Value and Impact of Recommender Systems

Dietmar Jannach and Markus Zanker

Abstract Recommender systems are an integral part of many of today's major web sites and online services and they widely exert their influence on which content users come across in the online world. From an organizational and economic perspective, recommender systems are designed to increase the quality of matches between users and items or content and this way create value for the different stakeholders. They do so, for instance, by reducing consumers' information overload or by helping to improve business-oriented performance indicators. However, notwithstanding of the actual achievement of concrete goals or targets, recommender systems have an impact on users' choice behavior as a matter of principle.
In this chapter, we first review the various ways recommender systems create value for different stakeholders and also discuss the possible risks and the potentially negative impacts of using such systems. We particularly focus on the organizational and business-oriented perspective by reporting how practical systems are evaluated and which effects have been observed in real-world deployments. In the final section of this chapter we discuss the limitations of academic research practices and emphasize the need to expand our research approaches to be able to address problems that are highly relevant in practice but still under explored in academia.[1]

1 Introduction

Automated recommendations are nowadays part of many websites and online services, and they are often a central part of the overall user experience, such as on

Dietmar Jannach
University of Klagenfurt, Austria, e-mail: dietmar.jannach@aau.at

Markus Zanker
Free University of Bozen-Bolzano, Italy, e-mail: markus.zanker@unibz.it

[1] Cite as: Dietmar Jannach, Markus Zanker: "Value and Impact of Recommender Systems", in Ricci et al. (eds.), Recommender Systems Handbook, 3rd edition, Springer, 2022, pp. 519–546

e-commerce and media streaming sites. The content we come across in the online world is therefore often highly individualized. The applied personalization strategy is usually centrally determined by the recommendation service provider who tailors its own content to consumers.[2]

The challenge therefore consists of striking a balance between economic-oriented goals such as conversion rates, increased sales, or customer retention and keeping users happy with personalized and tailored offerings [41, 69]. A commonly mentioned goal of a recommender system for the latter aspect is therefore to help users find items of interest in situations of information overload. This is according to literature typically assessed by its capability to accurately predict the relevance of individual items for individual users. As Abdollahpouri et al. [1] thus noted, in the best case, a recommender system therefore creates value for the consumers and potentially even further stakeholders while being economically sustainable for the service provider.

Independent of such specific goals and their achievement, recommender systems influence the users' choices and their behavior as a matter of principle. In particular in cases where the recommendations are central to the user experience of the service, a large fraction of the observed interactions stems from recommendations. This is, for instance, the case for recommendations at YouTube or Netflix [19, 30]. These fractions can even grow due to self-enforcing cycles in case most of the content that users come across during browsing has already been pre-filtered and personalized in one way or another.

The use of recommender systems therefore also bears some risks and can potentially even lead to negative and undesired effects, such as reinforcing extremist or unhealthy behavioral and consumption patterns. For instance, a hypothesized radicalization pipeline leading users to more extremist video content at YouTube [63], or suggestions of a food recommender that would mostly relate to unhealthy dishes [22].

In this chapter, we review the possible values and the impact of recommender systems for different stakeholders. We consider both *organizational and business value* (e.g., in terms of increased sales or customer retention) as well as different types of *consumer value* (e.g., in the form of reduced search efforts or better decisions), and we describe additional ways in which recommenders may positively or negatively impact the behavior of consumers.

Next, in Section 2, we first discuss the different purposes a recommender can serve and in which ways such a system can thereby create value for different stakeholders. Furthermore, we sketch potential risks that can derive from the use of recommendation technology. Section 3 then discusses how the impact of recommender systems is currently measured, with a particular focus on measures that are used in practice. Finally, in Section 4, we reflect on our predominant research approaches in academia, where we often largely abstract from the aforementioned general goals and therefore are only partially capable of quantifying the potential impact. Based on this dis-

[2] This is termed a *provider-centric* strategy in [5].

cussion, we correspondingly outline potential ways towards more impact-oriented research methodologies.

2 Stakeholders and Value Drivers of Recommender Systems

While a recommender system is effective whenever it creates some *utility* or *value* for one of its stakeholders, the focus of the research literature lies mostly on demonstrating the potential value of recommender systems for *consumers*, i.e., for those receiving recommendations. Correspondingly, the extent of value creation is mostly assessed in terms of consumer-oriented measures, such as the assumed capability of a recommender system to help users finding the most relevant items in their particular situation. However, we argue that recommender systems are applications residing in the core of many e-commerce business models since they are primarily concerned about exploitation of information and matchmaking on virtual markets. Therefore, the impact and created value of recommender systems cannot only be seen from a pure consumer or user perspective, but there are effects on all involved stakeholders. For instance, depending on the pursued recommendation strategy not only the satisfaction of the receivers of recommendations, but also the profitability of the platform or the sales distribution among the items in the product catalog will be potentially impacted. Moreover, the strategy will also have long-term effects on the perception of the integrity and benevolence of the recommendation system itself [9], which in turn will moderate its impact. Therefore, after shortly describing the stakeholders of recommender systems, we will introduce the value driver model of e-commerce business models [6] to structure the discussion on the impact of recommender systems in this chapter.

2.1 Stakeholders of Recommender Systems

For every recommender system, there are at least two different types of stakeholders [1]. There are *(i)* the consumers or users who receive the recommendations, and *(ii)* the organization that provides the recommendations as part of their service. The differentiation between these stakeholders is important. While a recommender system should in the best case create value in parallel for all involved stakeholders, there might be competing goals involved in the process. In other words, the best recommendation for consumers, e.g., one that helps users discover novel things, might not be optimal for the provider, for instance, in terms of generated short-term revenues.

In Table 1, we sketch four types of possible stakeholders, see also [39]. Besides *consumers* and *recommendation service providers*, we also identify *suppliers* and *society*. Suppliers are the providers of item offerings that can be recommended to users. Such suppliers could be content creators, manufacturers, service providers, or

also retailers who use the platform of the recommendation service provider for their business activities.

Table 1 Stakeholders of Recommender Systems

Stakeholder	Description
Consumers	Consumers are the end users of recommendation systems, and their behavior is potentially influenced by the system's recommendations.
Recommendation Service Providers	These are organizations that provide the recommendations as part of their services. They invest in recommendation technology and are the ones in control of the system.
Suppliers	These are organizations that provide (some of) the products or services that are recommended to consumers. The recommendation of their offerings can, for instance, influence the overall demand on the market.
Society	Depending on the size of the population of end users, recommender systems like on news or social media platforms, can have effects on parts of society as a whole. Note that the impact on society may be more than the aggregate impact on individuals, e.g., when the directly affected consumers act as multipliers of opinions in a society.

For some applications, no external suppliers might be involved at all, or their interests might not be taken into account in the design of the implemented recommendation strategy. In other scenarios, however, considering the supplier interests in an appropriate way may represent a central problem when designing the recommendation service. This is in particular the case when the recommendation service provider acts as a platform between different suppliers and consumers.

In the travel and tourism domain, for instance, websites of hotel booking platforms often have a recommendation component integrated. The interest of the suppliers—e.g., hotel chains or individual property owners—is to be frequently included in these recommendations, assuming that being listed leads to more sales. In case of suppliers being hotel chains, they might even have specific preferences which hotels are recommended, such as one with an overcapacity during a particular period of time. The recommendation service provider's interest might, in contrast, be to push those hotels where the expected profit margin or commission is highest. The consumers' interest, finally, is to find a hotel that best matches their preferences and interests. The recommendation service provider therefore may have to consider all these interests in parallel. Not considering the consumers' interests to a large enough extent may lead to a limited acceptance of the recommendations and to low conversion rates; resulting in reduced revenues/profit in the short-term and a loss in credibility and impact in the medium and long-term. Not considering the supplier's interest might, on the other hand, lead to a dissatisfaction on the supplier's side in the long run.

