). Typically, they comprise a reward model and a policy. The former governs the bandit’s understanding of the world, with common choices including linear regression [18] and neural networks [19]. The latter dictates how the bandit modulates between exploration and exploitation. Canonical examples are UCB (Upper Confidence Bound) [20], and Thompson sampling [21], with numerous applications across various domains such as online advertising [22], personalized news feeds [23], customer support [24], and e-commerce recommendations [25, 26].

Off-Policy Evaluation is the main paradigm through which the efficacy of different bandit approaches is evaluated. It is a counterfactual estimation procedure in which the logged policy – the one for which real-world data is observed and measured – is compared to a hypothetical target policy. Numerous estimators exist to facilitate this comparison, including inverse propensity scoring, the direct method, and doubly robust estimators [27–29]. These techniques allow for the assessment of new policies without the need for costly and time-consuming online experimentation.

In contrast to the "how good," *Explainable AI* attempts to elucidate the purer "how" and "why" of machine learning approaches [30]. Recent research in this field has focused on identifying new methods, different presentation approaches, and better ways to judge the goodness of these explanations. The vast majority focus on interpretability concerns with respect to the recipient of a recommendation, something that is out of the scope of this work. Of particular relevance are a number of *User Interfaces* designed to enable the proper understanding of different ML systems. For example, ActiVis and LSTMVis are two different visualisation interfaces for interpreting deep learning models and results [31, 32], and InterpetML is a holistic system to help understand ensembles of decision trees [33]. A number of similar

explanations and interfaces exist for more general reinforcement learning policies [34, 35]. Although many organisations deploy contextual bandits – and some offer them as services to other companies – we could not find any equivalent bandit interfaces in the literature.

3. Interface

Figure 1 shows the interface we designed, through which human operators can interpret the workings of a contextual bandit system. It contains multiple, ordered visualizations that aim to provide increasing detail about the performance of different aspects of the bandit.

