12

Musical Robots: Overview and Methods for Evaluation

Emma Frid

Musical robots are complex systems that require the integration of several different functions to operate successfully. These processes range from sound analysis and music representation to mapping and modeling of musical expression. Recent advancements in Computational Creativity (CC) and Artificial Intelligence (AI) have added yet another level of complexity to these settings, with aspects of Human–AI Interaction (HAI) becoming increasingly important. The rise of intelligent music systems raises questions not only about the evaluation of Human–Robot Interaction (HRI) in robot musicianship but also about the quality of the generated musical output. The topic of evaluation has been extensively discussed and debated in the fields of Human–Computer Interaction (HCI) and New Interfaces for Musical Expression (NIME) throughout the years. However, interactions with robots often have a strong social or emotional component, and the experience of interacting with a robot is therefore somewhat different from that of interacting with other technologies. Since musical robots produce creative output, topics such as creative agency and what is meant by the term "success" when interacting with an intelligent music system should also be considered. The evaluation of musical robots thus expands beyond traditional evaluation concepts such as usability and user experience. To explore which evaluation methodologies that might be

DOI: 10.1201/9781003320470-12

appropriate for musical robots, this chapter first presents a brief introduction to the field of research dedicated to robotic musicianship, followed by an overview of evaluation methods used in the neighboring research fields of HCI, HRI, HAI, NIME, and CC. The chapter concludes with a review of evaluation methods used in robot musicianship literature and a discussion of prospects for future research.

12.1 Background

The history of musical automata predates digital technology. Archimedes invented the first known humanoid musical automaton, an elaborate clepsydra (water clock) combined with a Byzantine whistle, in the 3rd century BC and attempts to mechanize musical instruments in the form of mechanically wind-fed organs were done as early as in the 4th century BC [43, 88]. Mechanical automatic musical instruments that play pre-programmed music with negligible human intervention can be traced back at least to the 9th century [55, 87]. The algorithmic thought in Western music composition goes even further back in time, to the beginning of notation [129]. In more recent years, advances in computational power, sound processing, electrical engineering, as well as Artificial Intelligence (AI) and Virtual/Augmented Reality (VR/AR), have paved the way for new interaction possibilities with robots that go beyond physical corporality. Today, technological developments have blurred the line between robots as tangible entities and robots as abstract intelligent agents. The emergence of such musical systems introduces a need to understand and evaluate robotic systems in the musical and socio-cultural context in which they are used. But how should these systems be evaluated, and which properties should be considered important, when pursuing such an activity? This is the focus of this chapter.

Before diving deeper into a discussion of evaluation methods, it is first important to define the term *"musical robot"*; which properties of a system are required to be considered a musical robot and – more importantly – what is *not* a musical robot? To be able to answer these questions, we may refer to standards and robot taxonomies. It has been suggested that the concept of a *"robot"* predates the word by several centuries, and that the history of robots has been intertwined with the arts [149]. In ISO Standard 8373:2012, a robot is defined as *"a programmed actuated mechanism with a degree of autonomy to perform locomotion, manipulation or positioning"* [66]. Autonomy in this context refers to the *"ability to perform intended tasks based on current state and sensing, without human intervention"*.

Given that music, or musicking [140], is an activity that is embedded in a social context, it is worth reviewing taxonomies from the field of Human–Robot Interaction (HRI) in this context. Several different taxonomies have been

TABLE 12.1
Overview of Onnasch and Roesler's taxonomy to structure and analyze Human–Robot Interaction (adapted from [115]). Abbreviations: a = anthropomorphic, z = zoomorphic, t = technical, N_H = number of humans, N_R = number of robots.

Interaction context		**Field of application**	**Exposure to**
		Industry, service, military & police, space expedition, therapy, education, entertainment, none	Robot: embodied, depicted Setting: field, laboratory
Robot	**Robot task specification**	**Robot morphology**	**Degree of robot autonomy**
	Information exchange, precision, physical load reduction, transport, manipulation, cognitive stimulation, emotional stimulation, physical stimulation	Appearance: a/z/t Communication: a/z/t Movement: a/z/t Context: a/z/t	Information acquisition: − to + Information analyses: − to + Decision making: − to + Action implementation: − to +
Team	**Human role**	**Communication channel**	**Proximity**
	Supervisor, operator, collaborator, cooperator, bystander	Input: electronic, mechanical, acoustic, optic Output: tactile, acoustic, visual	Temporal: synchronous, asynchronous Physical: following, touching, approaching, passing, avoidance, none
	Team composition $N_H = N_R$ $N_H > N_R$ $N_H < N_R$		

proposed (see e.g. [169,170]). A recent example is the taxonomy introduced by Onnasch and Roesler in [115], which divides HRI work into three clusters with different foci: (1) *interaction context* classification, (2) *robot* classification, and (3) *team* classification. An overview of the three different clusters, and their corresponding categories to specify an HRI scenario, is presented in Table 12.1. The *interaction context* cluster involves, for example, the field of application. For musical robots, relevant examples include entertainment, education, and therapy. The interaction context also relates to how you are exposed to the robot; exposure can be embodied, which is the case for a physical robot, or depicted, which is the case for a virtual agent. This exposure can be in a field versus laboratory setting. The *robot classification* cluster focuses on the robot's work context and design; robot task specification, robot morphology, and degree of robot autonomy. In this context, robot morphology refers to the appearance of the robot, among other factors. For example, a robot can be classified as anthropomorphic (human-like) or zoomorphic (animal-like). It could also be more task-driven than human, i.e. technical. Finally, the third cluster, *team classification*, focuses on the human role (supervisor, operator, collaborator, co-operator, or bystander), the team composition (number of robots versus humans), the communication channel (e.g. tactile or acoustic communication), and proximity (temporal or physical).

A framework focused on classification of *social robots* was presented in [8]. This classification characterizes robots along seven dimensions (somewhat overlapping with the categories discussed in [115]): *appearance, social capabilities, propose and application area, relational role, autonomy and intelligence, proximity,* and *temporal profile*. Although musical robots may find themselves on different points along these dimensions, some broader themes can be identified. For example, musical robots often have artifact-shaped or bio-inspired *appearance*. In other words, the design of musical robots is often inspired by acoustic instruments or features of, or even the entire, human body.[1] Different robots have different levels of *social capabilities*. For example, musical robots usually communicate using non-verbal modalities, producing sounds. They may also use motion, gestures, and lights. Some musical robots can model and recognize social aspects of human communication and respond accordingly. For example, they may interpret musical phrases played by a musician and adopt their musical response. The *purpose and application areas* of musical robots span across a wide range of different domains. Musical robots can be used for personal empowerment, to expand on human abilities, and to empower people to enhance creativity on an individual level. The *relational role* of the robot, i.e. the role that the robot is designed to fulfill within an interaction, can also take many forms. For example, musical robots can act as co-players in an ensemble, solo performers, and music teachers, among other roles. They may greatly vary when it comes to their *autonomy and intelligence*, for example, in terms of their ability to perceive environment-related and human-related factors such as physical parameters (speed, motion), non-verbal social cues (gestures, gaze, facial expression), and speech. They may also differ in terms of planning of actions and how much they can learn through interaction with humans and the environment, over time. When it comes to the spatial *proximity* of the interaction, the most common scenario for a musical robot is that the robot exists in a shared space, interacting directly with a human (but there are, of course, also other scenarios). Finally, the *temporal profile* of musical robots can vary when it comes to time span, duration, and frequency of the interactions.

To further narrow down what we mean by the term *musical robot*, we may turn to literature on *machine musicianship*, and *robot musicianship*, in particular. An influential book in this context is *"Machine Musicianship"* by Robert Rowe [129]. Rowe describes that the training of a musician begins by teaching basic musical concepts that underlie the musical skills of listening, performance, and composition. Such knowledge is commonly referred to as *musicianship*. Computer programs that are designed to implement any of these skills, for example, the skill to make sense of the music that is being heard, will benefit from a musician's level of musicianship. Another influential example from the literature is the work *"Robotic Musicianship – Embodied*

[1] Consider, for example, *"The Prayer"*, a singing mouth robot by Diemut Strebe [151], versus full-body humanoids such as the "Waseda Saxophonist Robot No. 2 (WAS2)" [148].

Artificial Creativity and Mechatronic Musical Expression" by Gil Weinberg et al. [163]. In this book, the authors describe robot musicianship research as work focused on *"the construction of machines that can produce sound, analyze and generate musical input, and interact with humans in a musically meaningful manner"* [162]. They define two primary research areas in this field: *Musical Mechatronics* [81] and *Machine Musicianship* [129]. The first relates to the study and construction of physical devices that generate sound through mechanical means, whereas the latter refers to the development of algorithms and cognitive models of music perception, composition, performance, and theory. The two disciplines are said to be brought together by *Robotic Musicianship*. Weinberg et al. describe that the ultimate goal of robotic musicianship is to design robots that can demonstrate musicality, expression, and artistry, while stimulating innovation and creativity in other musicians [162]. Rather than imitating or replacing human creativity, the goal of robotic musicianship is to supplement human creativity, and to enrich musical experiences for humans [162]. In this way, robot musicianship may advance music as an art form by creating novel musical experiences that can encourage humans to create, perform, and think of music in new ways.

Robot musicianship brings together perceptual and generative computation with physical sound generators to create systems capable of (1) *rich acoustic sound production*, (2) *intuitive physics-based visual cues from sound producing movements*, and (3) *expressive physical behaviors through sound accompanying body movements* [133]. Robotic musicians make use of various methods for music generation. This includes generative functions such as composition, improvisation, score interpretation and accompaniment, which in turn can rely on statistical models, predefined rules, abstract algorithms, or actuation techniques [162]. Going beyond sound-producing ability, an important aspect of robotic musicianship is the cognitive models and algorithms that enable the machines to act like skilled musicians. A robotic musician should have the ability to extract information relevant to the music or performance and be able to apply this information to the musical decision process. This is something that Weinberg et al. refer to as *Musical Intelligence*. As stressed by Ajay Kapur in [81], a robot must be able to sense what the human is doing musically, and the machine must deduce meaningful information from all its sensor data and then generate a valid response. Moreover, as described in [145], an idealized musical robot should integrate musical representation, techniques, expression, detailed analysis and control, for both playing and listening. Musical robots usually put emphasis on *Machine Listening*, i.e, on extracting meaningful information from audio signals using sensing and analysis methods [131]. To provide the robot with information about other musicians (for example, to be able to synchronize musical events) visual sensing and computer vision techniques, as well as multimodal analysis focused on inertia measuring units capturing acceleration and orientation of limbs, may also be used [162].

12.2 Musical Robots

The terms *"musical robots"* and *"robotic musical instruments"* can refer to a wide range of different types of musical machines [84]. Ajay Kapur [81] defines a robotic musical instrument as *"a sound-making device that automatically creates music with the use of mechanical parts, such as motors, solenoids and gears"*. Steven Kemper [84] suggested that although approaches lacking autonomy could more accurately fall under the term *"musical mechatronics"* (see e.g. [81, 171]), the popular conception of robots rooted in mythology includes any machines that can mimic human actions (citing [72, 153]). As such, he considers any approach in which an electromechanical actuator produces a visible physical action that models the human act of making music as *"musical robotics"*, regardless of level of autonomous control.

