Received: 15 April 2021 | Accepted: 18 May 2021

DOI: 10.1002/aaai.12056

SPECIAL TOPIC ARTICLE

Explainability in music recommender systems

Darius Afchar[1] | **Alessandro B. Melchiorre**[2] | **Markus Schedl**[2] | **Romain Hennequin**[1] | **Elena V. Epure**[1] | **Manuel Moussallam**[1]

[1]Deezer Research, Paris, France

[2]Johannes Kepler University and Linz Institute of Technology, Linz, Austria

Correspondence
Darius Afchar, Deezer Research, Paris, France.
Email: dafchar@deezer.com

Alessandro B. Melchiorre, Johannes Kepler University and Linz Institute of Technology, Linz, Austria.
Email: alessandro.melchiorre@jku.at

Funding information
Austrian Science Fund, Grant/Award Numbers: DFH-23, P33526; State of Upper Austria and the Federal Ministry of Education, Science, and Research, Grant/Award Number: LIT-2020-9-SEE-113

Abstract

The most common way to listen to recorded music nowadays is via streaming platforms, which provide access to tens of millions of tracks. To assist users in effectively browsing these large catalogs, the integration of music recommender systems (MRSs) has become essential. Current real-world MRSs are often quite complex and optimized for recommendation accuracy. They combine several building blocks based on collaborative filtering and content-based recommendation. This complexity can hinder the ability to explain recommendations to end users, which is particularly important for recommendations perceived as unexpected or inappropriate. While pure recommendation performance often correlates with user satisfaction, explainability has a positive impact on other factors such as trust and forgiveness, which are ultimately essential to maintain user loyalty.

In this article, we discuss how explainability can be addressed in the context of MRSs. We provide perspectives on how explainability could improve music recommendation algorithms and enhance user experience. First, we review common dimensions and goals of recommenders explainability and in general of eXplainable Artificial Intelligence (XAI), and elaborate on the extent to which these apply—or need to be adapted—to the specific characteristics of music consumption and recommendation. Then, we show how explainability components can be integrated within a MRS and in what form explanations can be provided. Since the evaluation of explanation quality is decoupled from pure accuracy-based evaluation criteria, we also discuss requirements and strategies for evaluating explanations of music recommendations. Finally, we describe the current challenges for introducing explainability within a large-scale industrial MRS and provide research perspectives.

Darius Afchar and Alessandro B. Melchiorre contributed equally to this study.

MUSIC RECOMMENDER SYSTEMS

Recommender system (RS) technology permeates our daily lives. Like Recommender Systems (RSs) in other domains (Ricci, Rokach, and Shapira 2015), music recommender systems (MRSs) (Schedl et al. 2015) have information filtering algorithms at their core, which select from a commonly huge catalog of music items (e. g., artists, albums, or songs) those identified as most relevant for a target user. Thus Music Recommender Systems (MRSs) guide users in the otherwise sheer overwhelming amount of music available at their fingertips nowadays.[1]

The raising awareness of, and ongoing discussion about, transparency of Machine Learning (ML) algorithms, including those used in RSs, has resulted in a substantial demand from users to receive explanations for why certain items have been recommended to them (Zhang and Chen 2020). Also from a RS provider's perspective, these aspects are important for building and maintaining trust of the users in the system. Therefore, equipping MRSs with capabilities to provide explanations to their users is of mutual interest.

Characteristics of music consumption and music recommender systems

While music recommendation shares some properties with other media recommendation tasks, such as videos or movies, there also exist pronounced differences. Among the ones identified in literature (e. g., Schedl et al. 2018), the following characteristics are relevant for explainability in MRSs, as we will elaborate in the subsequent sections:

- The *duration* of item consumption is commonly much shorter than in other domains, that is, songs have typical lengths of several minutes, whereas movies or even books take a much longer time to consume.
- Music data come in *manifold representations*, including audio, MIDI, and textual metadata (e. g., editorial metadata but also user-generated tags). Furthermore, music-related data that can be leveraged in MRSs is highly multimodal and includes images (e. g., album covers) and videos (e. g., music video clips), next to audio and textual metadata. Finally, user feedback is collected from various activities (e. g., *likes*, *favorites*, *song skips*).
- The listening *context* strongly affects music preferences. For instance, the listener's mood, location, and social situation (e. g., alone vs. with friends) have been shown to influence musical needs and demands (Ferwerda, Schedl, and Tkalcic 2015; Rentfrow, Goldberg, and Zilca 2011).
- Music is often consumed sequentially, that is, tracks in a listening session or playlist. Therefore, for music, we often focus on sequential recommendation tasks, such as automatic playlist creation or continuation (Bonnin and Jannach 2014; Zamani et al. 2019), that leverage both long- and short-term user preferences.

Common music recommendation tasks and methods

Various use cases of MRSs exist, centered around different tasks. Among these, the most important ones are *front page recommendation* (recommending content for thematic collections of music—also known as shelves or channels—presented to the user on the front page of the platform's user interface) (Bendada et al. 2020), *music exploration/discovery* (e. g., based on item similarity in terms of melody, rhythm, or lyrics) (Goto and Dannenberg 2019; Knees, Schedl, and Goto 2020), *automatic playlist generation* (commonly based on the user profile, but possibly only based on a seed description such as "music to relax"), and *automatic playlist continuation* (based on a sequence of seed tracks) (Jannach, Lerche, and Kamehkhosh 2015; Zamani et al. 2019).

To create a music recommendation engine, a variety of methods are adopted, depending on the use case. These include latent factor models (e. g., singular value decomposition or factorization machines), graph mining techniques (e. g., random walks or graph embeddings), and deep learning-based techniques (e. g., convolutional neural networks, recurrent neural networks, or autoencoders). Furthermore, techniques from audio signal processing and natural language processing are often used to create vector representations of music items or to annotate music items with relevant tags.

How these introduced characteristics, tasks, and methods inform explainability strategies of MRSs will be detailed in subsequent parts of the article. Examples and challenges in an industrial context will be provided in the last part.

GOALS AND DIMENSIONS OF EXPLAINABILITY FOR MUSIC RECOMMENDER SYSTEMS

Recent years have seen an upsurge of interest in explainable recommendations, even though the concept emerged in the 2000s (Herlocker, Konstan, and Riedl 2000). This evolution of explainable RSs has been accompanied by an increasing popularity of eXplainable Artificial Intelligence (XAI), with which it shares roots, approaches, and

terms. eXplainable Artificial Intelligence (XAI) represents the convergence of many research disciplines, including computer science, human–computer-interaction, philosophy, and psychology. Coherent and stable XAI definitions and terms have started to appear only recently (Arrieta et al. 2020; Guidotti et al. 2018; Lipton 2018). Meanwhile, RSs research has developed explanation-related concepts that are unknown to general XAI, some of which, however, rest upon elusive descriptions. This specificity is probably due to the nature of RSs themselves, which differ w. r. t. their tasks, inputs, and results from general trends in XAI. Linking these two explainability realms would not only result in a more standardized approach to explanations in RSs but also in a direct application of methods from XAI to MRS.

In this section, we review definitions and concepts of explainability in RSs. Subsequently, we compare and connect them with the ones of XAI. Note that this is not a survey of XAI or explainable RSs as other valuable resources exist on this matter (Arrieta et al. 2020; Guidotti et al. 2018; Nunes and Jannach 2017; Zhang and Chen 2020).

Definitions and goals of explainability for MRS

What does it mean to *explain a recommendation*? Within the RS field, Tintarev and Masthoff (2015) addresses this question with "to *make clear by giving a detailed description*," and Zhang and Chen (2020) with " *an explainable recommendation aims to answer the question of why*." We can thus discern a role of explanations as complementary information to the recommendation. But these definitions are limited; for instance, ensuring *fair recommendations* involves tracing the "why" of a recommendation, but only regarding certain critical aspects (e. g., potential gender biases) and it does not tell how to act upon them. As we develop next, "complementary information" and "fair recommendations" shape two of the many facets of explainability.

Borrowing general ideas from recent harmonization efforts of XAI terms, it is more convenient to distinguish between explanation objects and goals. *Explanations* designate the result of an explanation system, they form an "*interface between the system to explain and a target audience*" (Guidotti et al. 2018). Quite interchangeably with *explainability*, we will use the term *interpretability*, with a more passive characteristic: a system can be interpretable—for example, decision trees are often interpretable, neural networks are not. The opposite notion is *blackboxness*. We stress that automatically concluding that trees and linear regressions are interpretable and that neural networks are not is questionable.[2] As we will see next, this depends on a precise formulation of explanation tasks that do not admit one-size-fits-all rules.

The previous mention of "audience" is essential, since a given explanation type may only convey meaningful information to specific people. In RS research, the target audience of explanations is usually end-users as they are the targets of the recommendation decision and could be skeptical about it. Nevertheless, other stakeholders may be interested in receiving explanations, for example, system designers and data scientists may inquire wherever their system bases its decisions on discriminatory biases from the data.

We shall continue with a cautionary tale: the disparate notions of explainability have led to many misuses of XAI (Lipton 2018). Because we usually do not have access to ground-truth explanations in the wild, and realistically will not in industrial contexts, many XAI works have relied on intuitive notions of what their target explanations should be. This first makes evaluation difficult. As Doshi-Velez and Kim (2017) highlights, the relevance of explanations is often suggested in a "*you'll know it when you see it*" fashion, which paves the way to many *confirmation biases*. Second, several counter-intuitive results have been unveiled. For instance, the widely agreed-upon idea that an interpretable model is more desirable than a blackbox one has been challenged: produced explanations—similarly to model predictions—may be misleading or biased (Adebayo et al. 2018 ; Dinu, Bigham, and Kolter 2020; Kaur et al. 2020; Rudin 2019). Moreover, without clear formulations of explanation tasks, how can several XAI systems be compared? Can we actually quantify interpretability and explanation quality? Can the relevance of proposed interpretation metrics be assessed? How can we detect misinterpretations and explanations based on spurious mechanisms? All those questions circle back to the definition of explanations.