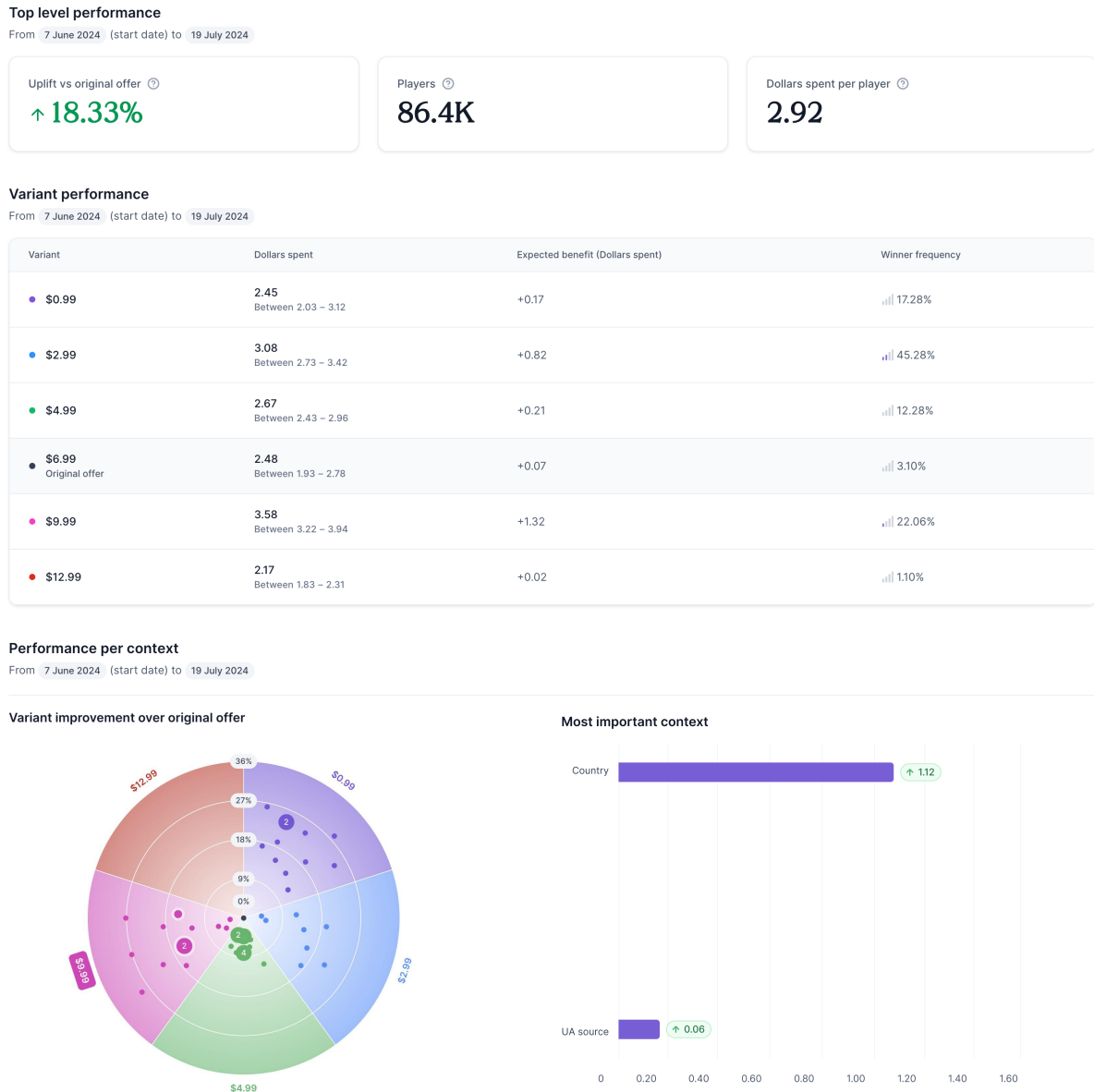


Figure 1: User interface for contextual bandits. The interface comprises three main components: top level performance, variance performance and performance per context. They each describe different elements of the performance of the bandit system, in increasing granularity of units.

The audience of the interface are the people responsible for launching – and potentially altering – the bandit. They need to understand not just its holistic performance, but also how each component of the bandit contributed to that performance. One way of providing this understanding is through comparison: supposing that component wasn't included in the system, how much less value would the

bandit generate? In other words, what is the value *gained* by the inclusion of that component.

To this end, we introduce the *value gain* metric – an estimate of the value of the production bandit, with respect to a simpler one in which certain components are ablated. Below, we define this metric in its general form. We then describe the interface itself in detail – considering as a prototypical example a use-case in which a mobile game wishes to increase dollars spent on in-app purchases. Although we consider this example for the purposes of the paper, the interface applies equally well to any use-case served by a contextual bandit.

3.1. Value gain

Let r denote the reward observed by the contextual bandit from one of its actions, and let π be the combination of policy-and-context-model that chooses actions yielding r . We can define the *value of π* as

$$v^\pi = \mathbb{E}_\pi [r]. \quad (1)$$

Here v is measured in the same units as the optimisation goal of the bandit. It describes, for example, the average revenue per user achieved by the bandit.

This policy-and-context-model π comprises multiple elements: the choice of exploration agent; the fields introduced to represent user context; the arms selected for inclusion in the bandit. Each of these individual components contributes to the value of the overall system. We attempt to measure that contribution through ablation. This is the principle behind the value gain metric.