Several overviews of the history of musical robots have been published throughout the years. A review of musical automata from classical antiquity to the 19th century was provided by Krzyzaniak in [88]. Kapur published a comprehensive overview of piano robots, percussion robots, string robots, wind robots, and turntable robots in [81]. Weinberg et al. [162] provided an overview of musical robots designed to play traditional instruments, with examples of robots playing percussive instruments, stringed instruments, and wind instruments. An introduction to research trends for musical performance robots was given by Solis and Takanishi in [146]. Finally, foundations of musical robotics and how such systems experienced a rebirth even in the face of loudspeaker technology dominance, thanks to their ability to serve as uniquely spatialized musical agents, was discussed in [97, 107].

Some of the earliest examples of musical robots include different types of musical automatons and automatic musical instruments driven by water and air, such as water clocks, wind-fed organs, and systems involving whistles or flutes, including mechanical birds (see [88, 109]). When it comes to programmable music machines, notable examples include Ismail ibn al-Razzaz al-Jazari's mechanical boat with four musical automata, as well as his early examples of percussion robots. Two programmable humanoid music robots that often reoccur in the literature are Jacques de Vaucanson's *"Flute Player"* automaton from 1738 [34], and Pierre Jaquet-Droz's *"Musical Lady"* from the 1770s [149, 168].

Kapur describes the *"Player Piano"* as one of the first examples of mechanically played musical instruments. On the topic of piano robots, he mentions the *"Pianista"* piano by Henri Fourneaux and *"Pianola"* by Edwin Scott Votey. Kapur also discusses humanoid techniques in which the entire human body is modeled to play the piano, for example, the *"WABOT-2"* from Waseda University [125]. Today, there are electronic systems for control of mechanical

pianos; automated pianos controlled by MIDI data can be purchased for example from Yamaha (Disklavier) and QRS Music (Pianomathon). When it comes to percussion robots, Kapur categorizes them into three subcategories: membranophones, idiophones, and extensions[2]. An example of a membranophone robot is *"Cog"*, which can hit a drum with a stick [167]. Idiophone examples include Gerhard Trimpin's robotic idiophones [157] and the LEMUR[3] *"TibetBot"* by Eric Singer, which plays on three Tibetan singing bowls using six arms [138]. The extension category includes, for example, combinations of many instruments, e.g. the LEMUR *"ModBots"*,[4] which are modular robots that can be attached anywhere. Kapur divides string robots into subcategories based on if they are plucked verus bowed. Examples from the plucked category include *"Aglaopheme"* by Nicolas Anatol Baginsky [7], the electric guitar robot from Sergi Jordà's *"Afasia"* project (see [73]), and *"GTRBOT666"* from Captured By Robots [19]. Two bowed robot examples are the *"MUBOT"* by Makoto Kajitani [80] and Jordà's *"Afasia"* violin robot [73]. A more recent example that would fall under this category is Fredrik Gran's cello robot [52]. Kapur defines wind robots as mechanical devices performing wind instruments like brass, woodwinds, and horn-type instruments. Examples mentioned include *"MUBOT"* [79] which performs on the clarinet, the Waseda University anthropomorphic robot playing the flute [142,143] and robotic bagpipes such as *"McBlare"* by Roger Dannenberg [30]. Humanoid woodwind robots, and challenges in designing such systems, are discussed more in detail in [147].

A classification framework based on the ways in which musical robots express creativity was introduced by Kemper in [84]. The framework distinguishes between six musical robot categories, see Table 12.2. Category 1 generally prioritizes versatile, humanoid robots engaging in quintessentially "human" activities over novel musical output. One example is the *"Toyota Partner Robot"* which can play trumpet, violin, and an electronic drum kit [154]. Category 2 is different from 1 in the sense that these robots model human actions, for example, by replicating humanoid organs such as lips or oral cavity, which in turn may affect the efficiency. They often involve complex mechanical models that can limit sonic possibilities (e.g. the ability to play at super virtuosic speed). Examples include pioneering work from Waseda University, such as piano keyboard - [125], flute - [142,143], and saxophone robots [148], as well as a robotic finger for harp plucking [20]. Robots in category 3 assume an anthropomorphic form but do not model the specific actions of human performance. They are generally focused more on musical output and appearance, with an anthropomorphic nature highlighted by their look, rather than an attempt to model the human actions of performance. Examples include robotic bands

[2]These are percussion robots that do not fall into the other two categories.

[3]LEMUR stands for "League of Electronic Musical Urban Robots". It is a group of artists and technologists who create robotic musical instruments, founded by Eric Singer (https://lemurbots.org/index.html).

[4]The *"ModBots"* were, for example, used in the multi-armed percussion Indian God-like robot *"ShivaBot"* [139].

TABLE 12.2
Summary of Kemper's musical robot classification system [84].

No	Category	Examples
1	Nonspecialized anthropomorphic robots that can play musical instruments	[154]
2	Specialized anthropomorphic robots that model the physical actions of human musicians	[20, 125, 142, 143, 148]
3	Semi-anthropomorphic robotic musicians	[33, 130, 141, 165]
4	Non-anthropomorphic robotic instruments	[128, 138]
5	Cooperative musical robots	[55, 163]
6	Individual actuators used for their own sound production capabilities	[117, 175]

such as *"The Trons"* [141] and *"Compressorhead"* [33], the robotic drummer *"Haile"* [165], and the robotic marimba player *"Shimon"* [130]. Category 4 are non-anthropomorphic instruments that are either mechatronic augmentations of acoustic instruments, e.g. Disklaviers, or new acoustic analog instruments. Such robots tend to focus more on sonic nuance than on modeling human performance actions. Examples mentioned by Kemper include the *"Expressive Machines Musical Instruments (EMMI)"* see e.g. (128) and LEMUR's musical robots [138]. Category 5 focuses on cooperative musical robots, i.e. systems that combine human performance and robotic actuation in a shared interface. Kemper mentions *"String Trees"* [55] and Georgia Tech's *"Robotic Drumming Prosthetic Arm"*, which robotically augments the capabilities of the human body [163]. Finally, category 6 includes projects focused on sound and movement of individual actuators, such as the arrangement of *"Imperial March"* from Star Wars by Pawel Zadrożniak played on floppy disc drives [117], and large-scale sound sculptures featuring individual motors actuating resonant objects by Zimoun [175].

Although most of the musical robots described in the literature are physical robots, there are also examples of virtual robot musicianship. Some HRI taxonomies (e.g. [61, 169]) have explicitly distinguished between exposure to embodied versus depicted robots, since there is a growing body of research suggesting that physically embodied robots are perceived differently than virtual agents (see e.g. [86]). The ethical dimensions of virtual musicians and machine (or robo) ethics were explored in [22]. Topics discussed were, for example, *"vocaloids"*[5] such as *"Hatsune Miku"*, *"Kagamine Rin"* and *"Kagamine Len"*, as well as *"Utatane Piko"*. Hatsune Miku [29] has gained widespread success as a virtual musician, performing in concerts as a 3D hologram (for an analysis of the recent popularity of three-dimensional holographic performances in

[5]Vocaloids make use of the Yamaha Vocaloid software [158] for speech and singing sound synthesis.

popular music, see [103]). Miku was also available as a voice assistant hologram from the company Gatebox [51], a product somewhat similar to Amazon Alexa, although that product is now discontinued. Going beyond Japanese vocaloid characters, attempts have also been made to recreate representational simulations of celebrity musicians. Such efforts include avatar simulations of Kurt Cobain and resynthesis of Freddie Mercury's singing voice [22], as well as holograms of Gorillaz, Mariah Carey, Beyoncé, Michael Jackson, Old Dirty Bastard of Wu-Tang Clan, Eazy-E of N.W.A., and Tupac Shakur[6] [103]. Yet another example of virtual robot musicianship is the attempt to recreate Jerry Garcia from The Grateful Dead using Artificial Intelligence, as mentioned in [21].

Certain musical robots focus specifically on improving access to music-making. Such robots can, for example, take the form of wearable technology or prosthetic devices that support people with disabilities in their musicking. These musical robots can be considered a subcategory of the type of Digital Musical Instruments (DMIs) that are called *Accessible Digital Musical Instruments (ADMIs)*. ADMIs can be defined as *"accessible musical control interfaces used in electronic music, inclusive music practice and music therapy settings"* [48]. Examples include the *"Robotic Drumming Prosthetic Arm"*, which contains a drumming stick with a mind of its own; the *"Third Drumming Arm"*, which provides an extra arm for drummers[7]; and the *"Skywalker Piano Hand"*, which uses ultrasound muscle data to allow people with an amputation to play the piano using dexterous expressive finger gestures [164]. Another example is *"TronS"*, a prosthetic that produces trombone effects; *"Eleee"*, a wearable guitar prosthetic; and *"D-knock"*, a Japanese drum prosthetic [60].

The above section has shed light on the breadth of work carried out within the fields of musical robotics and robot musicianship. Given the broad range of different interfaces that may be considered musical robots, it is to be expected that no unified framework for evaluation of robot musicianship exists. A strategy that has been used to find appropriate evaluation methods proposed in the fields of new musical interfaces [42] and creativity support tools [121] is to look into other disciplines to identify existing methods, and to adapt those to the specific use case (if required). Building on this idea, the following sections review evaluation methods used in neighboring research fields to inform the method selection strategy for evaluation of musical robots. More specifically, the succeeding sections will explore methods used in Human–Computer Interaction (HCI), Human–AI Interaction (HAI), New Interfaces for Musical Expression (NIME), and Computational Creativity (CC).

[6] The latter performing with Dr. Dre and Snoop Dogg in a famous concert in 2012.

[7] Although this robot was primarily designed to be used by people without disabilities.

12.3 Evaluation in Human–Computer Interaction

The term *"evaluation"* is commonly used to describe a range of different activities and goals in the field of Human–Computer Interaction (HCI). Traditionally, considerable focus has been on evaluating *usability*. The ISO Standard 9241-11:2018 defines usability as the *"extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [65]. Usability evaluation encompasses methodologies for measuring the usability aspects of a system's user interface and identification of specific problems (see e.g. [110]). Many different evaluation methodologies have been proposed, each with different limitations and disadvantages. Evaluation processes can roughly be divided into two broad categories: *formative* evaluations, which take place during a design process, and *summative* evaluations, which take an already finished design and highlights its suitability for a specific purpose [136]. In other words, some evaluation methods can be applied already at an early stage of the design process, whereas others are intended to be used when the final interface design has been implemented.

Nielsen describes four main ways of conducting a user interface evaluation: *automatically*, *empirically*, *formally*, and *informally* [110]. *Automatic* usability evaluation involves using programs that can compute usability measures based on user interface specifications. An overview of the state of the art in usability evaluation automation was presented in [68]. *Empirical* evaluation involves testing an interface with real users. Many different methods may be employed, ranging from questionnaires to observations. A commonly used empirical method is *usability testing* [39]. Usability tests have the following six characteristics: the focus is on usability; the participants are end users or potential end-users; there is some artifact to evaluate (a product design, a system, or a prototype); the participants think aloud as they perform tasks; the data are recorded and analysed; and the results of the test are communicated to appropriate audiences. *Formal (model-based)* evaluation uses a model of a human to obtain predicted usability measures by either calculation or simulation. The goal of this procedure is to get some usability results before implementing a prototype and testing it with human subjects. A notable example is GOMS (Goals, Operators, Methods, Selection) modelling, which involves identifying methods for accomplishing task goals and calculating predicted usability metrics [85]. Finally, *informal evaluation* methods are non-empirical methods for evaluating user interfaces. Nielsen uses the term *usability inspection* to describe such methods that are based on having evaluators inspect the interface [110]. Commonly used methods include heuristic evaluation [111] and cognitive walk-through [94].