In addressing these questions, the concept of *incompleteness* was proposed. Its purpose is to characterize the "missing piece" justifying the use of an explanation system (Doshi-Velez and Kim 2017). Here, the literature of explainable RS and general XAI diverges. A distinction of the goals of RS explanations is proposed in Tintarev and Masthoff (2015). We can enrich this discussion with goals identified in general XAI by Arrieta et al. (2020). Both sets of goals are displayed in Figure 1 with short definitions. We find that neither of the two may solely account for all MRS purposes: explainable RS goals mostly fall into the *informativeness* category. This has a broader scope than RS's *transparency* that feels, too focused on the decomposition of models' inner mechanisms. Furthermore, RS goals have been found to be intercorrelated (Balog and Radlinski 2020), in particular, *satisfaction* being

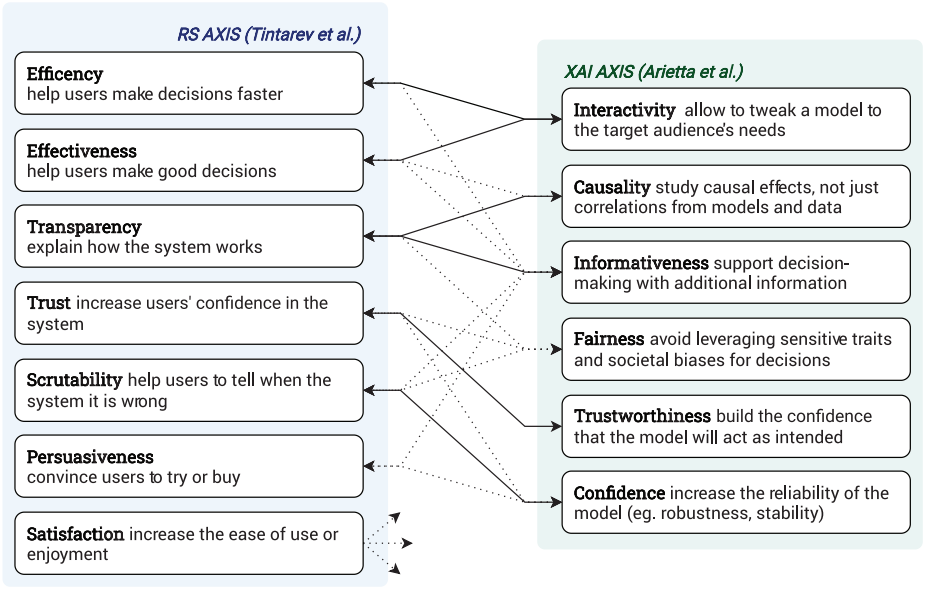

FIGURE 1 RS and XAI explanation goals and linking. We display "one-liner" definitions for conciseness. A solid line indicates a strong correspondence, a dotted line a weaker one that depends on the exact task and context. *Satisfaction* could have been linked to everything but links are omitted for clarity.

arguably a desired byproduct of any explanation method. That said, *persuasiveness* is a strong dimension of RS that is absent from general XAI (Ehrlich et al. 2011); when aiming at transparency, creating a persuasive system may appear contradictory.

Identifying those goals is crucial, because explanations may simply not be needed if incompleteness is not an issue. Evaluation should then be conducted with regard to each targeted incompleteness, to avoid mismatched objectives. Lastly, the concept of *understandability* simply bridges the gap between a chosen XAI system and this new notion of goal/incompleteness being addressed for a target audience. All these notions are illustrated and placed accordingly in Figure 2. We discuss additional taxonomic axes for XAI in subsequent paragraphs.

As a final note, explainability can be framed through an interesting take from Michael Jordan on the future of ML.[3] The goal of XAI is not only for decision-makers to understand model predictions, but to allow a back and forth interaction between the two. *Why* do you make this decision? *What if* this aspect was different? *Then* what if this aspect was different? "*Consequential decision involves thinking about new facts that were never put in the original data, that are relevant to the current situation."* The purpose of the discussion on *incompleteness* and *understandability* is to go beyond the view of explainability as a mere complementary prediction, but to allow this reciprocal gain of knowledge between several actors.

Local/Global scope

As MRSs commonly provide numerous recommendations, there is a focal distinction to make in the explanations scope: local ersus global (Doshi-Velez and Kim 2017). *Local* or instance-wise explanations target the decision of the model for a specific input-recommendation pair, for example, explaining that a track was recommended to an end-user because some of its features matched. Local explanations must be tailored to each individual prediction. This type of explanation is aligned with the European General Data Protection Regulation (GDPR) "Right to explanation" (Parliament 2016), which entitles users to inquire about the reasoning behind the outcome of an algorithm, hence supporting informativeness as an explainability goal.

In contrast, *global* explanations provide a big picture of the model logic, covering multiple model decisions. For instance, estimating clusters of model-learned user embeddings may help rationalize the behavior of the MRS within several general communities. This broad view of the model is necessary to detect systematic biases of the model (addressing fairness goals) and to examine wherever a model is suitable for deployment (addressing informativeness, trustworthiness, and confidence). Lastly, note that the two types may be linked: it is sometimes relevant to craft a global explanation by providing multiple local explanations (Ribeiro, Singh, and Guestrin 2016).

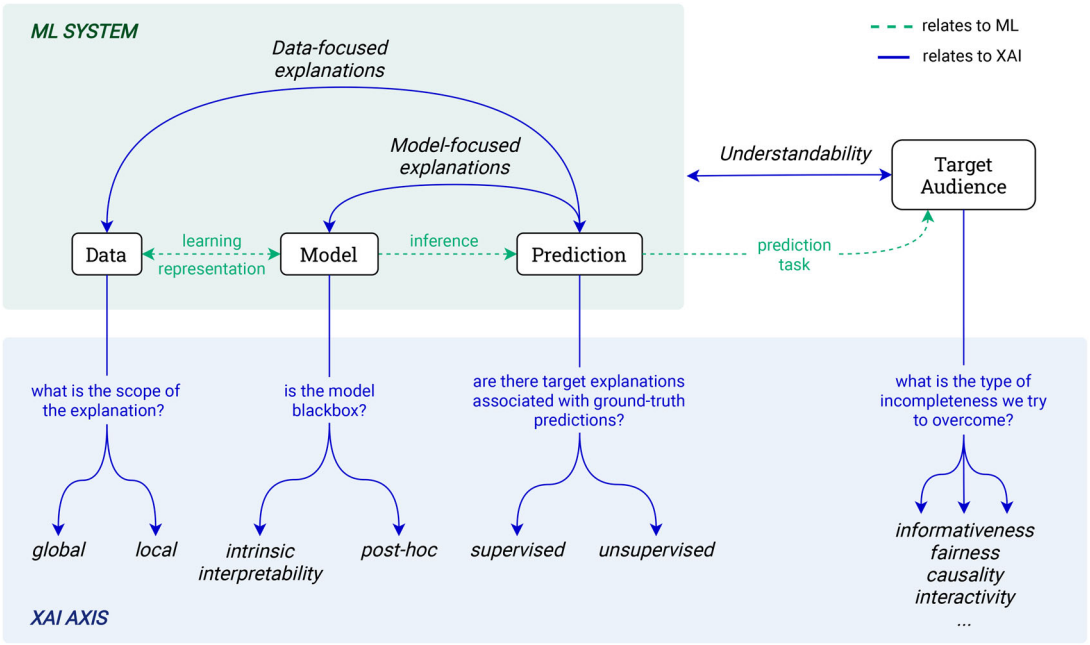

FIGURE 2 Overview of XAI notions. In the upper part, a ML model is trained on data and used to make predictions. Beyond prediction, if the model alone is insufficient w. r. t. an underlying human-grounded application, the use of an XAI method will be justified. The specification of the target audience delineates incompleteness to be addressed through explanations, along different explanation axis (lower part).

Intrinsic/Post hoc interpretability

We can also distinguish explanation systems w. r. t. whether interpretability should be an inherent part of the RS–*intrinsic* interpretability, or should be provided as an addition to an already working RS–*post hoc* interpretability.

Intrinsic interpretability refers to the ability of the RS to provide sufficient information to make its inner functioning clear to a specific audience (Arrieta et al. 2020). In this case, the explanations coincide with the model. Being inherent in the model, intrinsic interpretability has to be planned in advance, making it a component of the model design. For instance, an Item-*k-Nearest Neighbors* model recommends artists because they are similar to the ones the user listened to, thus allowing explanations such as "*We recommend you* `<artist>` *because it is similar to* `<artist(s)>`."

Post hoc or extrinsic interpretability refers to the use of external XAI to yield knowledge from a blackbox model. It can be considered as *reverse engineering* the model (Guidotti et al. 2018). For example, the recommendations of a blackbox model can be explained by making a post hoc selection of the relevant features that lead to the recommendation; they offer explanations such as "*We recommend you this because it has* `<feature(s)>` *you may like*." Both intrinsic and post hoc views are affiliated with the concept of transparency, thus supporting informativeness, causality, and confidence. However, post hoc explanations are hampered by their externalness and require an additional check of their faithfulness to the studied model. Yet, compared to intrinsic, they disentangle model design from explanation design, allowing to consider XAI systems in a later stage, or to apply them to already working models.

Un/supervised explanations

We often think of XAI methods as being unsupervised. Particularly on the end-user side, it is arduous to guess which could be the *ground-truth* explanations for the user since their judgment of what a good explanation is may be biased (Miller 2019). Nevertheless, target explanations are sometimes available (Balog and Radlinski 2020). Our goal is not only to make explanation predictions but to address an incompleteness; the relevance of the target predictions thus has to be questioned. Do these really address our needs w. r. t. to incompleteness? Or are they a proxy for it? In the latter case, how do we assert/evaluate their *understandability* w. r. t. our goal? We present two ideas from XAI for supervised explanations in the image domain that could be applied to MRS.