Suppose we had an alternative policy-and-context-model $\bar{\pi}$. And further suppose that $\bar{\pi}$ is equivalent to π in all but some elements (which we'll denote as τ) – ie, its components are a subset of those present in π . We define the *value gain* – the gain in value produced by those missing elements – as being given by

$$g(\tau) = v^\pi - v^{\bar{\pi}} = \mathbb{E}_\pi [r] - \mathbb{E}_{\bar{\pi}} [r]. \quad (2)$$

It is impractical to calculate online $v^{\bar{\pi}}$ across the range of elements in which we might be interested. Doing so would require diverting a large proportion of traffic to alternative assignment algorithms, reducing the volume of data from which the main policy can learn and exposing multiple users to a potentially inferior bandit. Instead, we use methods introduced in the off-policy evaluation literature [28, 29]. Although these methods are not perfect, they nonetheless represent the gold standard by which bandit policies are evaluated offline.

In practice any off-policy estimator can be used. To make concrete the ideas within this paper, we consider the inverse-propensity score estimator [28]. It estimates $v^{\bar{\pi}}$ as

$$v^{\bar{\pi}} = \frac{1}{n} \sum_{i=1}^n \frac{\bar{\pi}(a|x_i)}{\pi(a|x_i)} r(x_i), \quad (3)$$

where the functions in the fraction represent the probability of assigning arm a to context x_i for the ablated and actual policy respectively.

For example: suppose an e-commerce website launched a bandit to improve conversion rate on its landing page, by launching different variants of its home page. And suppose further that one of the variants is the pre-existing home page, ie, the one that existed before the launch of the bandit. Here, $g(\tau)$ can be used to measure the value gained from including all non-baseline arms. In this case, τ denotes the set of non-baseline arms, $\bar{\pi}$ is a trivial policy that assigns the baseline homepage to all incoming traffic, and π is the contextual bandit launched by the e-commerce website. The value gain $g(\tau)$ shows the extra conversion rate introduced by the bandit with respect to the original baseline experience.

3.2. User interface

The interface in Figure 1 comprises three main sections:

1. **Top level performance:** Short summary of the bandit’s overall performance. It provides an overview that can be digested in as short a time as possible.
2. **Variante performance:** Description of how each arm in the bandit is performing. It enables an operator to determine which arms performs best, and which are candidates for removal from the system.
3. **Performance per context:** More granular information about the relationship between expected performance and the different contexts used within the bandit. It provides further understanding of the way in which the bandit is personalising the underlying experience.

3.2.1. Top level performance

The top level performance section presents three distinct metrics:

1. *Uplift vs original offer:* What is the percentage increase provided by the bandit on the goal metric, when compared to a baseline arm?
2. *Players:* How many people have been exposed to the bandit so far?
3. *Dollars spent per player:* What is the performance of the bandit on its revenue-maximising goal metric so far?

The metrics are explicitly presented in order of importance. First, how much value is the bandit adding its use-case? Second, is it reaching a sufficiently large population? Third, what is its average performance?

The first metric is based off the previously defined *value gain*. Here, we assume that one of the arms can be considered as a “baseline” – it is the arm that would be shown to all traffic, if the recommendation policy was not in use. In this case, we show how the policy performs relative to another one that contains only that baseline arm. Following the notation above – and assuming the production policy comprises the set of arms $a_k \in \mathcal{A}$, we calculate $g(\tau)$ for

$$\tau = \{a_k \in \mathcal{A} : a_k \neq a_{\text{baseline}}\}, \quad (4)$$

ie, the counterfactual policy contains only the baseline arm a_{baseline} .

Players and *Dollars spent per player* are calculated directly from logged assignment and rewards data.

3.2.2. Variant performance

The variant performance table shows arm-level information. It is a table comprising one row per arm and four columns. One of the columns describes the arm itself (with a name or other distinguishing information), the other three summarise performance information about that arm.

The first metric column – *Dollars spent* – shows both (A) the expected performance of the arm for all people exposed to the bandit, as well as (B) a range of potential performance values. We calculate the expectation and range by first estimating for a given its reward across all players. Then, on that distribution of estimates, we compute three summary statistics: the mean, and the 10th and 90th percentiles.