It should be noted that the topic of evaluation has been extensively debated and reviewed in the field of HCI over the years (see e.g. [10,100] for an overview)

and that opinions about who to best evaluate interfaces greatly differ depending on who you ask. Some have even suggested that usability evaluations may be harmful [53] and proposed to not use such traditional methods to validate early design stages or culturally sensitive systems, advocating instead for the use of more reflective and critical methods. Others have emphasized that the increased interest in experience-focused (rather than task-focused) HCI brings forward a need for new evaluation techniques [83]. As mentioned in [105], discussions about user-orientated quality assessment of technology have lately moved away from a focus on usability, satisfaction, efficiency, effectiveness, learnability, and usefulness; instead, attempts have been made to shift focus to the user experience and the wider relationship between people and technology, exploring concepts such as engagement, pleasure, presence, and fun (see [13,102,126] for further reading). In [105], the authors suggest that three aspects should be considered simultaneously when designing and evaluating technology: *functionality*, *usability*, and *user experience*. In this context, *functionality* relates to technical issues (for example, which features that should be provided by the device, and aspects of performance, reliability, and durability). *Usability* relates to user issues. Finally, *user experience* focuses on the wider relationship between the product and the user, i.e. the individual's personal experience of using it [104,105].

Although the word *user experience* (UX) has been around since the 1990s, there is still no widely accepted definition of the term (see e.g. [59,93]). The ISO Standard 9241 defines user experience as *"a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service"* [65]. In a side note, this is said to include *"all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use"*. MacDonald and Atwood [100] suggest that a major challenge for UX evaluators is the lack of shared conceptual framework, despite that multiple models have been proposed (see e.g. [45,113]). An example of a tool for evaluation of user experience is the *User Experience Questionnaire*, which consists of 26 items arranged into six scales: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty (a shorter version of the UEQ with only 8 items is displayed in Table 12.3) [92,156]. A model of UX introduced by Hassenzahl suggested that products have both *pragmatic* (e.g. an ability to help users achieve goals) and *hedonic* attributes (e.g. an ability to evoke feelings of pleasure and self-expression) [58]. The authors of [100] stressed that UX methods tend to focus solely on hedonic attributes, while usability evaluation methods mostly are used to capture pragmatic attributes, voicing the need for methods that can seamlessly integrate both hedonic and pragmatic feedback.

Apart from being evaluated from a usability and user experience perspective, musical robots may also benefit from accessibility evaluation. This is particularly important for musical robots that find themselves at the intersection of robot musicianship and ADMIs. Accessibility can be conceptualized as usability for a population with the widest range of user needs, characteristics,

TABLE 12.3
Items in the shorter version of the User Experience Questionnaire (UEQ) [156].

1	Obstructive	□□□□□□□	Supportive
2	Complicated	□□□□□□□	Easy
3	Inefficient	□□□□□□□	Efficient
4	Confusing	□□□□□□□	Clear
5	Boring	□□□□□□□	Exciting
6	Not interesting	□□□□□□□	Interesting
7	Conventional	□□□□□□□	Inventive
8	Usual	□□□□□□□	Leading edge

and capabilities (see [65]). This fits within the *universal design* [26] or *design for all* [40] philosophies, which may be used as starting point for accessibility evaluations. The seven design principles of universal design include: *equitable use, flexibility in use, simple and intuitive use, perceptible information, tolerance for error, low physical effort*, and *size and space for approach and use* [26]. For an overview of methods for evaluating accessibility, usability, versus user experience, as well as a discussion of strengths and weaknesses of these concepts, see [119].

12.4 Evaluation in Human–Robot Interaction

Researchers have stressed that the experience of interacting with robots is different from interacting with other technologies and that such experiences often involve a strong social or emotional component [174]; people tend to treat robots similar to how they treat living objects and ascribe them life-like qualities [46, 152]. Robots' physical and social presence, and their tendency to evoke a sense of agency, i.e. a capacity to act that carries the notion of intentionality [37], creates a complex interaction different from the one involving other artifacts and technologies [174]. This poses certain challenges when it comes to evaluation. Even if interactions with robotic technology and themes discussed in HCI research have things in common [44], HRI researchers have emphasized that HRI is different from Human–Computer Interaction [31], and that evaluation methods from HCI should be applied to HRI with care [174].

Young et al. [174] published a summary of methodologies, techniques, and concepts from both HCI and HRI research, focusing on strategies that they deemed useful for the unique and deep social component of interactions between a person and a robot. The following evaluation approaches were discussed: (a) *task completion and efficiency*, (b) *emotion*, (c) *situated personal experience*, and (d) *frameworks for exploring social interactions with robots*. The authors suggest that (a) can be used as a wider part of evaluation of social

HRI but that other techniques are required for a more comprehensive view of the entire HRI experience. For (b), one suggestion was to monitor biological features (e.g. heart rate, blood pressure, and brain activity, number of laughs, and so on). However, given the holistic, rich, and multi-faceted nature of social interactions, such simplifications of emotion into quantities and discrete categories will have limitations. Other methods proposed within category (b) included self-reflection using think-aloud and interviews. For (c), the authors emphasized that the situated holistic experience of interacting with a robot includes aspects of social structure, culture, and context. Researchers have stressed the importance of accepting the complex nature of interactions [28,137] and proposed to focus on uncovering themes and in-depth descriptions of such complexities [11, 62, 67]. Young et al. (173) provide examples of how this can be tackled using qualitative techniques such as participant feedback and interviews, grounded theory, cultural- or technology probes, contextual design, and in situ context-based ethnographic and longitudinal field studies. They stress that context sensitive evaluation should value that individuals have unique, culturally grounded experiences; that one should take care when generalizing across people (citing [15,28]); and that evaluators themselves carry culturally rooted personal biases toward robots, participants, and scenarios (see [28]). Complementary to above-described methods, evaluators can use specific frameworks (d) when exploring social interaction with robots. For example, Norman's three-level framework for analysing how people interact with and understand everyday objects may be used [112]. This framework highlights different temporal stages of interaction with a product: *the initial visceral impact*, *the behavioral impact during use*, and *the reflective impact after* interacting with a product.

Based on above discussion, Young et al. (173) present an attempt to classify the rich interaction with a robot into articulated concepts: *visceral factors, social mechanisms,* and *social structures*. These perspectives can be integrated into existing HCI and HRI evaluation methods and provide a new vocabulary that encourages investigators to focus more on emotional and social aspects of interaction. The *visceral factors* of interaction involve the immediate and automatic human responses on a reactionary level that is difficult to control, for example, instinctual frustration, fear, joy, and happiness. The *social mechanics* involve the application of social language and norms. This includes higher-level communication and social techniques; for example, gestures such as facial expressions and body language, spoken language, and cultural norms such as personal space and eye-contact. Finally, *social structures* refer to macro-level social structures, i.e. the development and changes of the social relationships and interaction between two entities over a longer period of time. This can be seen as the trajectory of the two other perspectives, and relates to how a robot interacts with, understands, and modifies social structures. As an illustrative example, they mention how cleaning-robot technology in homes may shift who is responsible for cleaning duties [46]. The authors describe how these three perspectives can serve as tools throughout the evaluation process at various

TABLE 12.4
The USUS Evaluation Framework, as presented in [166].

Factor	Indicator	Methods
Usability	Effectiveness	Expert evaluation, user studies
	Efficiency	
	Learnability	
	Flexibility	
	Robustness	
	Utility	Questionnaires, interviews
Social Acceptance	Performance expectancy	Questionnaires, focus groups
	Effort expectancy	
	Self efficacy	
	Forms of grouping	
	Attachment	
	Attitude toward using technology	Questionnaires
	Reciprocity	
User Experience	Emotion	Questionnaires, physiological measurements, focus groups
	Feeling of security	
	Embodiment	Questionnaires, focus groups
	Co-experience	
	Human-oriented perception	Questionnaires
Social Impact	Quality of life	Questionnaires, focus groups, interviews
	Working conditions	
	Education	
	Cultural context	

stages, from designing a study to conducting it, to analysing collected data.

A comprehensive overview of evaluation methods in HRI was presented in [78]. This book discusses questionnaires for HRI research, processes for designing and conducting semi-structured interviews, standardized frameworks for evaluation, evaluation of user experience of human–robot interactions (see [96]), evaluations based on ethology and ethnography, as well as recommendations for reliable evaluations. One of the discussed frameworks is the USUS Evaluation Framework [166], which is described in Table 12.4. It consists of a theoretical framework based on a multi-level model involving factors split into indicators, extracted and justified by literature review, and a methodological framework consisting of a mix of methods derived from HRI, HCI, psychology, and sociology. The USUS Evaluation Framework was later reformulated by Wallström and Lindblom into the USUS Goals Evaluation Framework [159], after the authors identified a lack of evaluation methods that included UX goals, i.e. high-level objectives driven by the representative use of the system, in social HRI.

12.5 Evaluation in Human–AI Interaction

Researchers in the HCI community have proposed several guidelines and recommendations for how to design for effective human interaction with AI systems. Several large companies have also published white papers aiming

to serve as guidance for development of AI systems (see e.g. [116] and [64]). Amershi et al. proposed 18 generally applicable design guidelines for Human–AI Interaction (HAI) in [3]. These guidelines are intended for AI-infused systems, i.e. systems that have features harnessing AI capabilities that are exposed directly to an end user. The AI design guidelines in [3] are separated into categories depending on when during the user's interaction that they are applied: *initially, during interaction, when something goes wrong*, and *over time* [3]. A summary of the guidelines is presented in Table 12.5. Although described as guidelines rather than evaluation criteria, these points can be used for evaluation purposes (see e.g. [50]). A research protocol for evaluating Human-AI Interaction based on the guidelines was recently published in [95].

Eight of the guidelines presented in [3] overlap with the principles for Mixed-Initiative Systems by Horvitz [63]. A taxonomy for Mixed–Initiative Human–Robot Interaction was presented in [70][8]. Related areas of research that may be interesting to explore for the purpose of identifying evaluation methodologies suitable for musical robots include work on Mixed-Initiative Co–Creative Systems [172], Mixed–Initiative Creative Interfaces [36], and Human–AI Co–Creativity [122]. To the author's knowledge, there is still no unified framework for evaluation of Human–AI co-creative systems (although some attempts have been made, see e.g. [82]). Potential pitfalls when designing such systems were discussed in [17]. The pitfalls were identified starting from three speculation prompts: issues arising from (a) *limited AI*, (b) *too much AI involvement*, and (c) *thinking beyond use and usage situations*. The first category includes issues related to invisible AI boundaries, lack of expressive interaction, and false sense of proficiency; the second to conflicts of territory, agony of choice, and time waste; and the third to AI bias, conflict of creation and responsibility, and user and data privacy.

Another relevant area of research in this context is the field of EXplainable AI (XAI) [54]. XAI aims to interpret or provide a meaning for an obscure machine learning model whose inner workings are otherwise unknown or non-understandable by the human observer [4]. This relates to the notion of *"explainable agency"*, which refers to autonomous agents such as robots explaining their actions, and the reasons leading to their decisions [91]. In the context of robots, a range of other related terms are also used to explore similar concepts, e.g. understandability, explicability, transparency, and predictability (see [4] for an overview). A systematic review on explainable agency for robots and agents was presented in [4], highlighting a considerable lack of evaluations in the reviewed papers. A framework and paradigm for evaluation of explanations of AI was provided in [47]. This framework suggests that an explanation needs

[8] *"Mixed Initiative Interaction HRI (MI-HRI)"* is defined as *"A collaboration strategy for human–robot teams where humans and robots opportunistically seize (relinquish) initiative from (to) each other as a mission is being executed, where initiative is an element of the mission that can range from low-level motion control of the robot to high-level specification of mission goals, and the initiative is mixed only when each member is authorized to intervene and seize control of it".*

TABLE 12.5
The 18 Human-AI interaction design guidelines proposed by Amershi et al. [3], categorized by when they likely are to be applied during interaction with users.