Finally, *society* as a whole can be another stakeholder that may at least indirectly be affected by recommendations. Today, there are various online services with an

enormous reach, including social media sites such as Facebook or Twitter, news aggregation sites like Google News, or media platforms like YouTube. The selection of the content presented on such sites may have significant effects on the users' view of the world, e.g., in terms of political questions, potentially leading to phenomena of filter bubbles or the broad dissemination of fake news.

2.2 Value Dimensions of Recommender Systems

The value driver model [6] of e-commerce business models was developed to depict the value-creating transactions by networks of business actors on virtual markets. It basically clusters the value driving aspects of e-commerce business models into four groups: efficiency, complementarities, lock-in and novelty. *Efficiency* generally refers to savings in transaction costs or time due to speed gains, scaling effects and the reduction of information asymmetries on virtual markets. *Complementarities* represent synergy effects due to the combination of different sales channels or product catalogs which creates additional opportunities such as cross-selling of items or follow-up sales offers to customers. *Lock-In* effects are creating value due to customer retention deriving from the avoidance of switching costs or positive network externalities. The latter derives from network connections and the joint use of services. For instance, the fact that collaborative filtering systems become more accurate when exploiting more data and from larger groups of users is an example for such a positive network externality. Finally, *novelty* stands for new architectural configurations and opportunities that would have been unfeasible on non-virtual markets. In the following, we structure the reported value contributions of recommender systems according to these aforementioned categories of the value driver model. As can be seen, recommender systems are making contributions with respect to all value creating aspects of e-commerce business models.

2.2.1 Efficiency

One pillar for e-commerce success is the reduction in information asymmetries between buyers and sellers due to the up-to-date and abundant availability of information with an enormous reach and at nearly zero cost. Recommender systems may help to counterbalance the resulting potential overload of information by primarily creating transparency by (relatively) unbiased information offerings, lowering consumers' information search costs and facilitating their decision processes as enlisted in Table 2. However, the impact on users' decision processes is always moderated by the concrete domain and the properties of products as Lee and Hosanagar [53] demonstrated in their large-scale e-commerce study. According to their results utilitarian and experience products, for instance, enjoy a higher lift in awareness due to recommendations as compared to hedonic and search product categories.

Table 2 Efficiency and the Value of Recommender Systems

Inform in a balanced way: A recommender system can be tuned to ensure that the recommendations are not biased, e.g., towards certain items or content supporting only one particular opinion.
Help users find objects that match their assumed short-term intent and context: Recommend items that are relevant in the ongoing user session, sometimes even without long-term preference information.
Help users find objects that match their long-term preferences: The most explored problem in the literature. Assumes the existence of long-term user profiles, e.g., in the form of a user-item interaction matrix.
Improve decision making: Decision making can be improved, e.g., by reducing choice difficulty and pre-selecting a small set of options; better choices and choice satisfaction can also be supported by explanations.
Establish group consensus: In a group recommendation scenario, the purpose of the system can be to make suggestions that are agreeable for all group members.
Cost savings: Recommender systems allow providers to automate the sales-advisory function at large scale and with zero marginal costs.
Real-time transparency: Recommender systems allow providers to observe in real-time the immediate effects of their automated sales advisory.

Due to the networking of consumers the support of decision making processes may not only target single consumers, but can even address the decision making of groups of users such as in [56]. From the provider perspective the efficiency gains of recommendation technology are primarily in cost savings by automating the traditional sales advisory function being naturally inherent to offline distribution channels like retail or personal selling. In this respect, the recent advances in natural language processing also initiated current research efforts towards chat-based recommendation approaches [60, 36, 56]. For reasons of completeness, however, it needs to be mentioned that the provisioning of recommendation services also incurs some operating costs. Therefore, Goshal et al. [29] made a theoretical analysis on the optimal strategies of consumers that would need to make their choice between personalizing and non-personalizing firms under competition. One of their outcomes is that consumers would weight in lower prices of non-personalizing firms against higher fit costs. Therefore one of their main results is that also the prices and profits of a non-personalizing firm are impacted by the competitor's recommendation system.

Furthermore, the real-time transparency of consumer search also—at least hypothetically—enables providers to immediately react to recent trends and eventually influence them by adapting the recommendation strategies. Since assumptions about purposeful biasing of recommendation agents would undermine consumers' trust in these systems, no research in this respect has been performed or at least reported.

2.2.2 Complementarities

The ease of integrating diverse and enormously large product catalogs and service offerings in online marketplaces enables the exploitation of synergies in sales. Again,

recommender systems are therefore a central cornerstone in many e-commerce business models to exploit these complementarity effects as summarized in Table 3. Like in the efficiency dimension, a large number of consumer-centric value aspects have been described. Again, however, only anecdotal evidences on generated revenue gains of providers have been made public or scrutinized further in recommender systems research.

Table 3 Complementarities and the Value of Recommender Systems

Enable item "discoverability": A recommender system can help users discover items in the catalog they were not aware of. This can lead to increased demand over time, but also to more engagement.
Help consumers explore or understand the item space: Sometimes users are not aware of the available options. A recommender system can be tuned to show a diverse set of options.
Create additional demand: One typical goal of recommendations is to point users to additional items in the catalog, e.g., to stimulate cross-sales. This can lead to increased demand both in the short and in the longer term.
Show alternatives: Relates to the previous aspect. A recommender system can present alternatives in the context of a reference (currently viewed) item.
Show accessories: Instead of alternatives, a recommender can point users to accessories of given reference items.
Recommend in sequence: Create a logical continuation of previously observed user interactions, e.g., recommend next place to visit or next music track to listen to.
Actively notify users of new content: A recommender might proactively notify users, e.g., through push notifications, with a focus on novel content to stimulate consumption and interaction.
Revenue gain: based on generated additional business.

2.2.3 Lock-In

The *stickiness* or ability to lock-in participants is another core characteristic of online business models. Recommender systems again primarily focus on consumers or users and create value by personalizing their interaction experiences. Table 4 selectively lists reported value aspects of recommender systems. Technically this is facilitated by exploiting users' observed behaviors and data points and making the system an indispensable virtual alter ego that would be lost when switching platform or provider. This is analogous to the stickiness of traditional sales agents in retail knowing their clients for years. Clients over time therefore recognize them as a form of *old friends* whose viewpoints and recommendations they trust. A hidden value from the provider perspective is therefore in particular the corporate knowledge management aspect that becomes evident in the popularity of recommender systems as an application domain for data science efforts.

Table 4 Lock-In and the Value of Recommender Systems

Increase user engagement and activity on the site: User engagement is often used as a proxy to gauge the effectiveness of a recommender system. When consumers interact with a service, e.g., for music streaming, more frequently customer churn is expected to be lower.
Entertainment: A recommender system can be entertaining or emotionally satisfying, e.g., by supporting discovery of new content in a convenient way.
Remind consumers of already known items: In some domains, it can be helpful to remind the user of things they already know or have. The recommender might remind the user of the purchase of consumables or show items the user liked in the past.
Provide a valuable add-on service: A recommender system can serve as a tool to differentiate the provided service from competitors in the market. High-quality personalized recommendations may lead to increased customer retention and increase the switching costs for customers.
Learn more about the customers: Personalized recommendations require the collection of customer preference profiles and often a thorough understanding of the specific demands of certain consumer groups. Providing a recommendation service might therefore contribute to a better understanding of customers in general.