The second column is *Expected benefit (Dollars spent)*. It measures the achieved value that can be attributed to the arm in question. We again use *value gain* to evaluate this quantity. In this instance, we compare the production policy to a counterfactual one whose ablated element is $\tau = a_k$, where a_k is the arm being measured. Take variant \$0.99 as an example. It has an expected benefit of +0.17. This means the bandit gains 17 cents more per user thanks to its ability to show this arm.

The third column shows the proportion of contexts for which that arm was displayed to users.

3.2.3. Performance per context

The performance per context component contains two visualizations. The first is a radar chart in which a circle is split into multiple segments, with each segment representing a single arm. Dots are plotted on

to the segments. Each dot represents a distinct context vector encountered by the bandit. The dots are placed into the segment corresponding to the expected best arm for that context vector. Their distance from the chart’s origin is defined by the relative uplift of that arm compared to the original offer. We again use the *value gain* to calculate this distance.

The second visualisation is a bar chart shows the value gain attributable to each context field. Here we compare the production policy to a counterfactual one, in which the context field in question is removed. Each bar hence describes how much better is the bandit thanks to the inclusion of that context field.

4. User study

To better understand the ability of our interface to meaningfully represent interpretable results from a contextual bandit system, we conducted a user study.

Table 1

Results from the self-guided component of the user study.

Introduction to interview and page	All interviewees quickly discerned the meaning of the top-level performance metrics and how they would help in measuring performance.
Variant performance table	The variant performance table was the second element of the page at which they each arrived. All three understood the dollars spent and winner frequency columns, but needed some prompting with the latter. More difficult was the expected benefit column. Each interviewee correctly stated it denoted the value of the variant, and that the measure was comparing the variant to something else. No interviewee could state what that something else was. At first, each said it might be the baseline variant before deducing that to be impossible (as the baseline variant also had a non-zero value). Even after extensive prompting, they couldn’t correctly define the metric.
Radar chart	The radar chart was the last component of the page each interviewee discovered. They all found it somewhat daunting to explore at first, but quickly established (A) what each point represented and that (B) each segment related to an individual variant. Two of the interviewees stated the correct definition of the dots’ placement. All three worked out how to evaluate the different variants using the chart. One candidate noted the chart was pretty but potentially superficial.
Desire for contextualisation	Beyond the specific sub-components of the page, the three interviewees each expressed a desire for more context beyond the base numbers shown. All three explicitly requested information on the “significance” of the results. Two wanted to understand the number of observations relating to each number. One interviewee requested filters to gain more granular information about the data.

4.1. Format

As in [31] and [36], we performed a qualitative evaluation built off deep-dive interviews with candidates who would use this system as part of their daily workload. The interviews each took forty-five minutes. We started with a short explanation of how contextual bandits work, to ensure candidates had sufficient familiarity with the topic. Interviewees were then encouraged to explore the UI on their own and, whenever they focused on a specific component, they were asked to explain their perception of what it meant.

During the self-guided exploration, we also asked each interviewee the following specific knowledge-based questions to probe the extent of their ability to correctly interpret the bandit using the interface:

Table 2

Results from the knowledge-based questions in the user study.

Bandit value	All interviewees correctly stated the bandit provided value above the baseline. They relied only on the top-left uplift metric to make this point. When asked whether they'd let the bandit continue running, all three replied yes. They again depended on the top-left uplift metric.
Variant performance	All three interviewees could reason about which variants were worse-performing. They used a combination of the winner frequency, the radar chart and the expected benefit to answer – with no clear preference among these elements. They all determined a best set of variants (\$2.99 and \$9.99) using the same elements. None tried to contrast the quality of these two variants (using the expected benefit column and winner frequency, for example).
Context contribution	Two interviewees used the context contribution chart to reason that removing poor-performing context fields would improve the bandit (by avoiding opportunity cost and/or simplifying the system). One interviewee couldn't reason effectively about which contexts best contributed to the bandit. They instead lent on their practical experience (of user behaviour in different countries). They didn't try to use the context contribution chart for their answer.