No	AI Design guideline	Description	Timepoint
1	Make clear what the system can do.	Help the user understand what the AI system is capable of doing.	Initially
2	Make clear how well the system can do what it can do.	Help the user understand how often the AI system may make mistakes.	
3	Time services based on context.	Time when to act or interrupt based on the user's current task and environment.	During interaction
4	Show contextually relevant information.	Display information relevant to the user's current task and environment.	
5	Match relevant social norms.	Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	
6	Mitigate social biases.	Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	
7	Support efficient invocation.	Make it easy to invoke or request the AI system's services when needed.	When wrong
8	Support efficient dismissal.	Make it easy to dismiss or ignore undesired AI system services.	
9	Support efficient correction.	Make it easy to edit, refine, or recover when the AI system is wrong.	
10	Scope services when in doubt.	Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	
11	Make clear why the system did what it did.	Enable the user to access an explanation of why the AI system behaved as it did.	
12	Remember recent interactions.	Maintain short term memory and allow the user to make efcient references to that memory.	Over time
13	Learn from user behavior.	Personalize the user's experience by learning from their actions over time.	
14	Update and adapt cautiously.	Limit disruptive changes when updating and adapting the AI system's behaviors.	
15	Encourage granular feedback.	Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	
16	Convey the consequences of user actions.	Immediately update or convey how user actions will impact future behaviors of the AI system.	
17	Provide global controls.	Allow the user to globally customize what the AI system monitors and how it behaves.	
18	Notify users about changes.	Inform the user when the AI system adds or updates its capabilities.	

to (1) *provide knowledge*, (2) *be trustworthy*, (3) *be useful*, (4) *update the receiver's estimation about the probability of events occurring*, and (5) *change the receivers mental model*.

12.6 Evaluation of New Interfaces for Musical Expression

Evaluation of acoustic musical instruments was discussed by Campbell in [18]. The author posed the question *"if we are to optimize a musical instrument, who determines success?"* The author suggests that this should either be the musician playing the instrument, or the listener who hears the sound. Campbell points out that musicians and physicists approach evaluation differently, and they often lack a common vocabulary to discuss these differences. He emphasized the importance of understanding cross-modal interference and how this may influence judgment of instrument quality, mentioning for example differences in perceived tone quality of the piano (a phenomenon caused by cross-modal interference between auditory and haptic channels) and the importance of distracting visual cues when evaluating the quality of violins and brass instruments.

When it comes to the digital domain, the topic of evaluation has also been extensively debated in the fields dedicated to New Interfaces of Musical Expression (NIME) and Digital Musical Instruments (DMIs) for quite some time. Attempts have been made to explore what the word *"evaluation"* means for the NIME community, with findings suggesting that there are different understandings of the term [9]. A review of papers published in the NIME conference proceedings revealed that the word often is used to denote the process of collecting feedback from users to improve a prototype. Others use the term to assess the suitability of existing devices for specific tasks, or to compare devices. In addition, evaluation is sometimes used to describe emerging interaction patters when using devices. As pointed out by Barbosa et al. [9], complicating factors involved in NIME evaluation include that several stakeholders often are involved in the design of the instruments, and the requirements of one stakeholder may not necessarily intersect those of another. Barbosa et al. also stress that the time window of the evaluation[9], as well as the level of expertise, can influence the evaluation results.

Traditionally, evaluation methods for NIMEs and DMIs have largely been based on frameworks used in the field of Human–Computer Interaction (HCI). Wanderley and Orio [160] discuss the application of such methodologies in the evaluation of input devices for musical expression. In particular, they

[9]The experience of playing a musical instrument usually changes the more you play on the instrument.

focused on specific tasks used in HCI to measure the performance of an input device in the music domain. More specifically, Wanderley and Orio propose to use a set of musical tasks for *usability evaluation*. For a musical instrument, such tasks could focus on the production of musical entities and include the generation of, for example *isolated tones*, i.e. pitches at different frequencies and loudness levels; *basic musical gestures* like glissandi, trills, vibrato, and grace notes; and *musical phrases*, such as scales and arpeggios, as well as more complex contours with different speeds and articulations. In addition, tasks could focus on reproducing continuous timbral changes or different rhythms for such musical entities, given a specific loudness. Wanderley and Orio also propose a set of relevant features to be tested in usability evaluations of controllers used in the context of interactive music: *learnability, explorability, feature controllability*, and *timing controllability*.

In the framework for musical instruments proposed by Kvifte and Jensenius, the authors discuss three perspectives of instrument description and design: that of *the listener*, *the performer*, and *the instrument constructor* [89]. Barbosa et al. [9] highlight that, for example, playability might be important for a performer, but not for an audience. The idea that there are many perspectives from which one can view the effectiveness of an instrument was also stressed by O'Modhrain [114], who suggested that if performance is considered a valid means of evaluating a musical instrument, a much broader definition than what is typically used in HCI is required. O'Modhrain emphasizes that in addition to players and audiences, there are also composers, instrument builders, component manufacturers, and perhaps also customers, to consider. The different stakeholders can have different views of what is meant by the term evaluation. A complicating factor in this context is that the boundaries between roles usually are blurred in DMI design (this is usually not the case for design of acoustic musical instruments). Since DMI designs can be evaluated from multiple perspectives, different techniques and approaches are required. O'Modhrain aims to provide a structure to these competing interests by providing a framework for evaluation that enables performers, designers, and manufacturers to more readily identify the goal of an evaluation, and to view their methods in the light of prior work. Her framework includes a summary of methods that a given stakeholder might use to evaluate a DMI against a given design goal. Possible evaluation goals include *enjoyment, playability, robustness,* and *achievement of design specifications*.

Young and Murphy [173] suggest that the evaluation of DMIs should focus on *functionality*, *usability*, and *user experience*. The functionality testing should aim to highlight potential issues before longitudinal studies are carried out, i.e. the usability and user experience studies. The initial stages of a device's evaluation should focus on capturing low-level device characteristics, thus creating a generalized device description (for example, through evaluation of the musical tasks proposed in [160]). This should be followed by a process in which a device is reduced to its physical variables in terms of a taxonomy of input. In this step, you contextualize a device's evaluation in terms of

stakeholders, questioning who is evaluating the device and why. Several HCI paradigms exist that can be augmented to fit these processes.

Useful tools for evaluation and classification of NIMEs include, for example, the phenomenological dimension space for musical devices introduced by Birnbaum et al. [12]. This framework can be used to describe a musical instrument along a set of seven axes: *required expertise, musical control, feedback modalities (outputs), degrees of freedom (input), inter-actors, distribution in space,* and *role of sound*. Another example is the epistemic dimension space for musical devices presented by Magnusson in [101], which includes eight parameter axes: *expressive constraints, autonomy, music theory, explorability, required foreknowledge, improvisation, generality,* and *creative simulation*. Yet another useful framework was provided by Jordà in [74]. This framework focuses on the musical output diversity of the instrument and how the performer can control and affect this diversity, dividing instrument diversity into *macro-diversity*, *mid-diversity*, and *micro-diversity*. Macro-diversity determines the flexibility of an instrument to be played in different contexts, music styles or assuming varied roles; mid-diversity refers to how different two performances or compositions played with the instrument can be; and micro-diversity to how two performances of the same piece can differ.

When it comes to evaluation of user experience for NIMEs, a recent review was published by Reimer and Wanderley [120]. Findings suggested that UX-focused evaluations typically were exploratory and that they were limited to novice performers. The authors propose to use the *"Musicians Perception of the Experiential Quality of Musical Instruments Questionnaire (MPX-Q)"* to compare UX for different instruments [135]. This questionnaire is based on psychometric principles and consists of three interrelated subscales: (1) *experienced freedom and possibilities*, (2) *perceived control and comfort*, and (3) *perceived stability, sound quality, and aesthetics*. Referring to [173] and building on ideas previously discussed in [42], Reimer and Wanderley also propose that standardized frameworks to evaluate UX in other fields could be adapted for NIME evaluation (see e.g. [38] for an overview), mentioning for example the *"Gaming Experience Questionnaire (GEQ)"* from ludology[10].

The suitability of HCI evaluation tools, which put emphasis on technological aspects of musical instruments and describe them as *"devices"* with properties viewed from a *"usability"* and *"accessibility"* perspective, has been widely debated in the NIME community. Some have even questioned the use of the word *"evaluation"*, proposing to instead employ the term *"user experience study"*, thereby broadening the scope of such work to acknowledge that while ergonomics and efficiency are important, they are not the primary determinants of the quality of a musical interface [71]. Stowell et al. [150] suggested that while the framework proposed by Wanderley and Orio [160] is useful, it has drawbacks. For example, the reduction of musical interaction to simple tasks

[10]The study of gaming.

may compromise the authenticity of the interaction. Since musical interactions involve creative and affective aspects, they cannot simply be described as tasks for which aspects such as completion rates are measured [150]. Task-based methods may be suited to examine usability, but the experience of the interaction is subjective and requires alternative approaches for evaluation. Stowell et al. propose that the following questions should be considered when evaluating interactive music systems: (1) *Is the system primarily designed to emulate the interaction provided by a human, or by some other known system?* (2) *Is the performer's perspective sufficient for evaluation?* (3) *Is the system designed for complex musical interactions, or for simple/separable musical tasks?* (4) *Is the system intended for solo interaction, or is a group interaction a better representation of its expected use pattern?* (5) *How large is the population of participants on which we can draw for evaluation?* They also present two methods for evaluation of musical systems: a qualitative approach using structured discourse analysis and a quantitative musical Turing-test method. Finally, they suggest that the design of evaluation experiments should aim to reflect the authentic use context as far as possible.

Rodger et al. [127] questioned the adoption of tools from traditional HCI to understand what constitutes a good musical instrument. The implication of viewing musical activities as something compromised of a "device" and a "user" is that the instrument is considered an entity with a set of intended functional behaviors, known to the designer and employed by the user, for the purpose of a specific goal. This is a limiting view of how musicians interact with instruments. There are many examples in which musical instruments are used in manners that differ from the original intended design. The idea that the instrument should be assessed by how readily it supports an intended design function can also be questioned by what Rodger et al. call "instrument resistance", i.e. that the effortfulness of playing an instrument may serve as a source of creativity. Viewing musicians as users of musical devices results in conceptual issues, since musicians vary in their capabilities and histories of embodied knowledge. As such, the idea of a "prototypical user" doesn't suit this context. The functional properties of an instrument can only be meaningfully understood relative to the capabilities of specific musician at a specific period in her musical development. Moreover, it is hard to make sense of what a musician does with an instrument if divorced from both the immediate and extended socio-cultural context. Rodger et al. therefore propose an evaluation approach in which instruments are understood as processes rather than devices, and musicians are viewed as agents in musical ecologies, i.e. a system compromising an agent and environment, rather than users. Evaluations of instruments should align with the specificities of the relevant processes and ecologies concerned. In this context, a specifity is defined as *"the effective components of the musician-instrument system relative to the relevant musical activities and contexts of interest"*. In other words, the evaluation should be relative to its environmental context, and not focus on a generalizable methodology based on a prototypical user. The consequence of this stance is

that instruments may mean different things to different musicians.