2.2.4 Novelty

Novelty and the innovation potential of e-commerce business models generally refers to the introduction of new products or services or the establishment of novel process models transforming and creating businesses.

Note, however, that with the *novelty* dimension we do not refer to recommender systems being novel themselves, but that they potentially create novel opportunities and business scenarios.

While recommending novel or previously unknown items to users has already been mentioned as an example of value creation due to efficient information processing on virtual markets, the purposeful influencing and biasing of consumers' will by persuasive recommendation and nudging strategies [74] can fall into this novelty category as summarized in Table 5. Provisioning of automated recommendations is nowadays commonplace, particularly in e-commerce and on media streaming or (social) media sites. However, there still exists a number of potential application domains where recommendation techniques are not yet widespread. For instance, providing decision support in the domain of health and well-being, the potential for innovating by inducing behavior change through recommendations is only starting to be developed [65].

2.3 Risks of Recommender Systems

While recommender systems are designed to create value and have a positive impact on consumers and organizations, there exist also certain risks that come with the use of recommendation technology. In Table 6, we discuss examples for such risks.

Table 5 Novelty and the Value of Recommender Systems

Nudge toward desired behavior: A recommender might stimulate certain desirable behavior at the user's side, e.g., with respect to healthy behavior, through nudging techniques.
Increase (short-term) business success and promotion of content: A recommender system can be tuned to steer customer demand and to promote items that are favorable in terms of business-related figures such as revenue or profit.
Change user behavior in desired directions: By pre-selecting the items in a recommendation list, the choice set for the consumers can be reduced and certain items can be promoted. This may effectively change the consumers' behavior in desired directions, e.g., towards more profitable items.

Compared to the analysis of the benefits of recommender systems, potential negative effects have so far not been explored to a large extent in the literature. In [13], for instance, the authors investigated the effects of a malfunctioning music recommender system on consumer trust and behavior. Researchers from the field of Marketing [10] looked at the potential negative monetary effects of recommending the wrong items, such as items that consumers were already likely to purchase anyway. Recommending such items limits the opportunities to promote other items under the assumption that only a limited set of items can be recommended. The aspects of privacy and fairness, finally, have obtained increased interest in recent years in the recommender systems research community, see, for instance, [27] or [12].

3 Measuring the Impact of Recommender Systems

In the previous section, value contribution and impact of recommender systems in terms of the four categories of the value driver model have been presented. This section will now elaborate on approaches to assess and quantify the impact within these categories as well as their corresponding measures and Key Performance Indicators (KPIs). When seeing a recommender system purely as a machine learning task its evaluation is typically focusing on accuracy related measures. However, nowadays recommender systems research is coming to the conclusion that the impact of personalized content and item recommendations need to be assessed from a multi-disciplinary perspective with a plurality of methods and approaches [76]. Next, we will therefore shortly discuss methodological aspects before moving on to the most widespread measures in practice. They obviously focus on quantifying the value created for providers given that the value needs to surpass or at least match the efforts and investments. However, without creating value for consumers, no sustainable value can be created for providers. Thus, when users are consistently satisfied by, for instance, finding items of interest with ease, also the provider's KPIs will indicate a positive impact. Consequently, a longitudinal approach when quantifying the impact will lead to more reliable results.

Table 6 Potential Risks of Recommender Systems

Consumer Risks	
Poor Decisions / Choice Dissatisfaction	Ultimately, the pre-selection of items or the decision bias introduced by a recommender system may lead to bad decisions or to choice dissatisfaction.
Bad User Experience / Decision Difficulty	If the set of recommendations is chosen in an unfortunate way and, e.g., only contains very similar items, this might in parallel lead to a poor user experience and an increased decision difficulty for the consumer.
Biased Information State	In case the presented options emphasizes one particular range of the available options, the consumer might be left with an incomplete and biased information state.
Privacy	Recommendation providers may collect all sorts of user interactions in a comprehensive way and try to connect the information about users across services and sites. This may endanger the privacy of consumers.
Organizational Risks	
Loss of Consumer Trust	When recommendations are—for an extended period of time—not helpful for consumers or appear biased, they might lose the trust not only in the recommendations but in the organization as a whole.
Loss of Societal Trust	In particular recommendations that appear biased or unfair (e.g., appear to be only advantageous for the provider organization) may lead to a bad reputation of an organization and a loss of trust by society as a whole.
Missed Opportunities / Financial Loss	Recommender systems can have significant positive impacts on organizations in terms of business-related figures. A poorly designed recommender system might in contrast lead to missed opportunities, due to a low value of recommendations for users.
Societal Risks	
Filter Bubbles & Echo Chambers	Biased recommendations in particular on news sites or social media platforms may lead to filter bubbles and echo chambers, where the presented information mainly reflects pre-existing interests and viewpoints of individuals or user groups.
Algorithmic Bias and Discrimination	In some application scenarios of recommender systems, the underlying algorithms may reflect or even reinforce uneven distributions in the data, potentially leading to discrimination of certain user groups.

The academic literature identifies three main methodologies for evaluating recommender systems: *(i)* field studies (A/B tests), *(ii)* user studies, *(iii)* offline experiments. Field studies are run by recommendation service providers to test the effects of different recommendation strategies on their respective KPIs. We will discuss which KPIs are widespread in industry and which insights were obtained from field studies later in this section. User studies are typically executed in the form of controlled

experiments, either in the lab or online. Here, the study participants are randomly assigned to interact with different versions of a recommender system or they interact with the system under different conditions. Typical goals of such user studies are to assess the subjective quality perception of a system or of some of its components and in order to understand how these perceptions might influence the future behavioral intentions of users, see [49, 58]. Such studies often focus on user experience aspects of recommender systems. We will discuss the need for more user-centric studies, which also consider the provider perspective later in Section 4.

Offline experiments, finally, are purely data-based and do not require the active involvement of users for their execution. Such experiments are by far the most common ones in literature and we will focus on potential limitations of today's offline experimentation practices when it comes to the assessment of the impact and value of recommender systems.

Most often, offline experiments are used to compare different algorithms in terms of their capability of predicting held-out interaction or preference data. They build on the underlying assumption that when an algorithm is able to rank the presumably more relevant items higher in recommendation lists, the generated recommendations will more likely better match users' interests. Thus, algorithms with higher offline prediction accuracy are supposed to also outperform their weaker offline comparison partners in real-world settings.

However, assessing the impact of a recommender system based on offline experimentation is sometimes seen to be too simplistic and building on problematic assumptions. It is, for instance, relatively straightforward to predict based on past data that a lover of Star Wars movies will watch any new sequel being released. Therefore, recommending a new sequel to this user may—even though the prediction is perfectly accurate—not create much value, neither for the consumer already knowing about the sequel and watching it anyway, nor for the provider who could have promoted another content.

In terms of the evaluation measures, algorithmic research is strongly concentrated on prediction accuracy, i.e., the ability of an algorithm to predict held-out data. Evaluating recommendation algorithms in terms of accuracy is generally meaningful, as discussed above. However, it stands to question *(i)* if the often small increases in accuracy on selected datasets reported in research reports would truly make a difference if these new algorithms would be deployed in practice[3], and *(ii)* if the results from offline tests are generally predictive for outcomes of deployed systems [30, 41]. Furthermore, while some algorithms might yield good offline accuracy results, the resulting recommendations may have other, undesired properties such as a bias towards popular items generating only limited additional value for providers [42]. Finally, an orthogonal problem in the context of offline studies seems to lie in methodological problems in applied machine learning research, where we often only observe an "illusion of progress" [32]. While the proposed models become more

[3] One historical fallacy is that while in other application areas of machine learning, e.g., image classification, every small improvement in accuracy may lead to an improved system, it is less than clear that small improvements in prediction accuracy on past data have a positive impact at all with respect to the effectiveness of a recommender system.

and more complex, they sometimes actually do not outperform existing methods if evaluated independently [7, 61, 24, 23], which potentially leads to a certain stagnation in algorithm research.