In the field of image classification, some datasets gather images with textual descriptions. Each set of words can be matched against corresponding visual aspects in

the images, enabling to generate *visual explanations* for unseen instance class predictions through RNN-generated texts (Hendricks et al. 2016). The explanations are evaluated against held-apart test descriptions. Here, the concept of explanation is driven by two desiderata, first as a way to link different modalities of a same object—image and text, and second as a rationale that conveys useful information by yielding class-specific information that differentiates it from other classes. Obtaining this informative *discriminative* quality is tricky in an unsupervised setting. The multimodality of music data (e. g., audio, lyrics, users, playlists) makes it a good candidate for this paradigm.

We can identify another line of supervised explanations as linking different conceptual levels. The *TCAV* method (Kim et al. 2018), for instance, allows to check predictions against human-understandable concepts, for example, how much the model prediction for an image of a zebra is sensitive to "stripeness." Again, there is an interesting link to music: there is a known and unresolved *semantic gap* between low-level data (i. e., audio signal) and its correspondence to high-level descriptions (e. g., genre, mood) (Celma, Herrera, and Serra 2006).

Model/Data

We conclude this section with a paramount yet subtle distinction that is prone to be overlooked: Are the explanations related to the RS model processing or to the data it represents?

Model explanations, on one side, focus on a model-learned representation and parameters and aim at making sense out of it. With a mild exaggeration, to the question "*why is this track recommended by the MRS given my history?*" a model-focused answer for a RS might be "*it maximizes the probability of being co-listened with your history, considering all other users listening history*." *Data explanations*, on the other side, would rather focus on "*why are those items co-listened in the first place?*" The trained model by itself is less interesting than the goal of uncovering "natural mechanism in the world" (Chen et al. 2020). In practice, in the first case, the model inspection may expose irregularities and lead to adjust its architecture and regularization (e. g., balancing fairness trade-off parameters); in the second case, the model plays the role of a proxy representation of data, detected errors would more suitably be attributed to a misrepresentation of input data (e. g., feature engineering for a better matrix factorization), and the finality is to find a structure that is credible given prior knowledge of the problem.

These aspects are often entangled. Explaining the model provides little information with noisy data, and explaining the data may be misleading if the model assumptions do not capture salient aspects (e. g., correlation instead of causation). It is a widespread fallacy to explain a model (which is often easier, particularly when using interpretable models), when the objective is to explain data.

MAKING MUSIC RECOMMENDER SYSTEMS EXPLAINABLE

In the previous section, we have drawn links between explainability in RSs and XAI, and presented different definitions. Bearing these definitions in mind, we now study different ways MRSs can be made more explainable. We start with a general overview of possible explanation methods for MRSs, then discuss the adaptability of three relevant explanation paradigms to MRSs.

Overview of explanation methods for MRSs

We want to provide the reader with a short background on existing explanations methods for RSs, and then discuss how the latter are particularized for MRS.

Explanations of RS

Zhang and Chen (2020) characterize six RSs explanation types. First, *relevant item or user* explanations, also called example-based explanations, are bonded with item-based or user-based collaborative filtering. Thus, a recommendation is motivated either by the similarity of the item to other items previously liked by the user, or by the affinity that similar users have towards the recommended item.

Second, there are *feature-based* explanations, which are associated with content-based recommendation algorithms. Explanations are commonly shown as tags relevant to a user or an item (Zhang and Chen 2020). *Opinion-based* explanations focus on relevant aspects of the recommended item (Zhang et al. 2014; Wang et al. 2018), which can be enriched with a sentiment (Zhang et al. 2014). In contrast to feature-based explanations leveraging item metadata or user profiles, opinion-based aspects are mined from reviews or social media posts.

Further, we also distinguish sentence, visual, and social explanations. *Sentence* explanations can be predefined templates with placeholders regarding features or aspects/opinions filled on-the-fly depending on the recommendation or specific user (e. g., "*We recommend this item because its* `[good/excellent]` `[feature]` *matches with your* `[emphasize/taste]` *on* `[feature]`") (Wang et al. 2018). Alternatively, sentence explanations can be generated from scratch using language models trained on

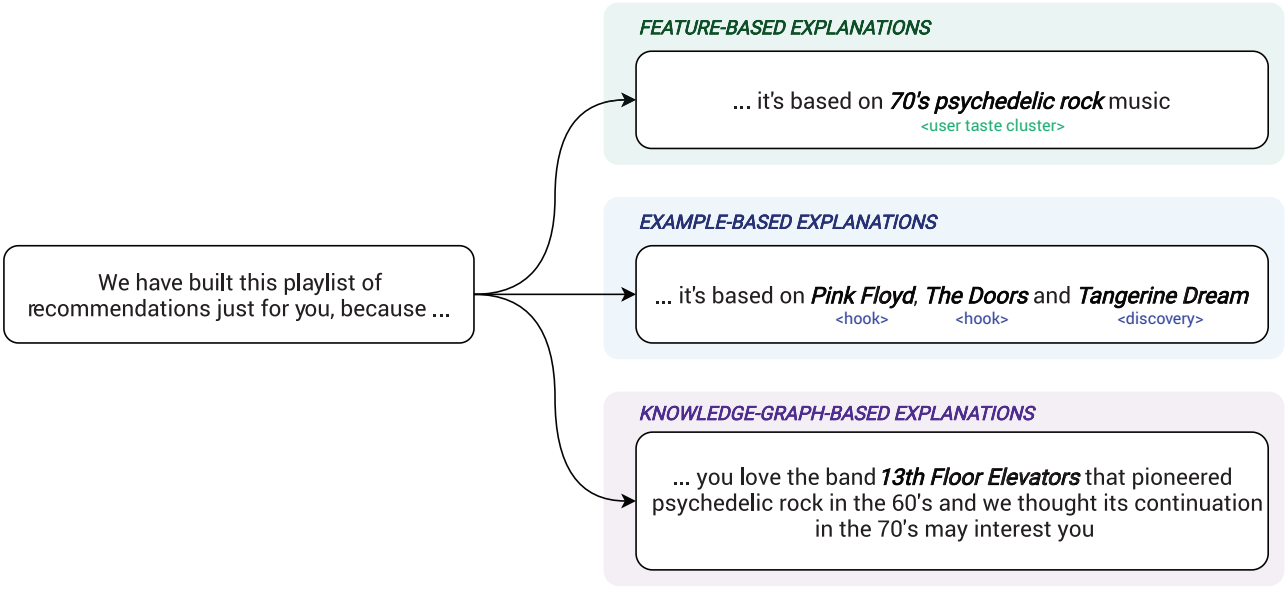

FIGURE 3 Summary of explanation types. We show some *informativeness*-oriented explanation examples that may be provided to an *end-user* for which a personalized playlist has been generated.

reviews (Costa et al. 2018). *Visual* explanations appear as images or visual elements often accompanied by text (Andjelkovic, Parra, and O'Donovan 2016; Chen et al. 2019). Image regions or caption words that explain the recommendation could be highlighted (Chen et al. 2019). *Social* explanations mention either the user's friends who liked the recommended item (Sharma and Cosley 2013) or their overall number.

Extension to MRSs

With voice assistants becoming increasingly popular, researchers are investigating audio-based explanations in MRSs. A first line of work proposes *listenable explanations* (Behrooz et al. 2019), inspired from radio shows in which hosts provide information about played tracks for creating transitions. Alternatively, item parts such as track snippets focusing on a particular audio source (e. g., instrument or voice Melchiorre et al. 2021) can be emphasized as reasons for recommendation.

Whenever recommendations are provided as collections of items (e. g., playlists), explanation generation can be modeled as playlist captioning (Choi, Khlif, and Epure 2020) or playlist stories generation (Behrooz et al. 2019). Existing work usually relies on predefined textual templates (Behrooz et al. 2019).

Music explanations are rarely informed by a unique data source. Knowledge graphs (KGs) are constructed from external sources and used for explanations (Oramas et al. 2016). Information sources leveraged in existing work are: user-generated text such as music descriptions (Zhao et al. 2019), existing knowledge bases like MusicBrainz or Wikipedia (Oramas et al. 2016), tags describing items or users (Kouki et al. 2019; Zhao et al. 2019), social information such as users' friends (Kouki et al. 2019; Sharma and Cosley 2013), audio features (Andjelkovic, Parra, and O'Donovan 2016), or pretrained tag embeddings (Andjelkovic, Parra, and O'Donovan 2016).

We next discuss in detail feature-based explanations, example-based explanations, and graph-based explanations. We refer to Figure 3 for examples of each explanation type.

Feature-based explanations

MRSs rely on multimodal information (or features) in order to provide personalized recommendations to users. It is, therefore, legitimate to ask *which features are most responsible for the generated recommendation*. Feature-based explanations aim at answering this question by identifying a minimal subset of features that are relevant for the recommendation. For instance, such explanation may be "*We recommend you this song because it is '90s rock, a combo of era and genre you enjoy listening to*." where the *genre* and *era* represent the relevant features.

Relevance

Feature-based explanations are only relevant if the features are themselves interpretable. Furthermore, feature

selection is an NP-hard problem (Natarajan 1995) and real-word applications necessarily rely on feature assumptions: for example, a limited number of interacting features (Chen et al. 2018), group or structure coherency (Afchar and Hennequin 2020), feature independence (Ribeiro, Singh, and Guestrin 2016), or first order approximations (Simonyan, Vedaldi, and Zisserman 2014).

Applications

Frequently, considered features are selected and ranked through a relevance score. More than just the top-contributing features, displaying or visualizing all the scores is a common practice among data scientists, acting as an encompassing explanation (Kaur et al. 2020), though the information overload may be misleading (Poursabzi-Sangdeh et al. 2021). Note that "relevance" for a feature is a polysemous term that inherently depends on the used selection method. As illustration, both SHAP (Lundberg and Lee 2017) and L2X (Chen et al. 2018) assign relevance scores to single features, however, while SHAP expresses relevance in terms of marginalized contributions of features across all possible subsets, L2X encodes relevance as a notion of informativeness on the response variable through maximizing mutual information. We refer to Covert, Lundberg, and Lee (2020) and Sundararajan and Najmi (2020) for surveys and further details on selection methods.