Also El-Shimy and Cooperstock [42] stressed that the nature of musical performance requires that designers re-evaluate their definition of user "goals", "tasks", and "needs". They stress the importance of creativity and enjoyment rather than efficiency [42]. El-Shimy and Cooperstock reviewed literature focused on user-driven evaluation techniques offered by HCI, ludology, interactive arts, and social-science research, exploring aspects such as affect, fun, pleasure, flow, and creativity. They present a set of principles for user-driven evaluation of new musical interfaces involving: (1) *validating the basis*, (2) *investigating suitable alternatives to "usability",* and (3) *tailoring evaluation techniques.* El-Shimy and Cooperstock argue that qualitative and mixed research methods are particularly suited for studies of non-utilitarian systems and propose to use qualitative experiments to develop hypotheses that later can be verified through quantitative studies (as opposed to the traditional approach in which hypotheses are formed before an experiment). Qualitative research methods mentioned include interviews, discussions, case studies, and diaries. Analysis methods include, for example, content analysis, which operates on the principle of grounded theory. El-Shimy and Cooperstock encourage designers to tailor existing evaluation techniques to their own needs or to devise new ones if necessary.

Based on the above discussion, as well as what was briefly mentioned in Section 12.3, we can conclude that there has been a shift from task-based and usability-driven design to more experience-based design and evaluation (so called third wave) HCI, especially within creative and artistic contexts [42]. Third wave HCI is said to be particularly suited to the design and evaluation of novel interactive musical interfaces [42]. Building on the work by Rodger et al. [127], as well as Waters' notion of *"performance ecosystems"*, in which music activities can be understood as a dynamical complex of interacting situated embodied behaviors [161], Jack et al. [69] propose to view DMIs as situated, ecologically valid artefacts, which should be evaluated using qualitative and reflective processes focusing on sociocultural phenomena, rather than first wave HCI techniques.

Finally, when it comes to accessibility of musical expression, it should be noted that the field of Accessible Digital Musical Instruments (ADMIs) still appears to lack a formal framework for evaluation [48], although attempts have been made to formulate design principles and classification methods (see [32, 49, 57]). A set of design guidelines that could be used for the purpose of evaluation were proposed by Frid [49], including *expressiveness, playability, longevity, customizability, pleasure, sonic quality, robustness, multimodality*, and *causality.* A dimension space for evaluation of ADMIs was proposed by Davanzo and Avanzini in [32], in which eight axes are grouped into two subsets: *target users and use contexts* (use context, cognitive impairment, sensory impairment, physical impairment), and *design choices* (simplification, adaptability, design novelty, physical channels). Moreover, Lucas et al. explored ecological perspectives of human activity in the use of DMIs and assistive

technology in [99]. The authors used the Human Activity Assistive Technology (HAAT) [27] and the Matching Person and Technology (MPT) [134] frameworks to design and evaluate bespoke ADMIs, concluding that a shortcoming of these tools is that they are biased toward describing persons with disabilities from an external perspective.

12.7 Evaluation in Computational Creativity

Computational Creativity (CC) can be considered a sub-field of Artificial Intelligence (AI) [118]. Computational Creativity is the philosophy, science, and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative [25]. Evaluation of CC systems focuses on determining whether a system is acting creatively or not [118]. However, evaluation attempts in this domain have been found to lack rigor; there is no consensus on how to evaluate creative systems, and the reliability and validity of the proposed methods are in question [90]. Musical Metacreation (MuMe), a subarea of CC which aims to automate aspects of musical creativity with the aim of creating systems or artifacts deemed creative by unbiased observers, has also been found to be characterized by little systematic evaluation [1]. A historical perspective on how CC researchers have evaluated, or not evaluated, the creativity of their systems was presented by Anna Jordanous in [77]. In this work, Jordanous also address the question of how to choose an evaluation method and how to judge its quality via five meta-evaluation standards for comparison of evaluation methods in creativity: *correctness, usefulness, faithfulness as a model of creativity, usability of the methodology,* and *generality.*

Several different theories of creativity evaluation have been proposed throughout the years. Lamb et al. [90] suggest to group these based on their theoretical perspective, building on the taxonomy known as the four Ps: *Person, Process, Product*, and *Press* (see [123]). The four Ps were introduced to Computational Creativity by Jordanous. *Person* or *Producer*[11] is the human or non-human agent that is judged as being creative. Person theories aim to discover which traits (personality, emotional, cognitive) that distinguishes a more creative person from a less creative one. *Process*, on the other hand, refers to a set of internal and external actions that the agent may take when producing creative artifacts. Process theories study the actions that are undertaken in such contexts. This perspective focuses on how creative products are made, i.e. the cognitive steps that must be taken for an activity to be creative. *Product* is an artifact, for example a musical piece, which is seen as creative

[11] Jordanous suggested to use the term Producer instead of Person to emphasize that the agent does not need to be a human.

or as having been produced by creativity. Product theories study what it is about a certain product that makes it creative. Finally, *Press* refers to the surrounding culture that influences the other Ps in the model. Press theories study what it is in a culture that leads to the view that something is creative, and what kind of social effect a product needs to have to be called creative.

Person methods include psychometric tests or the study of famous creative people. Some have also attempted to measure the personal traits of computers. *Process* theories are often useful when modelling human creativity. Process evaluations tend to either place the system in a category or use a qualitative analysis of the system's process strengths and weaknesses. They are often somewhat descriptive in their nature, i.e. not always easy to apply to evaluation. Examples include the FACE and IDEA models (see [24]). The FACE[12] model describes creative acts performed by a software, whereas the IDEA model describes how such acts can have an impact on an audience. Another example is the SPECS model, which evaluates systems based on 14 factors that were identified through studies of how humans define creativity [76].[13] The SPECS model is divided into three steps [77]. Step 1 focuses on identifying a definition of creativity that your system should satisfy to be considered creative. Step 2 uses step 1 and focuses on clearly stating what standards you use to evaluate the creativity of the system. Step 3 focuses on testing the creative system against the standards stated in step 2 and reporting on these results. Moving on to the *Product* perspective, the focus lies on the artifact itself (e.g. a music piece or performance) as creative or as having been produced by creativity [90]. Common criteria are "novelty" and "value". Lamb et al. [90] suggest that if using such terms, one should define the specific audience for whom the system's products should be valuable. Ritchie suggested that "typicality" should be used rather than novelty, since creative systems should reliably generate both typical and valuable output [124]. *Product* methods include the Consensual Assessment Technique (CAT), in which a team of human experts evaluates a product [2], and the modified Turing Test (see [90]). The latter focuses on a test in which human subjects are challenged to figure out which products that are human versus computer-created in a set; if they cannot do it, then the computer system is considered creative.[14] Finally, *Press* methods include the Creative Tripod [23] and strategies to measure audience impact (see [90]). The Creative Tripod focuses on whether a system demonstrates skill, imagination, and appreciation, three qualities that are required to be deemed creative.

A review of Creativity Supporting Tools (CSTs) was presented in [121]. In

[12]The FACE model has been found to rank musical improvisation systems in the opposite order of other evaluation methods, and its validity has therefore been questioned [75].

[13]These factors include: active involvement and persistence; dealing with uncertainty; domain competence; general intellect; generation of results; independence and freedom; intention and emotional involvement; originality; progression and development; social interaction and communication; spontaneity/subconscious processing; thinking and evaluation; value; and variety, divergence, and experimentation.

[14]However, the modified Turing Test has been criticized on a number of points, see e.g. [118].

this work, the authors discuss six major points that researchers developing CSTs should consider in their evaluation: (1) *clearly define the goal of the CST*; (2) *link to theory to further the understanding of the CST's use and how to evaluate it*; (3) *recruit domain experts, if applicable and feasible*; (4) *consider longitudinal, in-situ studies*; (5) *distinguish and decide whether to evaluate usability or creativity*; and (6) *as a community, help develop a toolbox for CST evaluation* [121]. Related to this, Karimi et al. [82] provided a framework for evaluating creativity in co-creative systems, mentioning four questions that could guide such evaluation: (1) *Who is evaluating the creativity?* (2) *What is being evaluated?* (3) *When does the evaluation occur?* (4) *How is the evaluation performed?*

Agres et al. provided a theoretical motivation for more systematic evaluation of Musical Meta-Creation and computationally creative systems in [1]. The authors present an overview of methods to assess human and machine creativity, dividing creative systems into three categories: (1) *those that have a purely generative purpose*, (2) *those that contain internal or external feedback*, and (3) *those that are capable of reflection and self-reflection*. They present examples of methods to help researchers evaluate their creative systems, test their impact on an audience, and build mechanisms for reflection into creative systems. Other relevant references in the context of Musical Meta-Creation include [41], which describes an evaluation study of several musical meta-creations in live performance settings, and [155], which presents a discussion on evaluation of musical agents. The latter divides evaluations of MuMe systems into informal evaluations and formal evaluations. Informal evaluations do not involve formalized research methodologies (they usually take place as part of the software development), whereas formal evaluations use formalized methodologies to assess the success of systems.

Finally, it should be noted that some question the mere idea of creativity evaluation, and whether this is possible at all. For example, Baer [5] suggested that there are many creative skills, but no underlying process which informs them all. To be creative in one domain does not necessarily imply that you are creative in other domains [6]. Calling a person or process creative without specifying the domain is therefore not considered scientific. Others have suggested that creativity cannot be quantified. For example, Nake [108] suggests that quantification of creativity is an American invention and that there are risks commodifying creativity by framing it as an object one must have a certain amount of, as opposed to considering it a quality that emerges in a social context. Some argue that computational creativity should not be measured by human standards and that it is more interesting to investigate what computers produce according to their own non-human standards [98]. On the other hand, others claim that creativity is inherently human and thus never can be present in computers (although many counterarguments have been presented through the years, see e.g. [14, 106]).

12.8 Evaluation of Musical Robots

To explore the range of different methods used for evaluation of musical robots, a comprehensive search for the keyword "robot" in the title and abstract fields of papers published in the Computer Music Journal, the Journal of New Music Research, and the Leonardo Music Journal was performed. The proceedings of the International Conference on New Interfaces for Musical Expression (NIME) were searched using the same strategy. Chapters from the books *"Musical Robots and Multimodal Interactive Systems"* [144] and *"Robotic Musicianship – Embodied Artificial Creativity and Mechatronic Musical Expression"* [163] as well as the PhD thesis *"Expressive Musical Robots: Building, Evaluating, and Interfacing with an Ensemble of Mechatronic Instruments"* [107] were also skimmed to identify texts that could be included in the review.

In total, 14 journal articles and 50 papers from the NIME proceedings were identified. A total of 7 chapters were selected from the first book, 5 from the second, and 7 from the thesis. From this initial dataset, studies were selected to be included in the review by comparing the information presented in the publication against an inclusion criterion. In order to be included, the studies had to fulfill the following requirements: the publication had to describe a musical robot (the authors had to explicitly define their interface as a robot, and the robot should produce sounds), and an evaluation method must have been used. For this review, I adopted a wide definition of the term *"evaluation"*, including for example performances and installations displayed at public venues as methods. The review focused primarily on summative evaluations performed at the end of a study. The application of the inclusion criteria reduced the dataset from 83 to 62 publications. Information about the employed evaluation strategies was subsequently summarized per publication to identify reoccurring themes.