- **Bandit value:** How is the bandit performing compared to a default experience? When do you think the optimization would/should stop?
- **Variant performance:** What are the best / worst performing variants? Why? Given the information presented, would you intervene to change anything about the variants? If so, what changes would you make?
- **Context contribution:** Given the information presented, would you intervene to change anything about the context fields being used? If so, what changes would you make?

4.2. Results

We conducted three such deep-dive interviews in total. Each interviewee was a marketing professional who would be the person responsible for interpreting bandit choices and outcomes to make operational decisions. All of them had extensive experience with A/B testing, but little practical background usage of a contextual bandit. Each interviewee was shown the interface as depicted in Figure 1. Results from the self-guided exploration and knowledge-based questions are summarised in Tables 1 and 2.

5. Conclusions

5.1. Guiding principles

We have presented a visual interface to explain the workings of an in-operation contextual bandit system, built using novel metrics underpinned by methods from off-policy evaluation.

Through this exercise, we can identify a number of broad, guiding principles to inform the useful design of a similar interfaces in future. These principles are outlined in Table 3. Future researchers and practitioners should use them to help direct their own design processes.

Of these, the two most crucial are the complementary pair: *Feel empowered to use technical tools / Use clear non-technical language*. In the context of machine learning and recommender systems, the most insightful metrics can be simple to understand but inherently complicated to calculate. If they add the most value – use them. For example, we introduced ideas from off-policy evaluation to our dashboard. Despite their relevance, off-policy evaluation is not particularly well-known outside the machine learning and statistics community. As we built the interface, we held concerns internally that our audience might feel uncomfortable with these metrics, and not trust them sufficiently. However none of our interviewees raised a problem.

Table 3

Guiding principles for an interpretable interface for a contextual bandit. Future researchers and practitioners should consider these when building their own, similar interfaces.

Principle	Explanation
Feel empowered to use technical tools	Sometimes the most relevant metrics are highly technical. Don't shy from their usage. We used techniques from off-policy evaluation, an alien field to our interviewees. No one raised a concern; our candidates trusted and accepted the information we shared.
Use clear, non-technical language	Describe results in a way the audience can easily reason about. We named one column "expected benefit", as it related to statistical expectation and indirectly conveyed meaning. The title made sense to us as statisticians; our interviewees didn't get it. By contrast, all candidates understood "uplift vs original".
Consciously order information	Different results exist within the hierarchy of complexity in a recommender system. Carefully consider what to show and when. Our interviewees could digest our most complex visualisation – an information-dense radar chart – precisely because they'd been carefully shown simpler results earlier on.
Contextualise results	A repeated criticism of our interface was a lack of statistical significance or volume information. People couldn't reason about the importance of the results they were seeing. Providing contextual information of this sort enables operators to respond proportionately to insights.
Facilitate decision-making	Fundamentally, insights are useful only if they lead to decision-making. Consider what will guide choices, and present it in complementary formats where useful. Our interviewees successfully answered our task questions by combining multiple elements of our interface. It was this action-oriented approach they praised most.

To caveat the above point, it's important to consider the audience that will read the results. Think about the implication the statistical machinery conveys, then use that as the description of the metric. Don't blindly use the jargon term that aligns best with the literature if an end-user will not understand its meaning – an error we made in our design.

5.2. Future work

In this paper, we explore ideas and designs to enable the useful interpretation of a single contextual bandit system containing a (relatively) limited number of meaningful options. Relaxing these two constraints – interpreting only a single bandit, and considering a much wider set of options – would require additional design considerations.

Take multiple bandits: our explorations here relate to providing a deep-dive on a narrow, single optimisation. This necessitates an abundance of information that becomes hard to parse when multiplied across use-cases. This is what would be found in practice – one bandit managing a search experience, another bandit the delivery of product details, and a third bandit the creatives to display. We've so far considered an initial solution of surfacing a single key metric for each bandit (the *uplift vs original offer* metric). Future research could improve on this by (A) more robustly determining which metric is the most salient to display, (B) ascertaining how to usefully triage among running bandits and (C) explaining the health of all systems in parallel.