Applied to MRSs, feature explanations may be related to users, items, to the context, or a combination of the previous. *User features* stretch from reasonably static characteristics (e. g., country of origin, age group, personality) to constantly changing traits (e. g., tastes, recent interests, mood). These features offer a fertile ground for tailored recommendations, and thus tailored explanations such as "*We recommend you this track because it suits your current emotional state*" or "*… because of your country of origin*." However, the effectiveness of these explanations may be hindered by unreliable estimates of some user's variables, notably dynamic ones. Fairness-wise, societal biases in RSs often stem from the usage of sensitive user features, an analysis of their impact on recommendations being crucial to be able to temper with them.

Alternatively, *item features* are usually more objective and come in different types and granularity. For example, audio features cover low-level features (e. g., spectrograms, beats) (Müller 2015), mid-level features (e. g., articulation, melodiousness) (Aljanaki and Soleymani 2018), or high-level features (e. g., danceability, emotion) (Kim et al. 2020; Melchiorre and Schedl 2020), as well as metadata (e. g., genre, social tags) (Knees and Schedl 2016). Explanations involving these features are strongly tied to content-based recommendation (Zhang and Chen 2020) as they directly match the user's preference profile (e. g., "*We recommend you this because it has the tempo/genre you like*").

Lastly, miscellaneous *context features* may be suited for generating personalized explanations. Time and location, being the most popular ones, provide a sound contextualization for the recommendation such as "*This techno masterpiece is perfect for tonight's Friday's party!*" or "*Since you are doing home-office these days, we recommend you this 'Work from Home' playlist*."

Evaluation

Feature-based explanations are tied to a chosen definition of relevance. Their approximations can be compared to full computation results, when affordable. However, it may be tricky to evaluate whether the relevance scores themselves translate to true relevance. What is relevant for a trained model indeed reveals correlated events in the data, with the risk of returning spurious relations instead of causal truth about the data. As another pitfall, many feature selection methods do not handle intercorrelated features well (Yu and Liu 2003), which are however common with MRSs.

Example-based explanations

In MRSs, example-based explanations are a very common type of explanation, that can be reduced to the use of the sentence template "*We recommend you* `<this new item>` *because of* `<its similarity>` *to* `<meaningful item(s)>`." They are conceptually tied to case-based RS.

Relevance

Similarly to feature-based explanations, they are only relevant if the given examples are themselves interpretable for the target audience. This includes returning items that are known to the user: for example, from a set of liked or previously interacted items, or from broadly known items.

Applications

Regarding *examples types*, it is common to see *artist* examples as they convey a general sense of genre, temporal period, or style. Relevant-user examples were popular in the past decades as they have the interesting social twist of fostering users' curiosity to find pairs with similar tastes. They have gradually vanished from most music and video streaming platforms since they were found less

convincing and accurate than item-based explanations, and in turn may have a negative impact on *trustworthiness* (Herlocker, Konstan, and Riedl 2000). Nevertheless, limiting social explanations to close circles was found more relevant (e. g., "*recommended tracks recently discovered by your friends*"). Other than textual modalities, explanations in MRSs include displaying album covers, which may convey information about the style or even allow to recognize record labels (e. g., *Deutsche Grammophon*, *Blue Note*). Short audio thumbnails are also a promising way to provide explanations that cannot be otherwise expressed with words (Melchiorre et al. 2021).

As for *similarity relations*, we note it may not be explicitly stated in the explanation, or could even be statable. This is particularly true for RSs basing their recommendation on co-listening data. With the same causality counterpoints as before; a co-listening may be coincidental, confounded by external factors, or, more pragmatically, may result from noisy metadata and inattentive users. With deep learning models that compute nonlinear similarity metrics (e. g., the NeuCF method He et al. 2017), it gets trickier as we are faced with an added blackboxness issue.

This can lead to explanation examples that feel cryptic to the user. Recent works in Knowledge Graph (KG)-based recommendations are a way to alleviate this issue, we will see them next. Another lead lies in the disentanglement of the embeddings' latent dimensions that help rationalize proximity according to explicit concepts (e. g., audio features, genre, instrumentation) (Lee et al. 2020). Attention-based mechanisms are also a promising way of providing recommendations based on the selection of a reasonably small and contextual subset of neighbors (Kang and McAuley 2018), though claims on the interpretability of attention are disputed (Serrano and Smith 2019).

Evaluation

It is useful to evaluate the *discriminativeness* of examples. Indeed, example-based explanations are affected by popularity biases, which hampers *informativeness*. As illustration, "*The Beatles*" are streamed by many users with diverse profiles, thus appearing in many co-listening relations and are likely to emerge as similar neighbors. But using them as examples consequently lacks representativeness. On the other end, examples of niche artists are double-edged: they can yield a powerful feeling of understanding of user taste, but if the recommended item fall too far off the example style—and as the user sensibly connects with it, it can feel deceitful.

Then, one should distinguish items coming from an explicit elicitation (e. g., liked artists) and implicit preferences. The former are often meaningful to users but may make them feel trapped in a recommendation bubble, while the latter are more diverse but potentially lack direct connection to users, affecting trust in explanations.

Examples can also be useful for persuasiveness goals. It may be interesting, for instance, to provide a set of examples that include an item that is well-known to the user (acting as a *hook*) and unknown or weakly interacted ones (acting as discoveries). This principle is quite common in radio "clock" programming, where alternating *power songs* and discoveries has been shown to be a powerful tool to keep users engaged.

Graph-based explanations

Canonically, RSs match users and items. It is, therefore, not surprising that graph-based approaches on *bipartite graphs* can be used, with users on one side and items on the other. The recommendation task may indeed be framed as *link prediction*: given links of interactions between users and items, which unseen links are then probable? Likewise, similar item recommendation can be formulated as the task of finding probable nearest neighbors in a graph of users.

Relevance

This framing can seem cumbersome, with a first strong hindrance that graph methods quickly get computationally expensive—though some works have demonstrated industrial-scale applicability (Ying et al. 2018). Second, notions of repeated music consumption, preference decay, and accounting for a temporal dimension for sequential recommendations are tricky to incorporate into graphs. Nevertheless, graph-methods possess an outstanding expressive power, especially for multi-relational data, enabling abundant new RS applications.

Further justification lies in the graph structure naturally found in MRS data. *Vertically*, there is a natural hierarchy for musical items: tracks are organized into albums, that are themselves children of artists, that can be regrouped into genre, style, and time period, or any complex multileveled music ontology. Users exhibit a similar hierarchy, we can often assign them to several clusters of interest, that are themselves linked to a given culture, country, or age category. *Horizontally*, music item clusters act as islands of connected components, with central nodes being representative of a given style and having influence on surrounding artists. Weakly connected nodes denote niche artists, and nodes in-between clusters fuse several influences. The same reasoning may apply to users' communities and hierarchies.

Applications

Detecting *cliques* and using *k-degeneracy* can help represent communities. *Tripartite* and generally *n*-partite formulations allow to generalize canonical recommendation by handling more actors than items and users, for example, considering artists and context. *Directional edges* can be leveraged to create graphs with asymmetrical relations and avoid recommending niche artists as being similar to very popular ones (Salha et al. 2019). Graph-specific *embedding* techniques may be applied, for example, using random walks to train embeddings on more diverse sequences than observed data (Grover and Leskovec 2016). Other approaches are promising, such as analysis of graph structure for *domain-transfer*; application of the *traveling salesman problem* to find fluid playlist tracks orderings; or *continual learning* by framing the addition of new users and items as new nodes that should not perturb far away regions of the graph. All these tools address the transparency issue, leading to more interpretable models.

Interpreting structure is mostly useful for an audience of researchers, but recent advances in the field of *knowledge-graph-based recommendations* show additional promising applications for end-users. The term KG refers to the use of external expert-knowledge to better understand the entities at hand within a RS task and how they relate to one another (Dong et al. 2014). In the context of MRSs, available knowledge may include intra-music relations (e. g., "*is sung by*," "*has music label*," "*belongs to genre*"), and collaborative information (e. g., "*often streamed with*," "*user taste belongs to cluster*"). The KG can be applied to enhance the representation of items before recommendation. For instance, the latent space that is usually learned to compactly represent items can be structured to align with each item relations of the graph (Bordes et al. 2013). However, RSs may still fail to leverage the full power of KGs, solely relying on enhanced representations. Instead, another approach is to directly incorporate KGs into the recommendation computation, which allows multi-hop reasoning. For that, all paths (with a fixed maximum length) between a pair of user and item can be extracted, and their relevance estimated (Wang et al. 2019). This enables to produce explanations corresponding to paths of high probability (e. g., a path $\text{User}_i \xrightarrow[\text{listened to}]{} A \xrightarrow[\text{sung by}]{} \text{artist}_A \xrightarrow[\text{belongs to}]{} \text{label}_L \xleftarrow[\text{belongs to}]{} \text{artist}_B \xleftarrow[\text{sung by}]{} B$ translates for user i as "*Track B is recommended to you because it's similar to A you listened to before, which is sung by an artist belonging to the same indie music label L as B.*"). For a complete survey of KG methods, we refer to Ji et al. (2021).

Evaluation

The efficiency of those techniques is conditioned on a good modeling of the involved entities (i. e., nodes and links), deep knowledge engineering, and an accurate estimation of the paths' relevance while ensuring their interpretability. As a counter-example, a generic "*similar to*" relation in a KG does nothing for *informativeness* as it is still a black-box information, no matter the transparent relations before and after in the path.