The publications were initially searched for the keywords "evaluation" or "evaluate", to identify sections describing evaluation methodologies. A total of 38 publications explicitly made use of these terms, corresponding to 61%.[15] Although not explicitly mentioning the term, the remaining publications did indeed describe methods that could be considered evaluation strategies. For example, case studies, measurements, performances, composition processes, and different types of empirical experiments were mentioned, without referring to "evaluation" explicitly.

Severeal evaluation strategies for musical robots could identified. First of all, the authors mentioned both so called "objective" and "subjective" evaluations. The objective evaluations usually focused on technical and mechanical aspects, e.g. sound quality, through measurements and analysis of audio or sensor data. Reoccurring variables were dynamic range, pitch, timbre, speed, repetition,

[15]Instances in which the word "evaluate" was used more generally, e.g. to evaluate a mathematical expression, were excluded.

and latency. Some explicitly mentioned that they measured variables relating to *musical expressivity*, e.g. timbre control, peak loudness, decay control, pitch control, etc. When it comes to evaluation of acoustic quality, this was often computationally done, using software and algorithms. However, in some cases humans also analyzed the sonic output through inspection. Building on the methodology proposed by Wanderley and Orio [160], objective evaluations were often based on programming the robot to perform simple musical tasks. For example, robots were instructed to play different polyrhythms or at different dynamics.

It has been suggested that musical robotic systems need to be tested in performance and installation settings for their functionality to be properly understood [107]. Many of the publications in the explored dataset described processes involving composing pieces specifically for the robot, performing with the robot in front of an audience, organising concerts with the robot, or showcasing the robotic system as a music installation open to the public. Such practices were generally described without framing them as evaluation methods. For performances and installations, accounts of methods going beyond the generic description of collecting "informal feedback" from musicians or exhibition visitors were rare. Many papers included general statements describing that the robot had been used in "various applications/contexts". The methodology for analysis of these processes was not clearly defined (one exception was the suggestion to use structured observation in [16]).

In general, there was an overall tendency to focus on "performance analysis" of the technical systems. Relatively few publications included descriptions of subjective evaluations. When it comes to the use of subjective – or qualitative – methods, a rather common strategy was to use questionnaires. Interviews, think-aloud, and observational methods seemed to be less common. Several authors described using different types of listening tests, followed by questionnaires, often involving ratings on Likert scales. When it comes to the use of standardized tools for questionnaires, the vast majority developed their own questions (one exception was the *"Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST)"* questionnaire (see [35]) used in [164]). Questions focused, for example, on rating aspects related to musical performance (e.g. gesture expressiveness) or agreement with statements about robot musicianship (e.g. "the robot played well", see [132]). Although aspects such as "(user) experience" and "usability" were mentioned by some authors, no standardized methods were used to evaluate such dimensions. However, characteristics such as strengths, weaknesses, and frustrations of a system, were sometimes discussed. Overall, relatively little attention was given to aspects concerning the interaction with and creative possibilities of the systems, such as co-creation and agency.[16]

Another common theme was to compare robot-produced sounds and humanly created ones, using methods reminiscent of the modified Turing tests.

[16]This might, however, be a direct result of the selection of publication venues.

Such procedures involved, for example, listening tests in which participants listened to a piece generated by an expert performer who played on an acoustic instrument and compared this to the performance of a robot performer playing the same piece. Participants were either asked which performance they thought was generated by a human or rated the music on a set of scales (or both).

Finally, apart from public installations and performances, most of the evaluation experiments were performed in lab settings. Few evaluations involved many participants; often less than 10 subjects took part. This was usually motivated by the fact that experts were invited as participants, or that there was a need for subjects with very specific skill sets (e.g. they needed to not only be able to improvise freely on an instrument but also have good computer skills).

12.9 Prospects for Future Research

To inform the selection of evaluation strategies for musical robots, this chapter has provided an overview of evaluation methods used in the fields of Human–Computer Interaction, Human–Robot Interaction, Human–AI Interaction, New Interfaces for Musical Expression, and Computational Creativity. The chapter has highlighted not only the breadth of systems that can be considered *musical robots* but also the heterogeneous methods employed to evaluate such systems, as well as the sometimes conflicting views on what constitutes an appropriate evaluation method. Based on this heterogeneity, it seems somewhat naive to suggest that it would be possible to develop a general evaluation framework that could be applied to all musical robots. There is no unified reply to the question "what to evaluate" and how to conduct such an evaluation, nor an undivided view of what the term "evaluation" actually means.

Evaluation criteria that are important for one system might not be relevant for another one, and different stakeholders will have different perspectives on what is relevant to explore. For example, the requirements for a physical performance robot playing on stage together with musicians or other robots might be significantly different from the requirements for a virtual agent involved in a collaborative composition task. Different robot designs also pose specific challenges that might not necessarily be relevant to other categories of musical robots. For example, musical robots that act as wearable devices or prosthetic devices raise questions that relate to the notion of cyborgs (see e.g. [56]), aspects that are perhaps less relevant for an intelligent improvising Disklavier piano. To conclude, the employed evaluation frameworks should not be considered "off the shelf" tools that can be readily applied to all settings. The methods should be adapted to different situations, and perhaps also be modified, to make sense for a specific musical context and stakeholder.

Despite the heterogeneous views on evaluation discussed above, there are some themes that re-occur across different research domains. For example, terms such as functionality, usability, and user experience were mentioned in literature from numerous fields. Several authors from various backgrounds also voiced the need for more holistic approaches that go beyond traditional HCI methods focused on standard usability metrics such as task completion rates. Moreover, several mentioned the need for qualitative methods to explore the complex nature of musical interactions and their situated nature, stressing the importance of focusing on emotions and social interactions, as well as cultural contexts. Other topics that reoccurred in the literature were attempts to more clearly define when to evaluate, since this might affect the choice of methods.

Considering the above, a reasonable suggestion would be to propose a workflow for evaluation of musical robots, rather than a set of evaluation metrics. Such a workflow could consist of the following steps:

1. *Classify the robot using existing taxonomies.* For example, is the robot a humanoid, is it physical or virtual, is it a wearable device, and how much AI and autonomous agency is involved? The purpose of this initial step is to place the musical robot in a context and to inform the subsequent steps. Taxonomies and strategies for classification described in Sections 12.1 and 12.2 can be used to help situate the robot in a historical context, and to better understand the breadth of previous work that has explored similar topics.

2. *Specify the context (and goal, if there is such a thing) of the musical interaction.* For example, is the musical robot intended to be used in a solo performance context or in ensemble play, is it performing in front of a large audience, is it composing music (without an audience), and so on. This step can also involve identifying high-level objectives driven by the representative use of the envisioned robot system (this might or might not involve UX goals, see [159]).

3. *Identify stakeholders.* This includes exploring what is the role of the human versus robot, as well as the level of co-creation and collaboration. It also involves situating the work in a socio-cultural context.

4. *Informed by steps 1–3, identify evaluation methods from the literature and adapt them, if necessary.* This involves identifying both objective and subjective measures, as well as formative and summative methods. Once a set of methods have been identified, they can be tuned to fit the specific musical setting, the socio-cultural context, and the stakeholders involved.

To conclude, it is worth noting that the review of evaluation methods presented in Section 12.8 is far from a full systematic review and that it should be expanded to include additional publication venues to be able to draw generalized conclusions about the entire field of musical robots. The review puts a strong emphasize on robotic musicianship in Computer Music and New

Interfaces for Musical Expression (NIME) research. Different tendencies would perhaps be identified if reviewing literature published in HRI journals. Finally, it is likely that the dataset to be reviewed would have become much larger if the search had been expanded to include the term "robot" in the main text (not only in the title and abstract).

Bibliography

[1] Agres, K., Forth, J., and Wiggins, G. A. Evaluation of Musical Creativity and Musical Metacreation Systems. *Computers in Entertainment (CIE) 14*, 3 (2016), 1–33.

[2] Amabile, T. *Componential Theory of Creativity*. Harvard Business School Boston, MA, 2011.

[3] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. Guidelines for human-AI interaction. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[4] Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2019), pp. 1078–1088.

[5] Baer, J. Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior 46*, 1 (2012), 16–29.

[6] Baer, J. The importance of domain-specific expertise in creativity. *Roeper Review 37*, 3 (2015), 165–178.

[7] Baginsky, N.A. The three Sirens: A self learning robotic rock band. http://www.baginsky.de/agl/

[8] Baraka, K., Alves-Oliveira, P., and Ribeiro, T. An extended framework for characterizing social robots. In *Human-Robot Interaction* (2020), Springer, pp. 21–64.

[9] Barbosa, J., Malloch, J., Wanderley, M. M., and Huot, S. What does "evaluation" mean for the NIME community? In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2015), pp. 156–161.

[10] Barkhuus, L., and Rode, J. A. From mice to men – 24 years of evaluation in CHI. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2007), pp. 1–16.

[11] Bates, J. The role of emotion in believable agents. *Communications of the ACM 37*, 7 (1994), 122–125.

[12] Birnbaum, D., Fiebrink, R., Malloch, J., and Wanderley, M. M. Towards a dimension space for musical devices. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2005), pp. 192–195.

[13] Blythe, M. A., Overbeeke, K., Monk, A. F., and Wright, P. C. *Funology: From Usability to Enjoyment*. Springer, 2004.

[14] Boden, M. A. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

[15] Boehner, K., DePaula, R., Dourish, P., and Sengers, P. How emotion is made and measured. *International Journal of Human-Computer Studies 65*, 4 (2007), 275–291.

[16] Buch, B., Coussement, P., and Schmidt, L. "Playing Robot": An interactive sound installation in human-robot interaction design for new media art. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2010), pp. 411–414.

[17] Buschek, D., Mecke, L., Lehmann, F., and Dang, H. Nine potential pitfalls when designing human-AI co-creative systems. *arXiv preprint arXiv:2104.00358* (2021).

[18] Campbell, D. M. Evaluating musical instruments. *Physics Today 67*, 4 (2014).

[19] Captured! by Robots. http://www.capturedbyrobots.com/.

[20] Chadefaux, D., Le Carrou, J.-L., Vitrani, M.-A., Billout, S., and Quartier, L. Harp plucking robotic finger. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (2012), IEEE, pp. 4886–4891.

[21] Chen, R., Dannenberg, R. B., Raj, B., and Singh, R. Artificial creative intelligence: Breaking the imitation barrier. In *Proceedings of International Conference on Computational Creativity* (2020), pp. 319–325.

[22] Collins, N. Trading faures: Virtual musicians and machine ethics. *Leonardo Music Journal 21* (2011), 35–39.

[23] Colton, S. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium: Creative Intelligent Systems* (2008), p. 7–14.

[24] Colton, S., Charnley, J. W., and Pease, A. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the International Conference on Innovative Computing and Cloud Computing* (2011), pp. 90–95.

[25] Colton, S., and Wiggins, G. A. Computational creativity: The final frontier? In *Proceedings of the European Conference on Artificial Intelligence (ECAI)* (2012), pp. 21–26.

[26] Connell, B. R., Jones, M., Mace, R., Mueller, J., Mullick, A., Ostroff, E., Sanford, J., Steinfeld, E., Story, M., and Vanderheiden, G. The Principles of Universal Design, 1997.

[27] Cook, A. M., and Polgar, J. M., Mosby (Elsevier), 2015. Assistive technologies: Principles and Practice (2002).

[28] Corbin, J., and Strauss, A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, 2014.