Considering a much wider number of arms is another interesting technical challenge. Components we've introduced here – the radar chart, the table of performance per variant – do not naturally extend to the case where there is more than, say, twenty options. But modern contextual bandit systems, particularly with the advent of generative AI, can be easily designed to have much larger numbers of meaningful variants. Conveying information from across a multitude of potentially differing variants is something we hope to consider in future work.

References

- [1] M. De Nadai, F. Fabbri, P. Gigioli, A. Wang, A. Li, F. Silvestri, L. Kim, S. Lin, V. Radosavljevic, S. Ghael, D. Nyhan, H. Bouchard, M. Lalmas, A. Damianou, Personalized audiobook recommendations at spotify through graph neural networks, Association for Computing Machinery, New York, NY, USA, 2024. URL: <https://doi.org/10.1145/3589335.3648339>. doi:10.1145/3589335.3648339.
- [2] G. Tang, J. Pan, H. Wang, J. Basilico, Reward innovation for long-term member satisfaction, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 396–399. URL: <https://doi.org/10.1145/3604915.3608873>. doi:10.1145/3604915.3608873.
- [3] X. Liu, Z. Li, Y. Gao, J. Yang, T. Cao, Z. Wang, B. Yin, Y. Song, Enhancing user intent capture in session-based recommendation with attribute patterns, in: NeurIPS 2023, 2023. URL: <https://www.amazon.science/publications/enhancing-user-intent-capture-in-session-based-recommendation-with-attribute-patterns>.
- [4] D. Afchar, A. Melchiorre, M. Schedl, R. Hennequin, E. Epure, M. Moussallam, Explainability in music recommender systems, *AI Magazine* 43 (2022) 190–208.
- [5] H. Steck, L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, J. Basilico, Deep learning for recommender systems: A netflix case study, *AI Magazine* 42 (2021) 7–18.
- [6] H. T. Nguyen, J. Mary, P. Preux, Cold-start problems in recommendation systems via contextual-bandit algorithms, 2014. URL: <https://arxiv.org/abs/1405.7544>. arXiv:1405.7544.
- [7] J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, M. Ghavamzadeh, C. Boutilier, Non-stationary latent bandits, 2020. URL: <https://arxiv.org/abs/2012.00386>. arXiv:2012.00386.
- [8] C. Li, Q. Wu, H. Wang, Unifying clustered and non-stationary bandits, 2020. URL: <https://arxiv.org/abs/2009.02463>. arXiv:2009.02463.
- [9] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, D. G. Altman, Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations, *European journal of epidemiology* 31 (2016) 337–350.
- [10] D. J. Biau, B. M. Jolles, R. Porcher, P value and the theory of hypothesis testing: an explanation for new researchers, *Clinical Orthopaedics and Related Research* 468 (2010) 885–892.
- [11] A. Fabijan, P. Dmitriev, B. Arai, A. Drake, S. Kohlmeier, A. Kwong, A/b integrations: 7 lessons learned from enabling a/b testing as a product feature, in: 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), IEEE, 2023, pp. 304–314.
- [12] D. K. Vasthimal, P. K. Srirama, A. K. Akkinapalli, Scalable data reporting platform for a/b tests, in: 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 2019, pp. 230–238. doi:10.1109/BigDataSecurity-HPSC-IDS.2019.00052.
- [13] R. L. Kaufman, J. Pitchforth, L. Vermeer, Democratizing online controlled experiments at booking.com, 2017. URL: <https://arxiv.org/abs/1710.08217>. arXiv:1710.08217.
- [14] C.-H. Tsai, P. Brusilovsky, Evaluating visual explanations for similarity-based recommendations: User perception and performance, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 22–30.
- [15] C. Musto, G. Rossiello, M. de Gemmis, P. Lops, G. Semeraro, Combining text summarization and aspect-based sentiment analysis of users' reviews to justify recommendations, in: Proceedings of the 13th ACM conference on recommender systems, 2019, pp. 383–387.
- [16] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, J. Zhang, Chat-rec: Towards interactive and explainable llms-augmented recommender system, arXiv preprint arXiv:2303.14524 (2023).
- [17] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, Y. Zhang, Counterfactual explainable recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1784–1793.
- [18] W. Chu, L. Li, L. Reyzin, R. Schapire, Contextual bandits with linear payoff functions, in:

Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 208–214.

- [19] O. Nabati, T. Zahavy, S. Mannor, Online limited memory neural-linear bandits with likelihood matching, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 7905–7915. URL: <https://proceedings.mlr.press/v139/nabati21a.html>.
- [20] Y. Abbasi-Yadkori, D. Pál, C. Szepesvári, Improved algorithms for linear stochastic bandits, *Advances in neural information processing systems* 24 (2011).
- [21] S. Agrawal, N. Goyal, Thompson sampling for contextual bandits with linear payoffs, in: International conference on machine learning, PMLR, 2013, pp. 127–135.
- [22] B. Han, J. Gabor, Contextual bandits for advertising budget allocation, *Proceedings of the ADKDD* 17 (2020).
- [23] L. Li, W. Chu, J. Langford, 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.
- [24] S. Sajeev, J. Huang, N. Karampatziakis, M. Hall, S. Kochman, W. Chen, Contextual bandit applications in a customer support bot, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3522–3530.
- [25] D. N. Hill, H. Nassif, Y. Liu, A. Iyer, S. Vishwanathan, An efficient bandit algorithm for realtime multivariate optimization, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1813–1821.
- [26] X. HE, B. An, Y. Li, H. Chen, Q. Guo, X. Li, Z. Wang, Contextual user browsing bandits for large-scale online mobile recommendation, in: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 63–72. URL: <https://doi.org/10.1145/3383313.3412234>. doi:10.1145/3383313.3412234.
- [27] M. Farajtabar, Y. Chow, M. Ghavamzadeh, More robust doubly robust off-policy evaluation, 2018. [arXiv:1802.03493](https://arxiv.org/abs/1802.03493).
- [28] M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, *arXiv preprint arXiv:1103.4601* (2011).
- [29] Y.-X. Wang, A. Agarwal, M. Dudík, Optimal and adaptive off-policy evaluation in contextual bandits, in: International Conference on Machine Learning, PMLR, 2017, pp. 3589–3597.
- [30] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11 (2021) 1–45.
- [31] M. Kahng, P. Y. Andrews, A. Kalro, D. H. Chau, Activis: Visual exploration of industry-scale deep neural network models, *IEEE Transactions on Visualization and Computer Graphics* 24 (2018) 88–97. doi:10.1109/TVCG.2017.2744718.
- [32] H. Strobelt, S. Gehrmann, H. Pfister, A. M. Rush, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, *IEEE transactions on visualization and computer graphics* 24 (2017) 667–676.
- [33] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, *arXiv preprint arXiv:1909.09223* (2019).
- [34] A. Mishra, U. Soni, J. Huang, C. Bryan, Why? why not? when? visual explanations of agent behaviour in reinforcement learning, in: 2022 IEEE 15th Pacific Visualization Symposium (PacificVis), IEEE, 2022, pp. 111–120.
- [35] S. Milani, N. Topin, M. Veloso, F. Fang, Explainable reinforcement learning: A survey and comparative review, *ACM Computing Surveys* 56 (2024) 1–36.
- [36] E. Purificato, B. A. Manikandan, P. V. Karanam, M. V. Pattadkal, E. W. De Luca, Evaluating explainable interfaces for a knowledge graph-based recommender system., in: *IntRS@ RecSys*, 2021, pp. 73–88.