Multi-hop reasoning that is permitted by graphs is a great opportunity to enhance discovery, which is known to impact *effectiveness* and *satisfaction* of RSs (Castells, Vargas, and Wang 2011). But this requires crafting new metrics for relevance evaluation, which is still an open research topic (Ge, Delgado-Battenfeld, and Jannach 2010).

KGs are also a promising lead for *causality*, as they can allow to model and estimate causal structures for data.

Perspectives

Drawing inspiration from the recent success of GANs (Goodfellow et al. 2014), we could consider *generative explanations* in MRSs. In particular, assuming the audio content is available, a GAN-generated explanation may provide a listenable explanation of what the user tastes are like according to the model. Indeed, the explanation may be conditioned on some priors (Mirza and Osindero 2014), for example, what the user likes about metal or jazz, to provide reasonable explanations. However, these types of explanations are hampered by the demanding resources required to generate audios (Dhariwal et al. 2020).

Another interesting direction is exploiting human concepts of musical understanding (Kim et al. 2018). For example, to understand how much the concept of "rock" or "happy" matters for the recommendation to a specific user. Beyond informativeness, this may also lead to uncovering bias in the datasets (e. g., how much the concept of male artist matters for the recommendation).

Lastly, counterfactual or contrastive explanations not only pinpoint the causes of a model decision but also provides users with actionable levers to change the recommendation (Miller 2019; Ustun, Spangher, and Liu 2019). Among the explanation types, counterfactual explanations may be considered best compliant to the GDPR (Parliament 2016) as they can provide a refined framework for fairness (Kusner et al. 2017).

EVALUATING EXPLANATIONS

Evaluating MRS explanations is paramount to assess whether the explanation goals are met by the explanation

methods. We have discussed some evaluation aspects in previous sections, specific to particular explanation dimensions and categories of methods. While there exists no one-size-fits-all evaluation strategy, in the following, we provide some general guidelines, tailored to the target audience of the explanation (end user vs. technical stakeholders).

Evaluating explanation from the end-user's perspective

Most RSs explanations target end-consumers. One straightforward way to evaluate such explanations is to conduct user studies (Knijnenburg et al. 2012) and assess if the explanations allow to address the targeted goals. We have argued in previous sections that an explanation ground-truth is an evasive concept. Nevertheless, user studies can provide cues for what explanation types are best suited in specific domains, and can also detect practical misuses (Kaur et al. 2020).

In the context of MRSs, user studies showed that visual explanations increase understandability (Andjelkovic, Parra, and O'Donovan 2016) while social or sentence explanations are more persuasive (Sharma and Cosley 2013). However, providing too many details results in cognitive overload and is negatively perceived (Kouki et al. 2019). Also, persuasiveness does not necessarily correlate with the value recommendations have for the user. For instance, a user following an artist recommendation because a friend likes it does not necessarily result in the user liking the artist. One suggestion to overcome this is by corroborating different types of explanations (e. g., social with feature-based explanations) (Sharma and Cosley 2013). Another solution is to enable conversations between user and system, so recommendations could be gradually improved with system's explanations and user's feedback (Zhao et al. 2019).

User studies in MRSs are typically either between-subject or within-subject. Studies of the first type split users in two groups: one does receive the explanation, the other does not (Millecamp et al. 2019). Hence, we can naturally quantify the effect of the explanation by comparing the results between groups. The prominent A/B testing frequently used in industry belongs to this study type, where a large basin of users is available and different interfaces can be tested simultaneously. In contrast, within-subject experiments are used when only few users are available, especially outside the industry context. In these studies, each user is presented with all explanation interfaces (Herlocker, Konstan, and Riedl 2000; Kouki et al. 2019; Millecamp et al. 2019; Oramas et al. 2016), and one containing no explanation. Such within-subject studies need to take care of possible confounding factors emerging from the subsequent interaction with different interfaces (e. g., a user may feel lost interacting with a complex interface after seeing a very simple one).

Another fundamental aspect of user studies are the type of measurements they employ (Knijnenburg et al. 2012), usually either behavioral, such as click-through-rates and time-spent-interacting (Andjelkovic, Parra, and O'Donovan 2016; Zhao et al. 2019), or attitudinal, for instance, surveys and semi-structured interviews (Behrooz et al. 2019; Kouki et al. 2019). Generally, the measurement should be carefully tailored to the explanation goal(s). For example, if *persuasiveness* and *trustworthiness* are the most relevant explanation goals, we can assess the first via click-through-rate and the second through specific questions for example, "*Do you trust the recommendation?*" In an industrial context, these measurements may be used as key performance indicators of the explanations, though little research has been carried out here beyond general users' satisfaction (e. g., streaming time and weekly active users count).

Lastly, music consumption is influenced by the user's personal characteristics and context, which also affect the reception of the explanations. It is, therefore, necessary to take them into account by ensuring a representative population sample. Research has considered different demographics (e. g., gender, age group, and country) (Behrooz et al. 2019; Sharma and Cosley 2013), musical sophistication (Millecamp et al. 2019), listening habits (Andjelkovic, Parra, and O'Donovan 2016; Oramas et al. 2016), and psychological traits such as personality (Kouki et al. 2019) and need for cognition (Millecamp et al. 2019).

Evaluating explanation from the technical stakeholders' perspective

Methods to evaluate explanations can also serve the technical stakeholder's side of MRSs, for example, engineers and data analysts. Technical—offline—evaluations, though more convenient to conduct than user studies, are prone to the adoption of sketchy intuitive metrics, which can result in confirmation biases (Doshi-Velez and Kim 2017; Lipton 2018). Fortunately, some metrics for explainability are widely agreed upon and seldom lead to misinterpretations. For instance, the *stability* of an explanation between re-estimations (Naman Bansal 2020), its *robustness* to small data changes (Kindermans et al. 2019; Alvarez-Melis and Jaakkola 2018), and its *consistency* across several similar models (Fel and Vigouroux 2020) appear to be reasonable minimal requirements for XAI. Similarly, *sparsity* is often desirable for explanations since fewer parameters in the explanation translate to better cognitive handling

(Rudin 2019). *Discriminativeness* is already a not-so-trivial requirement as some popular feature-based explanation methods were shown to result in the same explanations across several class predictions (Adebayo et al. 2018). Other subtle sanity checks are necessary: for example, some ML models tend to leverage out-of-distribution artifacts and thus provide nonsensical explanations (Kumar et al. 2020), which must be avoided.

In a semi-encouraging manner, some XAI goals seem harder to achieve than to check. For instance, fairness objectives often stem from measured biases (e. g., disparity), the impact of which a fairness-inducing system can thus be quantified (Frye, Feige, and Rowat 2020). Note that this gets trickier for less tractable objectives (e. g., minimizing environmental impact) or if a complete measurement is unavailable, costly, or requires time to witness a significant change. The same could be said for *interactivity*, for instance by tracking the variety of tracks a user listens to after adopting the system.

Not every method can generate explanations for all items or users of a MRSs. Thus, it is useful to measure the *coverage* of a method, for example, how many explainable items are recommended in the top-k list for each user (Abdollahi and Nasraoui 2016). Likewise, computational efficiency of explanation generation should be taken into account (Chen et al. 2018), particularly for time-sensitive use cases.

EXPLAINABILITY CHALLENGES IN AN INDUSTRIAL CONTEXT

In previous sections, we discussed that different ways MRSs can be made more explainable and how to evaluate explanations. We now focus on the inherent challenges that arise in a real industrial context when trying to implement these methods to explain recommendations to end-users.

Explanations in real MRS

Many commercial music streaming services design recommendation interface as *swipeable carousels* (Bendada et al. 2020), namely sequences of sections that users can scroll. These carousels have titles that convey information to end-users such as:

- Self-explanatory titles: for example, "Top 10," "Popular in your area," "Trending content" or "Recommended for you" that merely indicate the content selection process (Figure 4 top).
- Feature-based explanations: for example, "70's soul" or "Rock music" (Figure 4 middle).
- Example-based explanations: for example, "Because you like artist X," "Because you listened to album Y" (Figure 4 bottom).

Certainly, these simple and crude explanations are in contrast with the advanced explanation capabilities we have presented earlier. Graph-based explanations, for instance, do not easily fit the headline formatting constraint, due to their length and complexity. Therefore, they are quite uncommon in the industry though they represent a promising aspect of conversational MRS. In the following, we further analyze this discrepancy between the scientific state of the art and the industrial realm.

Overview of an industrial MRS

A simplistic view of an industrial MRS is given in Figure 5. Central to it is the *Core Recommendation Engine* that models users and items affinities. Usually trained offline on a vast amount of user–item interactions, the system is then used online to generate item recommendations for each user accessing the service. This core MRS is complemented by heuristic filters and pre/post-processing.

To train and query the Core module, only a fraction of all available information about items and users will be eventually used. For instance, users' metadata such as location, context or declared age can be used as-is, transformed (e. g., quantized into broad areas or age buckets) or discarded. Items' data can be even more heavily processed. The audio signal can be subsampled, compressed, bounded, or normalized. Contextual information about the device, time, and location may be collected or inferred. Additionally, some systems leverage continuous user feedback in a session for online adaptation.

Symmetrically, the direct output of the core RS is not what the final user will be confronted to. Heuristics may be added, for instance to remove items that were already presented recently. In some contexts, enforcing contractual or legal obligations can also be necessary. Finally, product constraints in terms of display space on the device, connectivity status, or content availability issues can impact recommendations.