In general, assessing the impact of recommender systems with offline experiments alone has its limits. There are, however, a number of reports on successful real-world deployments of recommender systems. Such reports typically summarize the outcomes of field studies on newly introduced recommender systems or on two system versions A/B tested in parallel. In the following, we provide an overview of selected findings from real-world deployments based on the survey presented in [41]. This overview helps us to understand both *(i)* which measures and business-oriented KPIs are used in practical environments, and *(ii)* which effect sizes are observed.

3.1 Value Dimensions and Measurements in Practical Applications

In field studies, relating the single value drivers to separately measurable effects of the impact of recommendation applications is not always possible, since they cannot be isolated and individually considered. For instance, streaming media providers like Netflix or Spotify have business models based on flat-rate subscription fees. In their case, an individual successful recommendation is not only a sign of an efficient presentation of the enormous catalog space and effective decision making leading to an immediate consumption, but also contributes to the lock-in of the particular subscriber. Taken together this will also effect customer attrition and finally revenues, even if it is not directly measurable. In other cases, e.g., in news recommendation, it can even be difficult to assess in the short term if a recommendation was actually a success. While clicks on recommended articles can be easily measured, we often cannot be entirely sure if the consumer actually liked the content or not or if the consumer would have actually read the article even when it was not recommended.

In Table 7 we relate the measures and KPIs to the value dimensions introduced in Section 2. However, note that neither of the mentioned measures does exactly capture and quantify the identified value contributions discussed in the previous section. Let us consider, for instance, the Click-Through-Rate (CTR): it measures that an item presumably caught the user's attention and that it therefore serves as a first and preliminary indicator for a successful match of either user's short-term or long-term interests. Such preference matching therefore relates to the efficiency value dimension that facilitates to cope with information overloading by lowering users' search costs. However, while a click-through is a necessary precondition for indicating a successful match of preferences it is not sufficient. For instance, also some follow-up purchase or consumption (such as reading) would need to be observed additionally in order to have a stronger indication of a successful preference match. Furthermore, if we observe changes in CTR over time we could interpret them also as a proxy for an increase/decrease in user engagement that is another value aspect associated to the lock-in dimension. Thus, Table 7 must not be misinterpreted by assuming that the enlisted measures holistically capture already all aspects of the

discussed value dimensions. However, the dimensions have to be seen as a structuring mechanism for the main types of measures as they are found in the literature [41]. This emphasizes the need for further research to more systematically develop and define measures that would also be sufficient indicators for the specific value contributions of recommendations.

Table 7 Value Dimensions and Measurements

Value Dimensions	Measurement
Efficiency	*Click-Through-Rate (CTR)*: measures how many clicks a recommendation has garnered. Thus, the CTR indicates that recommendations have been noticed and presumably influenced users' choice due to their timeliness and relevancy.
Efficiency	*Adoption and Conversion* measures and correlates observed user behavior other than clicks with recommendation success, e.g., when users watch a certain fraction of a movie. Conversion rates measure the fraction of users who take a desired action, e.g., submit a resume after receiving a job recommendation.
Efficiency, Complementarities	*Sales and Revenue*: when recommendations lead to purchases, one can measure corresponding business-related figures such as revenue, sales, or profit.
Lock-In	*User Engagement*: Increased user engagement is often considered as an indicator of customer retention. Engagement can, for instance, be measured in terms of the number or length of usage sessions.
Novelty	*Effects on Consumption Distributions*: Recommender systems can lead to desired or undesired changes regarding sales distributions, such as, increased sales of long-tail or already popular items. They can also inspire users to consume different items than they would without a recommender.

Our discussion will show that recommender systems can have substantial impact on all these measures, but also that all these measurements have their limitations. Relating measures to the value driving dimensions as done in Table 7 above, can also help to recognize completely uncharted aspects, such as comparing, for instance, the cost aspect of virtual sales advisory and automated recommendations with physical in-shop encounters.

3.2 Click-Through-Rate

The CTR is a wide-spread measure used, e.g., in the domain of news recommendation. Results from real-world deployments are reported both for larger news aggregation sites like Google News [18], for business-oriented sites such as Forbes.com [48] and for regional ones like swissinfo.ch [28]. Other domains where the CTR were reported, sometimes in combination with other measures, including recommendations of videos on YouTube [19] or the recommendation of similar offers at eBay [47].

These studies typically report the effects of introducing a new recommender system when compared to the existing one. In the news domain, the reported increase of the CTR is often around 35%. In the study at YouTube [19], however, the difference in CTR between two systems is as high as 200%. In another study at eBay, in contrast, only a 3% improvement was observed [11]. The increase in revenue was however higher (6%).

The huge differences in terms of the reported effects may be explained both by the fact that the systems were deployed in different domains and, more importantly, that different baseline algorithms were used for comparison. In [19], a personalized algorithm for YouTube recommendations was compared to a non-personalized one that recommends the most popular videos on the site to everyone. In [11], in contrast, two more sophisticated techniques were compared, leading to smaller effects.

Independently from this problem of interpreting absolute numbers, using the Click-Through-Rate as the only or main instrument for assessing the value of recommender systems can be misleading. Often, the CTR mainly measures if individual recommendations raised attention or interest in an item. The click counts however cannot tell us if the consumers actually liked the item or if the recommendations will increase the probability that the user will return to the site next time. Increases in the CTR can sometimes also be achieved quite easily, e.g., through clickbait headlines on news portals or by recommending items that are generally popular. Moreover, the visual positioning of the recommendations can have a substantial impact on the CTR as well, as reported in [28]. Here, a more prominent positioning of the recommendation widget doubled the CTR in the short term.

Overall, optimizing for the CTR in the short term might in fact have negative business impacts in the long run, when consumers repeatedly feel misled by the recommendations or when the recommender system fails to draw the consumers' attention to less popular items. Also, as reported in [78], there can be a trade-off between optimizing item rankings for clicks and optimizing according to relevance for the consumer.

3.3 Adoption and Conversion Measures

Adoption measures assess the commercial success of a recommendation in a way that is more precise than CTR. In most cases reported in the literature, these measures are specific to the application domain. In the video recommendation domain, providers such as YouTube [19] or Netflix [30] measure the number of recommended videos of which consumers watched at least a certain fraction. In other domains, various types of user actions are interpreted as success indicators for a recommendation, e.g., "add-to-wishlist" events, "bid-through" and "purchase-through" rates on eBay, "cite-through" rates for research paper recommendations [8], booking requests on tourism sites [75], or opened communications on a dating platform [71].

The improvements in terms of such measures are again difficult to interpret on an absolute scale due to different measurement methods, baselines, and application

domains. A/B tests sometimes indicate increases of a few percent in terms of the "purchase-through" rate. In other cases, where the user action is not directly leading to increased business value (e.g., "add-to-wishlist" events or "click-out" actions on a marketplace [43]), the differences between two algorithms can be substantial, e.g., 89% in the case of "add-to-wishlist" actions on eBay [47] or over 90% higher click-through rate on tv programs [68].