[29] Crypton Future Media. Hatsune Miku https://ec.crypton.co.jp/pages/prod/virtualsinger/cv01_us

[30] Dannenberg, R. B., Brown, H. B., and Lupish, R. McBlare: A robotic bagpipe player. In *Musical Robots and Interactive Multimodal Systems*. Springer, 2011, pp. 165–178.

[31] Dautenhahn, K. Some brief thoughts on the past and future of human-robot interaction, *Proceedings of the ACM Transactions on Human-Robot Interaction*, Vol. 7, No. 1, Article 4. 2018.

[32] Davanzo, N., and Avanzini, F. A dimension space for the evaluation of accessible digital musical instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2020), pp. 214–220.

[33] Davies, A., and Crosby, A. Compressorhead: the robot band and its transmedia storyworld. In *Proceedings of the International Workshop on Cultural Robotics* (2015), pp. 175–189.

[34] de Vaucanson, J. Le méchanisme du fluteur automate, 1738.

[35] Demers, L., Weiss-Lambrou, R., and Ska, B. The quebec user evaluation of satisfaction with assistive technology (QUEST 2.0): An overview and recent progress. *Technology and Disability 14*, 3 (2002), 101–105.

[36] Deterding, S., Hook, J., Fiebrink, R., Gillies, M., Gow, J., Akten, M., Smith, G., Liapis, A., and Compton, K. Mixed-initiative creative interfaces. In *Proceedings of the ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), pp. 628–635.

[37] Dewey, J. Art as experience. In *The Richness of Art Education*. Brill, 2008, pp. 33–48.

[38] Díaz-Oreiro, I., López, G., Quesada, L., and Guerrero, L. A. Standardized questionnaires for user experience evaluation: A systematic literature review. In *Proceedings of the International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI)* (2019), 14, p. 1–12.

[39] Dumas, J. S. User-based evaluations. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence Erlbaum, 2002, pp. 1093–1117.

[40] EIDD. The EIDD Stockholm Declaration. https://dfaeurope.eu/what-is-dfa/dfa-documents/the-eidd-stockholm-declaration-2004/, 2004.

[41] Eigenfeldt, A., Burnett, A., and Pasquier, P. Evaluating musical metacreation in a live performance context. In *Proceedings of the International Conference on Computational Creativity* (2012), pp. 140–144.

[42] El-Shimy, D., and Cooperstock, J. R. User-driven techniques for the design and evaluation of new musical interfaces. *Computer Music Journal 40*, 2 (2016), 35–46.

[43] Farmer, H. G. The Organ Of The Ancients From Eastern Sources (Hebrew, Syriac, And Arabic), *The New Temple Press*, 1931.

[44] Fernaeus, Y., Ljungblad, S., Jacobsson, M., and Taylor, A. Where third wave HCI meets HRI: Report from a workshop on user-centred design of robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2009), pp. 293–294.

[45] Forlizzi, J., and Battarbee, K. Understanding experience in interactive systems. In *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (2004), pp. 261–268.

[46] Forlizzi, J., and DiSalvo, C. Service robots in the domestic environment: A study of the Roomba vacuum in the home. In *Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (2006), pp. 258–265.

[47] Franklin, M., and Lagnado, D. Human-AI interaction paradigm for evaluating explainable artificial intelligence. In *Proceedings of the International Conference on Human-Computer Interaction* (2022), pp. 404–411.

[48] Frid, E. Accessible digital musical instruments – A review of musical interfaces in inclusive music practice. *Multimodal Technologies and Interaction 3*, 3 (2019), 57.

[49] Frid, E. *Diverse Sounds: Enabling Inclusive Sonic Interaction*. PhD thesis, KTH Royal Institute of Technology, 2019.

[50] Frid, E., Gomes, C., and Jin, Z. Music creation by example. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.

[51] Gatebox. INC https://www.gatebox.ai/.

[52] Gran, F. Cello Concerto No. 1. https://fredrikgran.com/works/cello-suite-no-1/.

[53] Greenberg, S., and Buxton, B. Usability evaluation considered harmful (some of the time). In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2008), pp. 111–120.

[54] Gunning, D. Explainable artificial intelligence (XAI). Tech. rep., Defense Advanced Research Projects Agency (DARPA), 2017.

[55] Gurevich, M. Distributed control in a mechatronic musical instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2014), pp. 487–490.

[56] Haraway, D. *A Cyborg Manifesto*. Socialist Review, 1985.

[57] Harrison, J. *Instruments and access: The role of instruments in music and disability*. PhD thesis, Queen Mary University of London, 2020.

[58] Hassenzahl, M. The thing and I: Understanding the relationship between user and product. In *Funology*. Springer, 2003, pp. 31–42.

[59] Hassenzahl, M., Law, E. L.-C., and Hvannberg, E. T. User Experience – Towards a unified view. In *Proceedings of the NordiCHI: COST294–MAUSE Workshop* (2006), pp. 1–3.

[60] Hatakeyama, K., Saraiji, M. Y., and Minamizawa, K. MusiArm: Extending prosthesis to musical expression. In *Proceedings of the Augmented Human International Conference* (2019), pp. 1–8.

[61] Hoffmann, L., Bock, N., and Roesenthal v.d. Pütten, A. M. The peculiarities of robot embodiment (EmCorp-Scale): Development, validation and initial test of the embodiment and corporeality of artificial

agents scale. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2018), pp. 370–378.

[62] Höök, K. User-centred design and evaluation of affective interfaces. In *From Brows to Trust*. Springer, 2004, pp. 127–160.

[63] Horvitz, E. Principles of mixed-initiative user interfaces. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (1999), pp. 159–166.

[64] IBM. IBM Design for AI. https://www.ibm.com/design/ai/, 2019.

[65] International Organization for Standardization. ISO 9241-11:2018 - Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. https://www.iso.org/standard/63500.html.

[66] International Organization for Standardization. ISO 8373:2021 Robotics – Vocabulary. https://www.iso.org/standard/75539.html, 2021.

[67] Isbister, K., Höök, K., Sharp, M., and Laaksolahti, J. The sensual evaluation instrument: Developing an affective evaluation tool. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2006), pp. 1163–1172.

[68] Ivory, M. Y., and Hearst, M. A. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR) 33*, 4 (2001), 470–516.

[69] Jack, R., Harrison, J., and McPherson, A. Digital musical instruments as research products. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2020), pp. 446–451.

[70] Jiang, S., and Arkin, R. C. Mixed-initiative human-robot interaction: Definition, taxonomy, and survey. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (2015), pp. 954–961.

[71] Johnston, A. J. Beyond evaluation: Linking practice and theory in new musical interface design. In *Proceedings of the International New Interfaces for Musical Expression Conference (NIME)* (2011).

[72] Jones, R. Archaic man meets a marvellous automaton: Posthumanism, social robots, archetypes. *Journal of Analytical Psychology 62*, 3 (2017), 338–355.

[73] Jordà, S. Afasia: The ultimate homeric one-man-multimedia-band. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2002), pp. 132–137.

[74] Jordà, S. Digital instruments and players: Part II-Diversity, freedom, and control. In *Proceedings of the International Computer Music Conference (ICMC)* (2004).

[75] Jordanous, A. *Evaluating Computational Creativity: a Standardised Procedure for Evaluating Creative Systems and its Application*. University of Kent (United Kingdom), 2012.

[76] Jordanous, A. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation 4*, 3 (2012), 246–279.

[77] Jordanous, A. Evaluating evaluation: Assessing progress and practices in computational creativity research. In *Computational Creativity*. Springer, 2019, pp. 211–236.

[78] Jost, C., Le Pévédic, B., Belpaeme, T., Bethel, C., Chrysostomou, D., Crook, N., Grandgeorge, M., and Mirnig, N., Eds. *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer, Germany, 2020.

[79] Kajitani, M. Development of musician robots. *Journal of Robotics and Mechatronics 1*, 1 (1989), 254–255.

[80] Kajitani, M. Simulation of musical performances. *Journal of Robotics and Mechatronics 4*, 6 (1992), 462–465.

[81] Kapur, A. A history of robotic musical instruments. In *Proceedings of the International Computer Music Conference (ICMC)* (2005), p. 4599.

[82] Karimi, P., Grace, K., Maher, M. L., and Davis, N. Evaluating creativity in computational co-creative systems. *arXiv preprint arXiv:1807.09886* (2018).

[83] Kaye, J. J. Evaluating experience–focused HCI. In *Proceedings of the ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems* (2007), pp. 1661–1664.

[84] Kemper, S. Locating creativity in differing approaches to musical robotics. *Frontiers in Robotics and AI 8* (2021), 647028.

[85] Kieras, D. Model-based evaluation. In *Human-Computer Interaction*. CRC Press, 2009, pp. 309–326.

[86] Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition 26*, 2 (2008), 169–181.

[87] Koetsier, T. On the prehistory of programmable machines: Musical automata, looms, calculators. *Mechanism and Machine Theory 36*, 5 (2001), 589–603.

[88] Krzyzaniak, M. Prehistory of musical robots. *Journal of Human-Robot Interaction 1*, 1 (2012), 78–95.

[89] Kvifte, T., and Jensenius, A. R. Towards a coherent terminology and model of instrument description and design. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2006), pp. 220–225.

[90] Lamb, C., Brown, D. G., and Clarke, C. L. A. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR) 51*, 2 (2018), 1–34.

[91] Langley, P., Meadows, B., Sridharan, M., and Choi, D. Explainable agency for intelligent autonomous systems. In *Proceedings of the Association for the Advancement of Artificial Intelligence* (2017).

[92] Laugwitz, B., Held, T., and Schrepp, M. Construction and evaluation of a user experience questionnaire. In *Proceedings of the Symposium of the Austrian HCI and Usability Engineering Group* (2008), pp. 63–76.

[93] Law, E., Roto, V., Vermeeren, A. P., Kort, J., and Hassenzahl, M. Towards a shared definition of user experience. In *Proceedings of the ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems* (2008), pp. 2395–2398.

[94] Lewis, C., and Wharton, C. Cognitive walkthroughs. In *Handbook of Human–Computer Interaction*. Elsevier, 1997, pp. 717–732.

[95] Li, T., Vorvoreanu, M., DeBellis, D., and Amershi, S. Assessing human-AI interaction early through factorial surveys: A study on the guidelines for human-AI interaction. *ACM Transactions on Computer-Human Interaction* (2022).

[96] Lindblom, J., Alenljung, B., and Billing, E. Evaluating the user experience of human–robot interaction. In *Human–Robot Interaction*. Springer Nature Switzerland, 2020, pp. 231–256.

[97] Long, J., Murphy, J., Carnegie, D., and Kapur, A. Loudspeakers optional: A history of non-loudspeaker-based electroacoustic music. *Organised Sound 22*, 2 (2017), 195–205.

[98] Loughran, R., and O'Neill, M. Generative music evaluation: Why do we limit to 'human'. In *Proceedings of the Conference on Computer Simulation of Musical Creativity (CSMC)* (2016).

[99] Lucas, A., Harrison, J., Schroeder, F., and Ortiz, M. Cross-pollinating ecological perspectives in ADMI design and evaluation. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2021).

[100] MacDonald, C. M., and Atwood, M. E. Changing perspectives on evaluation in HCI: Past, present, and future. In *Proceedings of the ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems* (2013), pp. 1969–1978.

[101] Magnusson, T. An epistemic dimension space for musical devices. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2010), pp. 43–46.

[102] McCarthy, J., and Wright, P. Technology as experience. *Interactions 11*, 5 (2004), 42–43.