Issues with explainability in industrial MRS

If we were to provide a detailed description of the internals of an MRS, destined to *end-users* and using *natural*

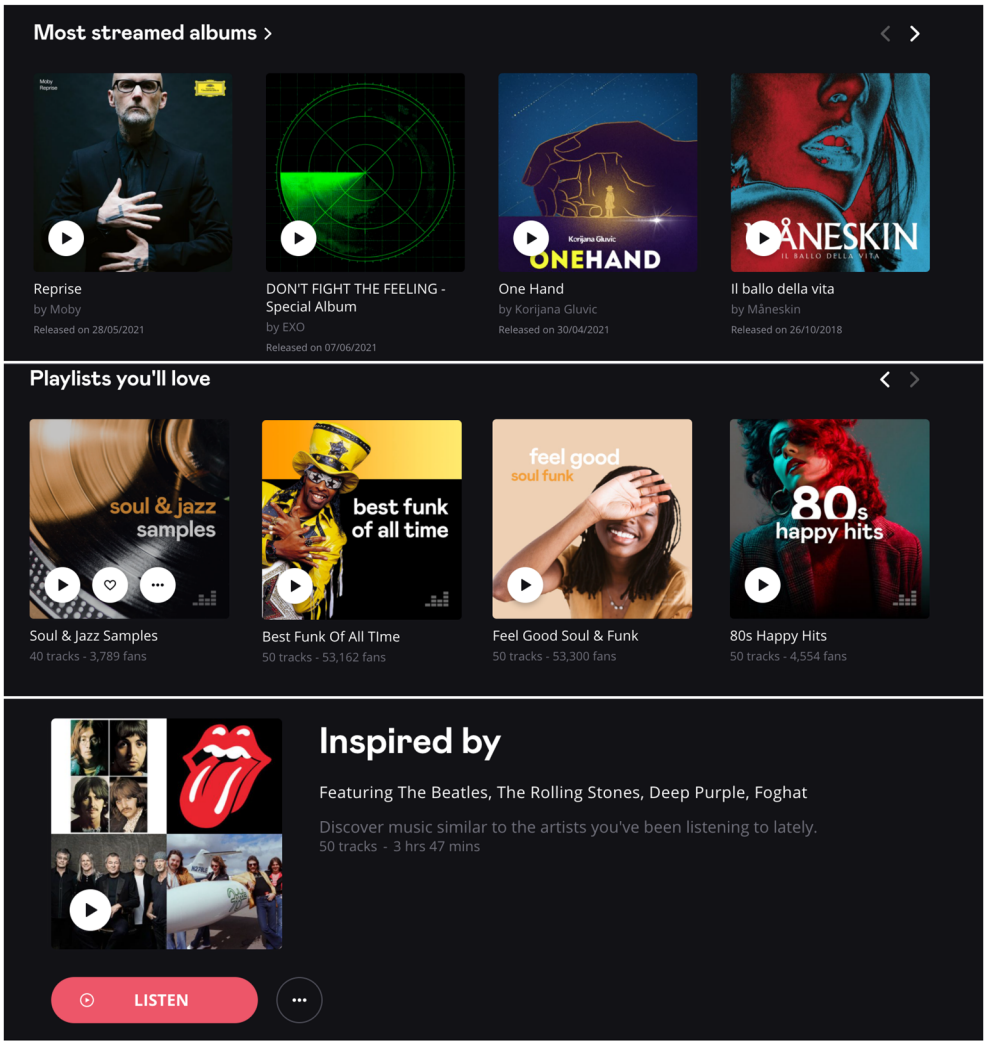

FIGURE 4 Real-world recommendations with explanations

language, it would probably look like the explanations provided in Table 1. While these may seem too detailed and almost provocative, they highlight a set of issues that we may face when trying to include explanations in an industrial MRS.

Issues with engineering assumptions and design choices

MRSs largely rely on implicit feedback and the engineering assumptions that come with their processing. For instance, most music services collect user feedbacks through basic interactions, namely *skips*, *likes*, *dislikes*, listening history and navigation outside what was provided by the MRS (such as music that was retrieved through the search engine). While *dislike* are rather self-explanatory, the intention of the user *liking* a recommended item may not be that clear as users may use it for bookmarking songs. The intention behind a *skip* is even more difficult to understand (Afchar and Hennequin 2020), while *skips* remain the most basic and common interactions. Thus, MRS designers usually want to take advantage of them and enforce heuristic rules, for example, negatively weighting *skips* and positively considering full-songs listenings (even though the music may have been played without someone actually listening to it).

Some design choices can also be made in order to make the system computationally efficient, notably limiting the amount of data: for instance, in item (1) of Table 1, the system needs to explain that old interactions were not taken into account for providing the recommendation, otherwise a user may not understand why some recurrent *skip* of an artist they dislike was not taken into account. It is worth noting that such design choices are usually optimized in the industrial context (e. g., through A/B testing), but are rarely considered in academic research.

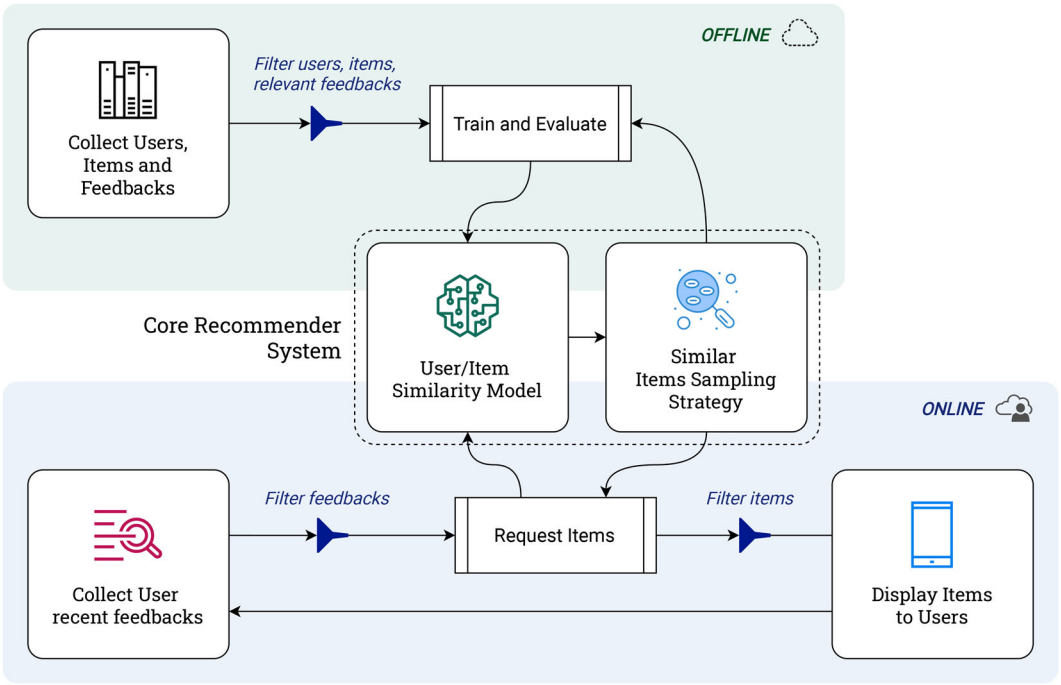

FIGURE 5 Overview of an industrial MRS. Explaining the recommendation may require more than an explainable core MRS.

TABLE 1 Honest recommendations explanation

	"We recommended the song A by artist B to you because:
(1)	*We considered your recent history (e. g., 3 months) and that older interaction may no longer be relevant. Also, considering a longer time period would have been too computationally costly.*
(2)	*We also considered the recent history of many more users (not all of them, some were excluded because, for instance they had too few interactions, or peculiar activity patterns) to learn a representation space encoding similarity between artists with a ML system.*
(3)	*The ML system learned to give close representations to artists that are co-listened by the same set of users and distant representations to artists that are not.*
(4)	*We saw that you listened to songs by artists similar to artist B and you did not skip them which we interpreted as positive feedback.*
(5)	*Eventually, you also explicitly liked artists similar to B or songs similar to A.*
(6)	*Some other songs that could have been very relevant in this context were discarded because you skipped them in a previous session.*
(7)	*We sampled items in our representation space that are close to items to which you gave positive feedback and far from those with negative feedback.*
(8)	*Song A and artist B also passed other heuristic filters (e. g., regarding redundancy of recommended content, or a user personal blacklist).*"

We could also mention that it is common, in large catalogs, to encounter metadata ambiguities such as homonym artist profiles or polysemous musical genres. The impacts of such ambiguities on the system can be large and put explanation at risk of being deceptive, for instance, if an example-based explanation "Because you listened to artist X" is displayed to a user that listened to a different artist named X.

In the artificial explanation presented in Table 1, items (4), (6), and (1) rely on pragmatical assumptions. This makes the explanation quite complex and may decrease user satisfaction, especially if some assumption is invalid: for example, when a song was skipped because played in an inappropriate context and not because it was disliked. Furthermore, providing such a detailed explanation may change the user behavior w. r. t. these interactions: for example, they may avoid skipping songs they like in order to prevent them from being discarded in future recommendations, which may bring dissatisfaction.

Trade-off between simplicity of the explanation and complexity of the RS

Among industrial actors, the *less is more* design pattern is widely adopted as a general good practice supported by the theory of cognitive load applied to user interface design (Zhou, Ji, and Jiao 2013). The latter suggests that unnecessary display of information goes against ergonomics principles (Scapin and Bastien 1997), and can thus be detrimental to users' satisfaction. Following these guidelines, end-user explanations should be carefully crafted to remain simple and concise, hence making cognitive overload less likely.

Additionally, industrial incentives are primarily driven toward highly accurate systems. This often requires complex MRS's components, making explanations not simple enough to be provided to the user. For instance, some constituting blocks can be based on black-box methods, as latent factor-based models or deep embeddings that are widely used in MRSs. These models embed users and items as multidimensional vectors and represent affinity as their relative distance in this space. While they usually provide good results in a large variety of recommendation tasks, the latent factors are very difficult to understand: for instance, in item (2) of Table 1, artist similarity is computed by a black-box system, which is barely explainable to the user.

Besides, several MRS processing blocks rely on parameter choices. For instance, one may not consider all past user–song interactions to train the user–item affinity model, but only those that are significant, (e. g., interactions when user listened to at least half of a song). But this threshold is arbitrary and may exclude interactions that are important to a user: for example, users only listening to the intro of a song many times because they like it a lot. Arguably, these parameters should be optimized, but in practice, they are so numerous that optimization becomes intractable.