Like for the CTR, the described adoption measures are often only proxies for the business value. An observed "add-to-wishlist" event for a recommendation is not yet a transaction, and one has to be careful not to overestimate the business value of such events. In [40], the results of A/B testing various algorithms for the recommendation of games for mobile phones were presented. The test included a variety of measures including different conversion rates (e.g., recommend-to-purchase), click behavior, game downloads, as well as actual purchases. Among other aspects, it turned out that the number of item view events (i.e., CTR of a game's detail page) incited by a recommendation was *not* indicative for the ultimate business value, which was measured in actually purchased games. In fact, even the stronger signal of a download of a demo version was not a strong predictor, which means that some algorithms created increased interest in certain items, but did not lead to a purchase at the end.

3.4 Sales and Revenue

Measuring changes in sales, revenue or similar business-related figures is the most direct way of determining the business value of a recommender system. Such measurements are common when the revenue of a provider is not based on a flat-rate subscription model but on individual item sales or fee-based transactions. In such situations, one cannot only measure the overall effects of different recommendation strategies in an A/B test, but also which recommendations actually led to successful transactions.

A few works in the literature exist that report the outcomes of A/B tests in terms of such business-related measures. In [52], for instance, the authors compared two recommendation algorithms—"view-based" and "purchase based" collaborative filtering—with a baseline condition where no recommendations were provided to customers of an online DVD retailer. The study revealed that the "purchase-based" best strategy led to an increase in sales by 35% ("*for those who purchase*" [52]). Interestingly, when the recommendations were solely based on item view events, no increase in sales was observed compared to the baseline condition. These observations emphasize the importance of algorithm choice and at the same time give an indication of the extent of missed sales opportunities when no recommender is used.

A much more modest, but still relevant increase in terms of revenue was reported for the mobile game recommendation field study mentioned above [40]. In this A/B test, the strongest increase was observed when a very simple content-based algorithm was used, leading to an increase of 3.6% in sales. The study however also revealed that different algorithms should be used in different navigational situations in order

to maximize the revenue. Thus, even slightly stronger increases can be achieved, e.g., when a switching hybrid strategy is applied.

A study that examined how context-awareness of content-based recommendations impacted business performance measures was conducted by Panniello et al. [57]. They identified that the increased accuracy and diversity of the context-aware recommendations for comic books positively affected trust in the system which in turn affected purchases.

A number of additional studies in the literature focus on other e-commerce domains like e-grocery or online book stores [66]. The overall direct effects of adding a recommender system are often small, e.g., 0.3% in [20] or 1.8% in [51]. However, there can be major indirect and inspirational effects, e.g., up to 26% in some categories [20]. An interesting observation is furthermore reported in [66], where sales decreased significantly after the recommender was removed from the site.

Overall, we see a strong spread in the reported effects, from almost no effect up to 500% increase in Gross Merchandise Value for a study done at eBay [16]. Comparing the absolute values reported in these studies is challenging because the individual observed changes depend on many factors that are specific to the application domain, business model, market situation etc. A typical limitation of reported studies is that it is not always entirely clear if the observed effects would last beyond the A/B testing period, which is often limited to a few weeks. Another aspect related to the revenue impact of recommender systems is price sensitivity and users' willingness-to-pay (WTP). In this context Adomavicius et al. [4] determined that online recommendations have a positive effect on the recipients' WTP that correlates with the predicted ratings. This effect was observed even in the case of perturbed or manipulated rating predictions.

3.5 User Engagement

Discovery support, as mentioned above, can be a main functionality of a recommender system, which in turn leads to increased user engagement and customer retention. User engagement can be assessed in a number of ways, e.g., by the number of visits to the site per month, by the number of consumed content on a media site, or by the length of the individual consumption sessions and the number of particular user actions in such sessions.

Various works in the literature report increased user activity on the site, e.g., in the news domain [28, 48], where the visit lengths sometimes more than double when a recommender system is in place. Substantially increased user activity levels were furthermore observed in [21] and [46] for the music domain, where the number of playlist additions was used as an indicator for user engagement. Additional examples with increased user activity levels due to the presence of recommendation systems can be finally found for social networks and platforms [67, 72].

In many applications, it is a reasonable assumption that higher engagement with a service leads to repeated use in the future. However, the choice of the measurement

that serves as a proxy for engagement can be crucial. More clicks (interactions) by a user may, for example, not always be a good sign. Like in the evaluation of a search engine, more interactions might indicate that the system was actually not able to present something relevant to the user. A typical example in the recommendation context would be recommendations that are explicitly designed for discovery, as can be found on music streaming sites. Here, the best case is that the user who is presented with the recommendation list, discovers a new artist and then leaves the discovery module to explore this new artist. As a result, fewer interactions with the recommendation list might indicate that the recommendations were actually good.

Given this need for interpreting the observations of user behavior, research needs to more frequently draw a bow from the underpinnings of cognitive and psychological science to the principled observations and experimentation in order to draw reliable conclusions for algorithm and system adjustments as has been postulated in [76].

3.6 Effects on Consumption Distributions

Increases in sales and revenue, as previously reported, are often simply the result of consumers buying *more*, e.g., due to cross-selling links provided by a recommender system. Recommender systems may, however, also inspire users to consume *different* things. Such changes in the distribution of consumption or sales of items can be desirable for different reasons. On the one hand, there might be direct business-related effects when using a price and profit aware recommender system. Such a system might, for example, aim to recommend items that are both a good match for the consumer's preferences and at the same time are more profitable than other items that could be similarly good matches [38].

On the other hand, there might be more indirect effects that come with changes in the consumption distribution due to either enabling customers making new discoveries or re-enforcing rich-get-richer effects [26].

Already the seminal paper of Resnick et al. [62] questioned if the peer groups created by collaborative recommender systems will be permeable or fracture the global village into tribes. 20 years later Hosanagar et al. [35] identified, at least for the domain of music recommendations, that users consumed in general more and thus also more commonalities among users arose due to the impact of recommendations.

In some application domains like e-commerce, a large fraction of the sales comes from a relatively small number of "blockbuster" items, leading to a long tail of less frequently sold items. Recommender systems have the potential to point consumers to such long tail items and to help them discover parts of the catalog that were previously unknown to them, such as observed in [75]. In transaction-based applications, e.g., on e-commerce sites, this can lead to the indirect effect of more sales in specific categories as discussed above [20]. Such indirect and inspirational effects of recommender systems were also found in the music domain [46]. Generally, supporting discovery and driving customers away from blockbusters is often considered a main

purpose of providing recommendations on media streaming services, assuming that discovery leads to more engagement and, ultimately, continued subscriptions [30].

Finally, changes in the consumption behavior of users might also be desirable from a societal perspective. The purpose of a recommender system might, for example, be to entice healthier consumer behavior. In this context, the authors of [22] explored the potential of nudging users towards more healthy dishes through recommendations. Another goal may be to stimulate users to enrich their information consumption behavior, e.g., by presenting news content that covers more than one opinion on a controversial topic to avoid filter bubbles and echo chambers. The study of [25], for instance, identified an increase of individuals' exposure to news from their less preferred side as well as a (modest) increase of ideological distance between individuals when analyzing the news consumption behavior of 50K users based on their browsing histories.

Generally, influencing *short-term* consumer behavior through recommendations to some extent might not be too difficult in a number of application domains. This is in particular the case when the system actually filters out certain options, which cannot be chosen by the consumers anymore. Often, however, a main challenge is to achieve positive long-term effects. Only recommending items that are highly profitable but not too relevant for consumers might, for instance, result in increased short-term profits, but lead to a loss of consumer trust in the long run.

4 Towards More Value-Oriented and Impactful Research

Our discussion has shown that there are different dimensions that drive the value creation of recommender systems for their stakeholders. Given this richness, the focus in academic research appears to be narrow. First, there is a strong focus on algorithmic proposals mostly evaluated on historical data. This focus on accuracy—although important—touches however only few value creating dimensions, while, in contrast, questions related to the users' experience or long-term effects of sales distributions are underexplored [45, 39].