[103] McLeod, K. Living in the immaterial world: Holograms and spirituality in recent popular music. *Popular Music and Society 39*, 5 (2016), 501–515.

[104] McNamara, N., and Kirakowski, J. Defining usability: Quality of use or quality of experience? In *Proceedings of the International Professional Communication Conference (IPCC)* (2005), pp. 200–204.

[105] McNamara, N., and Kirakowski, J. Functionality, usability, and user experience: Three areas of concern. *Interactions 13*, 6 (2006), 26–28.

[106] Minsky, M. L. Why people think computers can't. *AI Magazine 3*, 4 (1982), 3–3.

[107] Murphy, J., Kapur, A., and Carnegie, D. Musical robotics in a loudspeaker world: Developments in alternative approaches to localization and spatialization. *Leonardo Music Journal 22* (12 2012), 41–48.

[108] Nake, F. Construction and intuition: Creativity in early computer art. In *Computers and Creativity*. Springer, 2012, pp. 61–94.

[109] Ness, S. R., Trail, S., Driessen, P. F., Schloss, W. A., and Tzanetakis, G. Music information robotics: Coping strategies for musically challenged robots. In *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)* (2011), pp. 567–572.

[110] Nielsen, J. Usability inspection methods. In *Conference Companion on Human Factors in Computing systems* (1994), pp. 413–414.

[111] Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1990), pp. 249–256.

[112] Norman, D. A. *The Design of Everyday Things*. Basic Books, 1988.

[113] Norman, D. A. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic, 2004.

[114] O'Modhrain, S. A framework for the evaluation of digital musical instruments. *Computer Music Journal 35*, 1 (2011), 28–42.

[115] Onnasch, L., and Roesler, E. A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics 13*, 4 (2021), 833–849.

[116] Google PAIR, G. People + AI Guidebook. https://pair.withgoogle.com/guidebook/, 2019.

[117] Pawel Zadrozniak. Floppy music DUO – Imperial March. https://youtu.be/yHJOz_y9rZE. See also http://silent.org.pl/home/.

[118] Pease, A., and Colton, S. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the AISB Symposium on AI and Philosophy* (2011), pp. 15–22.

[119] Petrie, H., and Bevan, N. The evaluation of accessibility, usability, and user experience. In *The Universal Access Handbook* (2009), 20:1–16. See https://www.routledge.com/The-Universal-Access-Handbook/Stephanidis/p/book/9780805862805.

[120] Reimer, P. C., and Wanderley, M. M. Embracing less common evaluation strategies for studying user experience in NIME. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2021).

[121] Remy, C., MacDonald Vermeulen, L., Frich, J., Biskjaer, M. M., and Dalsgaard, P. Evaluating creativity support tools in HCI research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), pp. 457–476.

[122] Rezwana, J., and Maher, M. L. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction* (2022).

[123] Rhodes, M. An analysis of creativity. *The Phi Delta Kappan 42*, 7 (1961), 305–310.

[124] Ritchie, G. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines 17*, 1 (2007), 67–99.

[125] Roads, C. The Tsukuba musical robot. *Computer Music Journal 10*, 2 (1986), 39–43.

[126] Robert, J.-M., and Lesage, A. From usability to user experience with interactive systems. In *The Handbook of Human-Machine Interaction*. CRC Press, 2017, pp. 303–320.

[127] Rodger, M., Stapleton, P., Van Walstijn, M., Ortiz, M., and Pardue, L. What makes a good musical instrument? A matter of

processes, ecologies and specificities. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2020), pp. 405–410.

[128] Rogers, T., Kemper, S., and Barton, S. MARIE: Monochord - aerophone robotic instrument ensemble. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2015), pp. 408–411.

[129] Rowe, R. *Machine Musicianship*. MIT Press, 2004.

[130] Savery, R., and Weinberg, G. Shimon the robot film composer and DeepScore. In *Computer Simulation of Musical Creativity (2018)* (2018), pp. 1–14.

[131] Savery, R., and Weinberg, G. Robotics - Fast and curious: A CNN for ethical deep learning musical generation. In *Artificial Intelligence and Music Ecosystem*. Focal Press, 2022, pp. 52–67.

[132] Savery, R., Zahray, L., and Weinberg, G. Shimon the Rapper: A real-time system for human-robot interactive rap battles. In *International Conference on Computational Creativity* (2020), pp. 212–219.

[133] Savery, R., Zahray, L., and Weinberg, G. Shimon sings – Robotic musicianship finds its voice. In *Handbook of Artificial Intelligence for Music*. Springer, Cham, 2021, pp. 823–847.

[134] Scherer, M. J. Matching person & technology: Model and assessment process, 2007.

[135] Schmid, G.-M. *Evaluating the Experiential Quality of Musical Instruments*. Springer, 2017.

[136] Scriven, M. Beyond formative and summative evaluation. In *Evaluation and Education: At Quarter Century*, M. W. M. D. C. Phillips, Ed. University of Chicago Press, 1991, pp. 18–64.

[137] Sengers, P., and Gaver, B. Staying open to interpretation: Engaging multiple meanings in design and evaluation. In *Proceedings of the Conference on Designing Interactive Systems* (2006), pp. 99–108.

[138] Singer, E., Feddersen, J., Redmon, C., and Bowen, B. LEMUR's musical robots. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2004), pp. 181–184.

[139] Singer, E., Larke, K., and Bianciardi, D. LEMUR GuitarBot: MIDI robotic string instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2003), pp. 188–191.

[140] Small, C. *Musicking: The Meanings of Performing and Listening.* Wesleyan University Press, 1998.

[141] Snake-Beings, E. The Do-it-Yourself (DiY) craft aesthetic of The Trons – Robot garage band. *Craft Research 8*, 1 (2017), 55–77.

[142] Solis, J., Bergamasco, M. Isoda, S., Chida, K., and Takanishi, A. "Learning to Play the Flute with an Anthropomorphic Robot", Proceedings of the International Computer Music Conference, Miami, Florida, 2004.

[143] Solis, J., Chida, K., Taniguchi, K., Hashimoto, S. M., Suefuji, K., and Takanishi, A. The Waseda flutist robot WF-4RII in comparison with a professional flutist. *Computer Music Journal* (2006), 12–27.

[144] Solis, J., and Ng, K., Eds. *Musical Robots and Interactive Multimodal Systems.* Springer, 2011.

[145] Solis, J., and Ng, K. Musical robots and interactive multimodal systems: An introduction. In *Musical Robots and Interactive Multimodal Systems.* Springer, 2011, pp. 1–12.

[146] Solis, J., and Takanishi, A. An overview of the research approaches on musical performance robots. In *Proceedings of the International Conference on Computer Music (ICMC)* (2007), pp. 356–359.

[147] Solis, J., and Takanishi, A. Wind instrument playing humanoid robots. In *Musical Robots and Interactive Multimodal Systems.* Springer, 2011, pp. 195–213.

[148] Solis, J., Takanishi, A., and Hashimoto, K. Development of an anthropomorphic saxophone-playing robot. In *Brain, Body and Machine.* Springer, 2010, pp. 175–186.

[149] Stephens, E., and Heffernan, T. We have always been robots: The history of robots and art. In *Robots and Art.* Springer, 2016, pp. 29–45.

[150] Stowell, D., Robertson, A., Bryan-Kinns, N., and Plumbley, M. D. Evaluation of live human–computer music-making: Quantitative and qualitative approaches. *International Journal of Human-Computer Studies 67*, 11 (2009), 960–975.

[151] Striebe, D. The Prayer. https://theprayer.diemutstrebe.com/.

[152] Sung, J.-Y., Guo, L., Grinter, R. E., and Christensen, H. I. "My Roomba is Rambo": Intimate home appliances. In *Proceedings of the International Conference on Ubiquitous Computing* (2007), pp. 145–162.

[153] Szollosy, M. Freud, Frankenstein and our fear of robots: Projection in our cultural perception of technology. *AI & Society 32*, 3 (2017), 433–439.

[154] Takagi, S. Toyota partner robots. *Journal of the Robotics Society of Japan 24*, 2 (2006), 208–210.

[155] Tatar, K., and Pasquier, P. Musical agents: A typology and state of the art towards musical metacreation. *Journal of New Music Research 48*, 1 (2019), 56–105.

[156] Team UEQ. User Experience Questionnaire. https://www.ueq-online.org/.

[157] Trimpin. Wikipedia. https://en.wikipedia.org/wiki/Trimpin.

[158] Yamaha Corporation. Vocaloid. https://www.vocaloid.com/en/.

[159] Wallström, J., and Lindblom, J. Design and development of the USUS goals evaluation framework. In *Human-Robot Interaction* (2020), Springer, pp. 177–201.

[160] Wanderley, M. M., and Orio, N. Evaluation of input devices for musical expression: Borrowing tools from HCI. *Computer Music Journal 26*, 3 (2002), 62–76.

[161] Waters, S. Performance ecosystems: Ecological approaches to musical interaction. *EMS: Electroacoustic Music Studies Network* (2007), 1–20.

[162] Weinberg, G., Bretan, M., Hoffman, G., and Driscoll, S. Introduction. In *Robotic Musicianship: Embodied Artificial Creativity and Mechatronic Musical Expression*, vol. 8. Springer Nature, 2020, pp. 1–24.

[163] Weinberg, G., Bretan, M., Hoffman, G., and Driscoll, S. *Robotic Musicianship: Embodied Artificial Creativity and Mechatronic Musical Expression*, vol. 8. Springer Nature, 2020.

[164] Weinberg, G., Bretan, M., Hoffman, G., and Driscoll, S. "Wear it"—Wearable robotic musicians. In *Robotic Musicianship*. Springer, 2020, pp. 213–254.

[165] Weinberg, G., and Driscoll, S. Toward robotic musicianship. *Computer Music Journal* (2006), 28–45.

[166] Weiss, A., Bernhaupt, R., Lankes, M., and Tscheligi, M. The USUS evaluation framework for human-robot interaction. In *Proceedings of the Symposium on New Frontiers in Human-Robot Interaction* (2009), pp. 11–26.

[167] Williamson, M. M. *Robot arm control exploiting natural dynamics*. PhD thesis, Massachusetts Institute of Technology, 1999.

[168] Wood, G. *Edison's Eve: A Magical History of the Quest for Mechanical Life*. Alfred A. Knopf, 2002.

[169] Yanco, H. A., and Drury, J. Classifying Human-Robot Interaction: An updated taxonomy. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (2004), vol. 3, pp. 2841–2846.

[170] Yanco, H. A., and Drury, J. L. A taxonomy for human-robot interaction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Human-Robot Interaction* (2002), pp. 111–119.

[171] Yang, N., Savery, R., Sankaranarayanan, R., Zahray, L., and Weinberg, G. Mechatronics-driven musical expressivity for robotic percussionists. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (2020).

[172] Yannakakis, G. N., Liapis, A., and Alexopoulos, C. Mixed-initiative co-creativity. In *Proceedings of the International Conference on Foundations of Digital Games* (2014), pp. 1–8.

[173] Young, G. W., and Murphy, D. HCI models for digital musical instruments: Methodologies for rigorous testing of digital mmsical instruments. *arXiv preprint arXiv:2010.01328* (2020).

[174] Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H. I., and Grinter, R. E. Evaluating human-robot interaction. *International Journal of Social Robotics 3*, 1 (2011), 53–67.

[175] Zimoun. Zimoun selected works 4.2 (video). https://www.zimoun.net/ and https://vimeo.com/7235817.