Finally, industrial MRS are built upon several subblocks that are glued together and that rely on various sources of data: user modeling, content modeling, user–item affinity modeling, and so forth. The recommendation is made on top of all those blocks that may each influence the final recommendation. The impact of each block on this final recommendation is quite difficult to assess and, consequently, it is hard to generate a simple explanation on top of these unclear impacts. Feature selection may appear as a solution, but as long as several features are significantly impacting the prediction, the explanation would need to be either complex or incomplete. The overall complexity of explanations in Table 1 illustrates this issue.

Issues of transparency with respect to company competition

One of the main goals of explanations for RS is to increase transparency. While transparency can increase user satisfaction, it can possibly disclose some critical aspects of the system. Then, making sure that explanations do not bring insights about the system internals can be necessary. For instance, releasing the information that the MRS uses artist embeddings (item (3) of Table 1) or a specific hyperparameter of the system such as the considered timeframe in the history (item (1)) can be sensitive information that a private company may be reluctant to make public to competitors.

Perspectives for explainable MRSs

Improving the level of explanation of MRSs while keeping strong simplicity constraints for the user remains a challenge. Though, the end-user is not the only stakeholder to be impacted by MRSs. For instance, the revenue of music producers is impacted, too. Global explanations may thus be relevant for such an audience, in terms of fairness and transparency (explanations would not be about single recommendations but probably about explaining why an artist was recommended to a particular group of people). As there are no simplicity constraints for this kind of stakeholder, explanations could possibly be much more elaborated.

Another aspect is that keeping a system explainable is important for constantly improving its performance. For instance, receiving user complaints or feedback about bad recommendations can only be leveraged for improving the system if the RS engineers can understand the reason for these mis-recommendations. A RS that relies on black-box blocks prevents understanding bad recommendations and, therefore, hinders improving the system.

Finally, advanced users may want more control and simplicity constraints may be less important to them: for instance, Jin, Cardoso, and Verbert (2017) argues that, as opposed to the *less is more* design pattern, giving users additional control over the RS does increase cognitive load, but also increases user satisfaction for users who have a deep understanding of how it works. Controls on the RS make possible a positive feedback loop: explanations can be explicitly leveraged by the user to act on the RS and mitigate future spurious recommendations.

Interestingly, the increasing usage of voice-controlled devices to pilot music streaming services creates a promising new playground for deploying explainable MRSs and beyond, to create fully interactive experiences

where recommendations can be challenged, and eventually improved.

ACKNOWLEDGMENTS

This work received financial support by the Austrian Science Fund (FWF): P33526 and DFH-23; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through Grant LIT-2020-9-SEE-113.

CONFLICT OF INTEREST

The authors declare that there is no conflict.

ORCID

Darius Afchar https://orcid.org/0000-0002-4315-1461
Alessandro B. Melchiorre https://orcid.org/0000-0003-1643-1166
Markus Schedl https://orcid.org/0000-0003-1706-3406
Romain Hennequin https://orcid.org/0000-0001-8158-5562
Elena V. Epure https://orcid.org/0000-0002-6930-9482
Manuel Moussallam https://orcid.org/0000-0003-0886-5423

ENDNOTES

[1]The catalogs of music streaming platforms such as Deezer, Spotify, or Pandora include several tens of million music pieces.

[2]The existence of a general interpretability/accuracy trade-off seems a myth (Lipton 2018; Rudin 2019), despite its persistent mentions in some XAI papers.

[3]Math and IA seminar: https://vimeo.com/522733917

REFERENCES

Abdollahi, B., and O. Nasraoui. 2016. "Explainable Matrix Factorization for Collaborative Filtering." In *Proceedings of the 25th International Conference Companion on World Wide Web*, 5–6.

Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. 2018. "Sanity Checks for Saliency Maps." In *Proceedings of the Advances in Neural Information Processing Systems* 31, 9505–15.

Afchar, D., and R. Hennequin. 2020. "Making Neural Networks Interpretable with Attribution: Application to Implicit Signals Prediction." In *Proceedings of the Fourteenth ACM Conference on Recommender Systems*, 220–9. New York: ACM.

Aljanaki, A., and M. Soleymani. 2018. "A Data-driven Approach to Mid-level Perceptual Musical Feature Modeling." In *Proceedings of the International Society for Music Information Retrieval Conference*.

Alvarez-Melis, D., and T. S. Jaakkola. 2018. "On the Robustness of Interpretability Methods." In *ICML Workshop on Human Interpretability in Machine Learning*.

Andjelkovic, I., D. Parra, and J. O'Donovan. 2016. "Moodplay: Interactive Mood-based Music Discovery and Recommendation." In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 275–9. New York: ACM.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115.

Balog, K., and F. Radlinski. 2020. "Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 329–38 New York: ACM.

Behrooz, M., S. Mennicken, J. Thom-Santelli, R. Kumar, and H. Cramer. 2019. "Augmenting Music Listening Experiences on Voice Assistants." In *Proceedings of the International Society for Music Information Retrieval Conference*.

Bendada, W., G. Salha, and T. Bontempelli. 2020. "Carousel Personalization in Music Streaming Apps with Contextual Bandits." In *Proceedings of the Fourteenth ACM Conference on Recommender Systems*, 420–5. New York: ACM.

Bonnin, G., and D. Jannach. 2014. "Automated Generation of Music Playlists: Survey and Experiments." *ACM Computing Surveys* 47(2): 26:1–26:35.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. "Translating Embeddings for Modeling Multi-relational Data." In *Proceedings of the Advances in Neural Information Processing Systems*, 1–9.

Castells, P., S. Vargas, and J. Wang. 2011. "Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance." In *Proceedings of the International Workshop on Diversity in Document Retrieval*, 881–918.

Celma, Ò., P. Herrera, and X. Serra. 2006. "Bridging the Music Semantic Gap." In *Proceedings of the Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, Volume 187, Budva, Montenegro: CEUR.

Chen, H., J. D. Janizek, S. Lundberg, and S.-I. Lee. 2020. "True to the model or true to the data?" In *ICML Workshop on Human Interpretability in Machine Learning*.

Chen, J., L. Song, M. Wainwright, and M. Jordan. 2018. "Learning to Explain: An Information-theoretic Perspective on Model Interpretation." In *Proceedings of the International Conference on Machine Learning*, 883–92 PMLR.

Chen, X., H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha. 2019. "Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 765–74 New York: ACM.

Choi, J., A. Khlif, and E. Epure. 2020. "Prediction of User Listening Contexts forMmusic Playlists." In *Proceedings of the 1st Workshop on NLP for Music and Audio*, 23–7 Association for Computational Linguistics.

Costa, F., S. Ouyang, P. Dolog, and A. Lawlor. 2018. "Automatic Generation of Natural Language Explanations." In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, New York: ACM.

Covert, I., S. Lundberg, and S.-I. Lee. 2020. "Feature Removal is a Unifying Principle for Model Explanation Methods." In *Proceedings of the NeurIPS 2020 ML-Retrospectives, Surveys & Meta-Analyses Workshop*.

Dhariwal, P., H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. 2020. "Jukebox: A Generative Model for Music." *arXiv preprint arXiv:2005.00341*.

Dinu, J., J. Bigham, and J. Z. Kolter. 2020. "Challenging Common Interpretability Assumptions in Feature Attribution Explanations." In *Proceedings of the NeurIPS 2020 ML-Retrospectives, Surveys & Meta-Analyses Workshop*.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. 2014. "Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 601–10.

Doshi-Velez, F., and B. Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv preprint arXiv:1702.08608*.

Ehrlich, K., S. E. Kirk, J. Patterson, J. C. Rasmussen, S. I. Ross, and D. M. Gruen. 2011. "Taking Advice from Intelligent Systems: The Double-edged Sword of Explanations." In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 125–34.

Fel, T., and D. Vigouroux. 2020. "Representativity and Consistency Measures for Deep Neural Network Explanations." *arXiv preprint arXiv:2009.04521*.

Ferwerda, B., M. Schedl, and M. Tkalcic. 2015. "Personality & Emotional States: Understanding Users' Music Listening Needs." In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization*, Volume 1388 CEUR-WS.

Frye, C., I. Feige, and C. Rowat. 2020. "Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-agnostic Explainability." In *Proceedings of the 34th Conference on Neural Information Processing Systems*.

Ge, M., C. Delgado-Battenfeld, and D. Jannach. 2010. "Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity." In *Proceedings of the Fourth ACM Conference on Recommender Systems*, 257–60.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Networks." In *Proceedings of the Advances in Neural Information Processing Systems*.

Goto, M., and R. B. Dannenberg. 2019. "Music Interfaces based on Automatic Music Signal Analysis: New Ways to Create and Listen to Music." *IEEE Signal Processing Magazine* 36(1): 74–81.

Grover, A., and J. Leskovec. 2016. "node2vec: Scalable Feature Learning for Networks." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–64.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51(5): 1–42.

He, X., L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. 2017. "Neural Collaborative Filtering." In *Proceedings of the 26th International Conference on World Wide Web*, 173–82.

Hendricks, L. A., Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. 2016. "Generating Visual Explanations." In *Proceedings of the European Conference on Computer Vision*, 3–19 Springer.

Herlocker, J. L., J. A. Konstan, and J. Riedl. 2000. "Explaining Collaborative Filtering Recommendations." In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241–50.

Jannach, D., L. Lerche, and I. Kamehkhosh. 2015. "Beyond"Hitting the Hits": Generating Coherent Music Playlist Continuations with the Right Tracks." In *Proceedings of the 9th ACM Conference on Recommender Systems*, 187–94 ACM.

Ji, S., S. Pan, E. Cambria, P. Marttinen, and P. S. Yu. 2021. "A Survey on Knowledge Graphs: Representation, Acquisition and Applications." *IEEE Transactions on Neural Networks and Learning Systems* 33, 494–514.