Generally, there is no doubt that industry has always picked up proposals from the academic world and that they have successfully implemented and further improved novel machine learning models to serve given organizational goals. A prominent example is the use of matrix factorization techniques, which were explored already in the late 1990s, became popular in the context of the Netflix Prize, and were later broadly adopted in industry. However, as a result of the above-described phenomena, the question arises if—despite the many papers that are published every year—major parts of today's academic research on recommender systems actually have a strong impact in practice.

To be more impactful and relevant in practice, our research approach should therefore be refocused on two dimensions.

1. In terms of the *research scope*, it is important to put more emphasis on the user experience of recommender systems than on algorithms [50], acknowledging that

most aspects of recommender systems cannot be evaluated without involving users and without considering its context of use [44].
2. Regarding the *research methodology*, the focus on algorithms led to a certain overreliance on offline experimental designs that are common in applied machine learning research, but which are insufficient to assess the impact of a complex information system like a recommender.

Clearly, it would generally be desirable that more academic proposals were evaluated in real environments. There are, however, a number of other research instruments that are available, and there are several ways in which the research efforts of the community could be refocused in order to be more impactful.

4.1 Recommender Systems Research with a Purpose

Looking at the research scope, many of the works published today—in particular in the predominant area of algorithms—is not based on theory or explicitly stated research hypotheses. Higher prediction accuracy of a machine learning model is implicitly equated with progress towards better systems, even in cases where better accuracy is only demonstrated by a very specific experimental configuration of datasets, baseline algorithms, and evaluation procedures. In the midst of such a "leaderboard chasing" culture [54], the fundamental and underlying question "*What is a good recommendation (in a given context)?*" is too rarely asked [37].

As the discussions in Section 2 indicate, the same set of recommendations can be useful or not, depending on the goals that one tries to achieve, the perspective that is taken (e.g., consumer vs. provider), and the individual user's preferences and context. Therefore, when evaluating a recommender system, these surrounding factors and the goal that one tries to achieve should be made explicit in order to ensure that the chosen evaluation approach is appropriate [34]. In [37], a four-layer conceptual framework is proposed to guide researchers, both in academia and industry, towards a more goal and value oriented approach when designing and evaluating a recommender system. We summarize the main layers of this framework in Table 8, where we also give examples both from the consumer and provider perspective at each level.

Note the importance of making suitable choices at the lowest layer of this framework, i.e., how the system is evaluated. The framework should help to ensure that there is a "metric-task-purpose" fit, i.e., measuring if the system is actually able to fulfill the goals it intended to serve. Today's research far too often seems to focus solely on the two layers at the bottom of the framework. The system's task usually is considered to help the user to "find good items" [34]. Moreover, how the evaluation of algorithms is done is largely standardized, based on offline protocols and metrics such as Precision, Recall, and RMSE. However, such metrics, while generally useful, can only inform us about differences between algorithms in terms of item retrieval and ranking performance, but not about the created value for individual stakeholders. The proposed framework should therefore help researchers to look beyond the common "find good items" task and extend their research scope to the many other and

Table 8 A Conceptual Framework for Goal and Value Oriented Research

Goals and Values	At the top-most strategic level, it is important to understand or define the goals of the system and which value it drives for which group of stakeholders. From an organizational (provider) perspective, this goal could for instance focus on recommending complementarities and increasing revenues or on the lock-in value driver dimension and customer retention. For a consumer, e.g., of a media service, the value could for example simply be entertainment.
Recommendation Purpose	Depending on these strategic considerations, the specific purpose of the recommender system in this context has to be clarified, i.e., how the recommender system can help to achieve the described goals. If, for instance, one goal is customer retention, one purpose of the system may be to increase the "discoverability" of specific items for consumers and to thereby increase their engagement.
System Task	At the next, more operational level, the question has to be answered how the system or its components, e.g., an algorithm, has to be designed to support the intended purpose. In case the goal is to support discovery, an algorithm may try to intentionally diversify the recommendations. This might involve prioritizing items that do not have the highest consumption probabilities according to past observations.
Evaluation Method	The three upper layers finally determine how the system should be evaluated. In the given example of discovery support, one could use a combination of *(i)* computational metrics to objectively assess the level of novelty of the recommendations, and *(ii)* subjective measures collected through a user study regarding, e.g., the subjective perception of the recommendations and the participant's intention to use the system in the future.

more specific purposes recommender systems can serve, both from the perspective of the consumer and the provider.

4.2 Utilizing a Richer Methodological Repertoire

Extending the research scope in the described ways, e.g., to consider the impact of recommender systems on different stakeholders, requires a more comprehensive methodological approach than we often observe today. First and foremost, recommender systems are much more than retrieval systems. They exert influence on the choices of users, and whether a recommendation leads to follow-up actions of users or not depends on a variety of factors including individual decision heuristics and biases. Therefore, various aspects of a recommender system can only be addressed with research designs with humans in the loop, such as controlled experiments with users or qualitative and observational studies. On the other hand, recommenders are e-commerce information systems with a clear business impact and even implications

for society, which is why we need appropriate evaluation methodologies to assess these effects as well.

In the following, we will outline potential ways forward in terms of how we can address and evaluate the sometimes complex interplay between consumers, organizations, and information systems that is embedded into a societal context.

4.2.1 Evaluating with Humans in the Loop

When providing consumers with a recommendation service, we might ask ourselves a number of questions regarding its value for the users, e.g.:

- *How are recommendations perceived by users—do users actually find them helpful?*
- *Do users find recommendations diverse and surprising enough—do they help them discover new things?*
- *Would they appreciate more information justifying why certain items are shown and others not—do users trust recommendations to be fair?*
- *Would they be interested in receiving more recommendations—would they use the service in the future?*

None of these questions, can be confidently answered without involving humans when evaluating the system. Remember here that such user-centric questions are also very important from a provider's perspective, assuming in general that added consumer value directly or indirectly leads to increased value for the operator.

Evaluating from a user-centric or human-computer interaction (HCI) perspective has a long tradition in recommender systems research [50]. User centric research is typically guided by explicitly stated research questions, e.g., to what extent users would value a certain type of explanation. One main first challenge in such research efforts usually is to develop an appropriate experimental design, e.g., a randomized controlled experiment, that helps to reliably answer the stated questions. As a result, user-centric research is much less standardized than typical offline experiments that are conducted to compare the prediction accuracy of algorithms.

Nowadays, at least two problem-independent frameworks for user-centric recommender systems evaluation exist [59, 49]. Researchers benefit from such frameworks for their own works in different ways. These frameworks, for example, provide validated questionnaires for various general quality dimensions of a recommender system, and they also demonstrate how to use broadly-used statistical analysis techniques like Structural Equation Modeling. Nonetheless, user-centric research remains difficult and requires often more effort than offline experimentation in recommender systems research in many ways. Research questions must be stated on theoretical considerations, specific experimental designs have to be designed and defended against reviewers, participants must be recruited that are at least representative for some group of users, and finally there is not always consensus in the community regarding which statistical methods are appropriate for the subsequent analysis.

These difficulties may have led to a certain over-reliance on offline experimentation in our field. As a result, we observe that even though the number of research questions that can be reliably answered with offline experiments alone is actually small, such designs dominate the research landscape. While there are efforts to use computational metrics to capture some of the aspects mentioned above, e.g., regarding the diversity of a set of recommendations, many of the proposed metrics were not validated with humans in the loop. It is, for example, less than clear if the many metrics proposed for novelty, diversity, or serendipity actually correspond to the human perception.