Jin, Y., B. Cardoso, and K. Verbert. 2017. "How do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations?." In *Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, Volume 1884, 35–42 CEUR.

Kang, W.-C., and J. McAuley. 2018. "Self-attentive Sequential Recommendation." In *Proceedings of the 2018 IEEE International Conference on Data Mining*, 197–206 IEEE.

Kaur, H., H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. 2020. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning." In *Proceedings of the 2020 Conference on Human Factors in Computing Systems*, 1–14.

Kim, B., M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres. 2018. "Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." In *Proceedings of the International Conference on Machine Learning*, 2668–77 PMLR.

Kim, J., A. M. Demetriou, S. Manolios, M. S. Tavella, and C. C. Liem. 2020. "Butter Lyrics over Hominy Grit": Comparing Audio and Psychology-based Text Features in MIR Tasks." In *Proceedings of the International Society for Music Information Retrieval Conference*.

Kindermans, P.-J., S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. 2019. "The (un) Reliability of Saliency Methods." In *NIPS workshop on Interpreting, Explaining and Visualizing Deep Learning*.

Knees, P., and M. Schedl. 2016. *Music Similarity and Retrieval: An Introduction to Audio- and Web-based strategies*, Volume 36. Springer.

Knees, P., M. Schedl, and M. Goto. 2020. "Intelligent User Interfaces for Music Discovery." *Transactions of the International Society for Music Information Retrieval* 3(1): 165–79.

Knijnenburg, B. P., M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. 2012. "Explaining the User Experience of Recommender Systems." *User Modeling and User-Adapted Interaction* 22(4): 441–504.

Kouki, P., J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. 2019. "Personalized Explanations for Hybrid Recommender Systems." In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 379–90. New York: ACM.

Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, and S. Friedler. 2020. "Problems with Shapley-value-based Explanations as Feature Importance Measures." In *Proceedings of the International Conference on Machine Learning*, 5491–500 PMLR.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva. 2017. "Counterfactual Fairness." In *Proceedings of the Advances in Neural Information Processing Systems*, Volume 30 Curran Associates, Inc.

Lee, J., N. J. Bryan, J. Salamon, Z. Jin, and J. Nam. 2020. "Disentangled Multidimensional Metric Learning for Music Similarity." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6–10 IEEE.

C Lipton, Z. 2018. "The Mythos of Model Interpretability." *Queue* 16(3): 31–57.

Lundberg, S., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the Advances in Neural Information Processing Systems*.

Melchiorre, A., V. Haunschmid, M. Schedl, and G. Widmer. 2021. "Lemons: Listenable Explanations for Music Recommender Systems." In *Proceedings of the 43rd European Conference on Information Retrieval*.

Melchiorre, A. B., and M. Schedl. 2020. "Personality Correlates of Music Audio Preferences for Modelling Music Listeners." In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 313–7.

Millecamp, M., N. N. Htun, C. Conati, and K. Verbert. 2019. "To Explain or Not to Explain: The Effects of Personal Characteristics when Explaining Music Recommendations." In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 397–407.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38.

Mirza, M., and S. Osindero. 2014. "Conditional Generative Adversarial Nets." *arXiv preprint arXiv:1411.1784*.

Müller, M. 2015. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Switzerland: Springer.

Naman, Bansal, C. Agarwal, and A. Nguyen. 2020. "Sam: The Sensitivity of Attribution Methods to Hyperparameters." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

K Natarajan, B. 1995. "Sparse Approximate Solutions to Linear Systems." *SIAM Journal on Computing* 24(2): 227–34.

Nunes, I., and D. Jannach. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems." *User Modeling and User Adapted Interaction* 27(3-5): 393–444.

Oramas, S., L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra. 2016. "Information Extraction for Knowledge Base Construction in the Music Domain." *Data & Knowledge Engineering* 106: 70–83.

Parliament, E. U. 2016. "Regulation (eu) 2016/679."

Poursabzi-Sangdeh, F., D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. 2021. "Manipulating and Measuring Model Interpretability." In *Proceedings of the Conference on Human Factors in Computing Systems*.

Rentfrow, P., L. R. Goldberg, and R. Zilca. 2011. "Listening, Watching, and Reading: The Structure and Correlates of Entertainment Preferences." *Journal of Personality* 79: 223–58.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?" Explaining the Predictions of Any Cassifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44.

Ricci, F., L. Rokach, and B. Shapira. 2015. *Recommender Systems Handbook*. New York: Springer.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5): 206–15.

Salha, G., S. Limnios, R. Hennequin, V.-A. Tran, and M. Vazirgiannis. 2019. "Gravity-inspired Graph Autoencoders for Directed Link Prediction." In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 589–98.

Scapin, D. L., and J. C. Bastien. 1997. "Ergonomic Criteria for Evaluating the Ergonomic Quality of Interactive Systems." *Behaviour & Information Technology* 16(4-5): 220–31.

Schedl, M., P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas. 2015. "Music Recommender Systems." In *Recommender Systems Handbook*, 453–92. Boston: Springer.

Schedl, M., H. Zamani, C. Chen, Y. Deldjoo, and M. Elahi. 2018. "Current Challenges and Visions in Music Recommender Systems Research." *International Journal of Multimedia Information Retrieval* 7(2): 95–116.

Serrano, S., and N. A. Smith. 2019. "Is Attention Interpretable?." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–51.

Sharma, A., and D. Cosley. 2013. "Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems." In *Proceedings of the 22nd International Conference on World Wide Web*, 1133–44. New York: ACM.

Simonyan, K., A. Vedaldi, and A. Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." In *Proceedings of the Workshop, ICLR*.

Sundararajan, M., and A. Najmi. 2020. "The Many Shapley Values for Model Explanation." In *Proceedings of the International Conference on Machine Learning*, Volume 119, 9269–78.

Tintarev, N., and J. Masthoff. 2015. "Explaining Recommendations: Design and Evaluation." In *Recommender Systems Handbook*, 353–82. Boston: Springer.

Ustun, B., A. Spangher, and Y. Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–9. New York: ACM.

Wang, N., H. Wang, Y. Jia, and Y. Yin. 2018. "Explainable Recommendation via Multi-task Learning in Opinionated Text Data." In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 165–74 New York: ACM.

Wang, X., D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua. 2019. "Explainable Reasoning over Knowledge Graphs for Recommendation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, 5329–36.

Ying, R., R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. 2018. "Graph Convolutional Neural Networks for Web-scale Recommender Systems." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 974–83.

Yu, L., and H. Liu. 2003. "Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution." In *Proceedings of the 20th International Conference on Machine Learning*, 856–63.

Zamani, H., M. Schedl, P. Lamere, and C. Chen. 2019. "An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation." *ACM Transactions on Intelligent Systems and Technology* 10(5): 57:1–57:21.

Zhang, Y., and X. Chen. 2020. "Explainable Recommendation: A Survey and New Perspectives." *Foundations and Trends in Information Retrieval* 14(1): 1–101.

Zhang, Y., H. Zhang, M. Zhang, Y. Liu, and S. Ma. 2014. "Do Users Rate or Review? Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification." In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1027–30. New York: ACM.

Zhao, G., H. Fu, R. Song, T. Sakai, Z. Chen, X. Xie, and X. Qian. 2019. "Personalized Reason Generation for Explainable Song

Recommendation." *ACM Transactions on Intelligent Systems and Technology* 10(4): 1–21.
Zhou, F., Y. Ji, and R. J. Jiao. 2013. "Affective and Cognitive Design for Mass Personalization: Status and Prospect." *Journal of Intelligent Manufacturing* 24(5): 1047–69.

AUTHOR BIOGRAPHIES

Darius Afchar is a Ph.D. student working at Deezer Research and in the MLIA team at Sorbonne-Université. His research topics include interpretability and music recommender systems. His published work specifically focuses on feature attribution both from a theoretical and practical standpoint.

Alessandro B. Melchiorre is a Ph.D. student at the Institute of Computational Perception at the Johannes Kepler University Linz (JKU) and the Linz Institute of Technology (LIT). His research interests revolve around the topics of recommender system algorithms, explainability in AI, and algorithmic bias and fairness, especially in the music domain.

Markus Schedl is a full Professor at the Johannes Kepler University Linz (JKU), affiliated with the Institute of Computational Perception and the Linz Institute of Technology (LIT), where he leads the Multimedia Mining and Search and the Human-centered AI groups, respectively. He graduated in Computer Science from the Vienna University of Technology and earned his Ph.D. in Computer Science from JKU. His main research interests include recommender systems, user modeling, web and social media mining, as well as text, music, and multimedia information retrieval, research fields in which he (co-)authored more than 240 refereed articles in journals and conference proceedings.

Romain Hennequin is a Research Scientist at Deezer where he heads the researchers team. He graduated in Computer Science from Ecole Polytechnique, UPMC (now Sorbonne Université), and Telecom Paris and earned a Ph.D. in signal processing from Telecom Paris. He has been working for more than 10 years in industrial research, addressing various topics such as source separation, music information retrieval, recommender systems, and graph mining.

Elena V. Epure is a Research Scientist specialized in Natural Language Processing at Deezer, a global music, podcasts, and audiobooks streaming service located in Paris. Her research interests span a broad range of topics in NLP: information extraction and retrieval; semantics especially cross-cultural; pragmatics and conversation modeling; and generation, particularly for music captioning. She is also interested in topics related to Recommender Systems and User Modeling.

Manuel Moussallam is Director of research at Deezer. With a background in music information retrieval and signal processing, he has been working on topics ranging from music catalog organization to recommender systems and music consumer modeling. He has a Ph.D. in Applied Mathematics and a MS in Computer Science.

How to cite this article: Afchar, D., A. B. Melchiorre, M. Schedl, R. Hennequin, E. V. Epure, and M. Moussallam. 2022. "Explainability in music recommender systems." *AI Magazine* 43: 190–208. https://doi.org/10.1002/aaai.12056