As a result, it remains important that the community focuses more on user-oriented aspects of recommender systems. This is particularly important as there exists a number of research works that indicate that offline metrics like Precision, Recall, or RMSE do *not* necessarily correlate with the quality perception of users [64, 17, 28, 55, 8]. The work of Adomovicius et al. [4] goes even beyond these works by demonstrating in controlled behavioral studies that random recommendations or recommendations that artificially increase the predicted rating lead to a significant increase in the participants' willingness-to-pay for the received song recommendations. Thus, even if in field research the ultimate goal of actual conversion or purchase is reached, we still cannot be entirely confident that the actual recommendation was accurately matching the users' tastes. Here, *multi-modal* evaluation approaches may be advisable that compare results from offline experiments with the user's perceptions from a user study. With respect to the type of user involvement in evaluations, note that researchers are not limited to randomized controlled experiments when it comes to involving humans in the loop. Interviews, surveys, focus groups and various other types of *qualitative* research methods are used in various other fields outside of computer science, and should be considered in recommender systems research more often as well.

4.2.2 Re-thinking Data-based Research

Despite their limitations, offline and data-based research will remain relevant in the future and will not be limited to the assessment of computational aspects such as scalability. Typical offline experiments that are done, e.g., regarding the prediction accuracy of algorithms on historical data, can still serve us to identify algorithms or algorithm variants that we may rule out from A/B testing. However, instead of asking if a new machine learning model can outperform another one by a few percent on given datasets in terms of accuracy, the focus should be shifted to different types of questions. For instance, can we analyze in which ways the recommendations generated by one algorithm are different from those of another one? Evidence exists that algorithms with the same performance in terms of accuracy often recommend very different things to users [42]. Given the characteristics of such recommendations, we may then ask questions if these differences are actually desirable from an organizational perspective. For instance, is recommending already popular items of interest from the business perspective or not? A number of research works were published

that go into that direction, including a variety of works that aim at understanding aspects like diversity, novelty, or serendipity. Assuming that the used computational metrics correspond to user perceptions, offline experiments can help us to compare algorithms in these respects and to design new approaches that are able to balance potential trade-offs.

A general limitation of existing offline evaluations is that often little is known about the provenance of the data. Researchers in the domain of music recommendations often base their works on listening logs collected from music services like last.fm. In the e-commerce domain, on the other hand, datasets are often used that contain user interaction logs like item views or purchase events. One major issue with such datasets can be, for instance, that a recommender system or a personalized filtering algorithm was already in place. Such additional attention biases exist both in log-based datasets [14] as well as in rating datasets [3], and researchers have begun in recent years to deal with such phenomena that ultimately lead to biased, and thus not informative evaluation results, see, e.g., [73, 15]. In that context, new forms of offline evaluation approaches have emerged in recent years, including "off-policy" evaluation and "counterfactual reasoning" approaches [15, 31] that promise to lead to results that are more reliable predictors of A/B test outcomes.

Another area where offline experiments could be further explored are simulations. Almost all research today is focusing on short-term effects of recommendations, i.e., if a recommender is able to provide helpful item suggestions in the given situation. In contrast simulation-based approaches, such as agent-based modeling, were already successfully applied in other disciplines like managerial science [70]. Using such simulation approaches allows researchers to investigate, for instance, *longitudinal effects* of recommender systems such as potentially unexpected emergent agent behavior [77] or the impact on choice diversity [33]. These simulation approaches may furthermore also be promising choices to analyze the effects of different recommendation strategies in multi-stakeholder environments and eventually even make predictions on societal implications.

Finally, we can often observe a tendency towards over-generalization from results of offline experiments, such as one method being able to improve the state of the art even though this is only shown for a very specific experimental setting. While researchers are generally interested in generalizable solutions, it is very often imperative to closely look at the data and consider the specifics of a domain or application to reach reliable insights. Should we, for instance, in an e-commerce setting remind customers of items they have already inspected in previous sessions? Should we recommend them items that are currently on sale or trending in the community? Answering such research questions often requires an analytical, data science approach to recommender systems research. In some cases, such analytical insights can guide the design of domain-specific algorithms considering the characteristics of successful past recommendations [43], or help to understand the business impact of recommender strategies based on econometric models [2].

5 Conclusions

This chapter introduced the different value driving dimensions of e-commerce business models and used them to categorize and structure the various types of value that recommender systems create for their stakeholders. In this chapter we reviewed the most important studies demonstrating the strong influence recommender systems exert on the decision-making behavior of users as well as the potential risks that come with them. While the business value of recommenders in general is undoubted, deeper analysis on the specific value aspects and how to make them transparent with adequate measurements is still in its infancy. The review on the most widely used metrics revealed that they are only loosely related to the various value types associated to recommender systems. Thus, this chapter might also stimulate research to identify the blind spots of value and impacts that are not yet quantifiable with metrics as well as incite efforts to develop more prescriptive advice for practical settings. Furthermore, we found that today's academic research on value and impact of recommender systems suffers from a somewhat limited methodological basis. An over-reliance on offline experimentation and accuracy measures raises questions on the impact of current research work in practice. We therefore argue that a paradigmatic shift is necessary in how we evaluate recommender systems and outline potential ways of re-focusing research in the field both in terms of research scope and applied research methodology.

References

1. H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modelling and User-Adapted Interaction*, 30:127–158, 2020.
2. P. Adamopoulos and A. Tuzhilin. The business value of recommendations: A privacy-preserving econometric analysis. In *Proceedings of the International Conference on Information Systems (ICIS '15)*, 2015.
3. G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. Reducing recommender systems biases: An investigation of rating display designs. *MIS Quarterly*, 43, 2019.
4. G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang. Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, 29(1):84–102, 2018.
5. G. Adomavicius and A. Tuzhilin. Personalization technologies: a process-oriented perspective. *Communications of the ACM*, 48(10):83–90, 2005.
6. R. Amit and C. Zott. Value drivers of e-commerce business models. In C. Lucier and R. D. Nixon, editors, *Creating value: Winners in the new business environment*, pages 15–47. Blackwell Publishers, Oxford, UK, 2002.
7. T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 601–610, 2009.
8. J. Beel and S. Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Proceedings of the 22nd International Conference on Theory and Practice of Digital Libraries (TPDL '15)*, pages 153–168, 2015.
9. I. Benbasat and W. Wang. Trust in and adoption of online recommendation agents. *Journal of the AIS*, 6, 03 2005.
10. A. V. Bodapati. Recommendation systems with purchase data. *Journal of Marketing Research*, 45(1):77–93, 2008.
11. Y. M. Brovman, M. Jacob, N. Srinivasan, S. Neola, D. Galron, R. Snyder, and P. Wang. Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 199–202, 2016.
12. R. Burke, N. Sonboli, and A. Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214, 2018.
13. P. Y. Chau, S. Y. Ho, K. K. Ho, and Y. Yao. Examining the effects of malfunctioning personalized services on online users' distrust and behaviors. *Decision Support Systems*, 56:180–191, 2013.
14. H.-H. Chen, C.-A. Chung, H.-C. Huang, and W. Tsui. Common pitfalls in training and evaluating recommender systems. *SIGKDD Explor. Newsl.*, 19(1):37–45, 2017.
15. M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 456–464, 2019.
16. Y. Chen and J. F. Canny. Recommending ephemeral items at web scale. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1013–1022, 2011.
17. P. Cremonesi, F. Garzotto, and R. Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *Transactions on Interactive Intelligent Systems*, 2(2):1–41, 2012.
18. A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 271–280, 2007.

19. J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The YouTube Video Recommendation System. In *Proceedings of the 4th Conference on Recommender Systems, RecSys '10*, pages 293–296, 2010.
20. M. B. Dias, D. Locher, M. Li, W. El-Deredy, and P. J. Lisboa. The value of personalised recommender systems to e-business: A case study. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 291–294, 2008.
21. M. A. Domingues, F. Gouyon, A. M. Jorge, J. P. Leal, J. Vinagre, L. Lemos, and M. Sordo. Combining usage and content in an online recommendation system for music in the long tail. *International Journal of Multimedia Information Retrieval*, 2(1):3–13, 2013.
22. D. Elsweiler, C. Trattner, and M. Harvey. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 575–584, 2017.
23. M. Ferrari Dacrema, S. Boglio, P. Cremonesi, and D. Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems*, 39(2), 2021.
24. M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pages 101–109, 2019.
25. S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.
26. D. Fleder and K. Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712, 2009.
27. A. Friedman, B. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky. Privacy aspects of recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Recommender systems handbook*, pages 649–688. Springer Nature, second edition, 2015.
28. F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 169–176, 2014.
29. A. Ghoshal, S. Kumar, and V. Mookerjee. Impact of recommender system on competition between personalizing and non-personalizing firms. *Journal of Management Information Systems*, 31(4):243–277, 2015.
30. C. A. Gomez-Uribe and N. Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *Transactions on Management Information Systems*, 6(4), 2015.
31. A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette. Offline evaluation to make decisions about playlist recommendation algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 420–428, 2019.
32. D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.
33. N. Hazrati, M. Elahi, and F. Ricci. Simulating the impact of recommender systems on the evolution of collective users' choices. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 207–212, 2020.
34. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *Transactions on Information Systems*, 22(1):5–53, Jan. 2004.
35. K. Hosanagar, D. Fleder, D. Lee, and A. Buja. Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. *Management Science*, 60(4):805–823, 2014.
36. A. Iovine, F. Narducci, and G. Semeraro. Conversational recommender systems and natural language: A study through the ConveRSE framework. *Decision Support Systems*, 131:113250–113260, 2020.
37. D. Jannach and G. Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 7–10, 2016.
38. D. Jannach and G. Adomavicius. Price and profit awareness in recommender systems. In *Proceedings of the 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation (VAMS) at RecSys 2017*, 2017.

39. D. Jannach and C. Bauer. Escaping the mcnamara fallacy: Towards more impactful recommender systems research. *AI Magazine*, 41(4):79–95, Dec. 2020.
40. D. Jannach and K. Hegelich. A case study on the effectiveness of recommendations in the mobile internet. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '09, pages 205–208, 2009.
41. D. Jannach and M. Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems*, 10(4), Dec. 2019.
42. D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25(5):427–491, 2015.
43. D. Jannach, M. Ludewig, and L. Lerche. Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts. *User Modeling and User-Adapted Interaction*, 27(3):351–392, 2017.
44. D. Jannach, P. Resnick, A. Tuzhilin, and M. Zanker. Recommender systems - beyond matrix completion. *Communications of the ACM*, 59(11):94–102, 2016.
45. D. Jannach, M. Zanker, M. Ge, and M. Gröning. Recommender systems in computer science and information systems - a landscape of research. In *Proceedings of the International Conference on Electronic Commerce and Web Technologies, EC-WEB '12*, pages 76–87, 2012.
46. I. Kamehkhosh, G. Bonnin, and D. Jannach. Effects of recommendations on the playlist creation behavior of users. *User Modeling and User-Adapted Interaction*, 30:285–322, 2019.
47. J. Katukuri, T. Könik, R. Mukherjee, and S. Kolay. Recommending similar items in large-scale online marketplaces. In *IEEE International Conference on Big Data 2014*, pages 868–876, 2014.
48. E. Kirshenbaum, G. Forman, and M. Dugan. A Live Comparison of Methods for Personalized Article Recommendation at Forbes.com. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases*, ECMLPKDD'12, pages 51–66, 2012.
49. B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22:441–504, 2012.
50. J. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.
51. R. Lawrence, G. Almasi, V. Kotlyar, M. Viveros, and S. Duri. Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery*, 5(1):11–32, 2001.
52. D. Lee and K. Hosanagar. Impact of recommender systems on sales volume and diversity. In *Proceedings of the 2014 International Conference on Information Systems*, ICIS '14, Dec. 2014.
53. D. Lee and K. Hosanagar. How do product attributes and reviews moderate the impact of recommender systems through purchase stages? *eBusiness & eCommerce eJournal*, 2018.
54. J. Lin. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51, Jan. 2019.
55. A. Maksai, F. Garcin, and B. Faltings. Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 179–186, 2015.
56. T. N. Nguyen and F. Ricci. A chat-based group recommender system for tourism. *Information Technology & Tourism*, 18(1-4):5–28, 2018.
57. U. Panniello, M. Gorgoglione, and A. Tuzhilin. Research note—in carss we trust: How context-aware recommendations affect customers' trust and other business performance measures of recommender systems. *Information Systems Research*, 27, 01 2016.
58. P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, 2011.
59. P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th Conference on Recommender Systems (RecSys '11)*, pages 157–164, 2011.

60. M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL'17*, pages 498–503, 2017.
61. S. Rendle, L. Zhang, and Y. Koren. On the difficulty of evaluating baselines: A study on recommender systems. 2019. arXiv:1905.01395.
62. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186, 1994.
63. M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 131–141, 2020.
64. M. Rossetti, F. Stella, and M. Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 31–34, 2016.
65. H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, and C. Trattner. Towards health (aware) recommender systems. In *Proceedings of the 2017 International Conference on Digital Health*, pages 157–161, 2017.
66. G. Shani, D. Heckerman, and R. I. Brafman. An MDP-Based Recommender System. *Journal of Machine Learning Researh*, 6:1265–1295, 2005.
67. E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating Similarity Measures: A Large-scale Study in the Orkut Social Network. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 678–684, 2005.
68. P. Symeonidis, A. Janes, D. Chaltsev, P. Giuliani, D. Morandini, A. Unterhuber, L. Coba, and M. Zanker. Recommending the video to watch next: An offline and online evaluation at youtv.de. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, page 299–308, New York, NY, USA, 2020. Association for Computing Machinery.
69. A. Tuzhilin. Personalization: The state of the art and future directions. *Business Computing*, 3(3):3–43, 2009.
70. F. Wall. Agent-based modeling in managerial science: an illustrative survey and study. *Review of Managerial Science*, 10(1):135–193, 2016.
71. W. Wobcke, A. Krzywicki, Y. Sok, X. Cai, M. Bain, P. Compton, and A. Mahidadia. A deployed people-to-people recommender system in online dating. *AI Magazine*, 36(3):5–18, 2015.
72. Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin. Modeling professional similarity by mining professional career trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1945–1954, 2014.
73. L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, and D. Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 279–287, 2018.
74. K.-H. Yoo, U. Gretzel, and M. Zanker. *Persuasive recommender systems: conceptual background and implications*. Springer Science & Business Media, 2012.
75. M. Zanker, M. Bricman, S. Gordea, D. Jannach, and M. Jessenitschnig. Persuasive online-selling in quality and taste domains. In *Proceedings of the 7th International Conference on E-Commerce and Web Technologies*, EC-Web '06, pages 51–60, 2006.
76. M. Zanker, L. Rook, and D. Jannach. Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies*, 131:160–168, 2019.
77. J. Zhang, G. Adomavicius, A. Gupta, and W. Ketter. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*, 31:76–101, 2020.
78. H. Zheng, D. Wang, Q. Zhang, H. Li, and T. Yang. Do clicks measure recommendation relevancy?: An empirical user study. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 249–252, 2010.