Design Principles for Generative AI Applications

JUSTIN D. WEISZ, IBM Research AI, USA
JESSICA HE, IBM Research AI, USA
MICHAEL MULLER, IBM Research AI, USA
GABRIELA HOEFER, IBM, USA
RACHEL MILES, IBM, USA
WERNER GEYER, IBM Research AI, USA

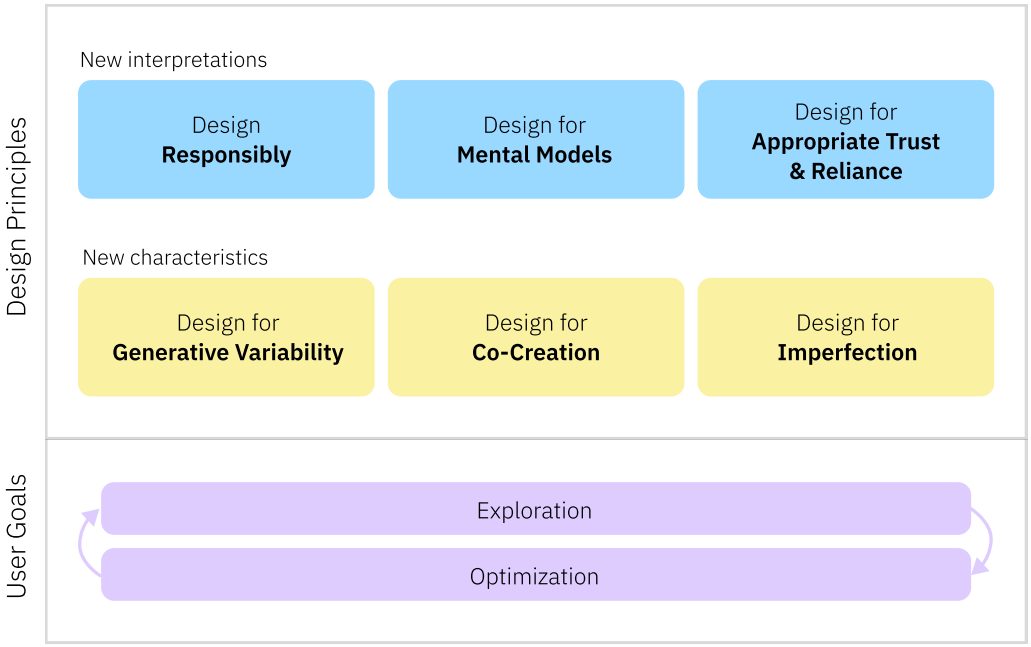

Fig. 1. Six principles for the design of generative AI applications. Three principles offer new interpretations of known issues with AI systems through the lens of generative AI, and three principles identify unique characteristics of generative AI systems. The principles support two user goals: optimizing a generated artifact to satisfy task-specific criteria, and exploring different possibilities within a domain.

Generative AI applications present unique design challenges. As generative AI technologies are increasingly being incorporated into mainstream applications, there is an urgent need for guidance on how to design user experiences that foster effective and safe use. We present six principles for the design of generative AI applications that address unique characteristics of generative AI UX and offer new interpretations and extensions of known issues in the design of AI applications. Each principle is coupled with a set of design strategies for implementing that principle via UX capabilities or through the design process. The principles and strategies were developed through an iterative process involving literature review, feedback from design practitioners, validation against real-world generative AI applications, and incorporation into the design process of two generative AI applications. We anticipate the principles to usefully inform the design of generative AI applications by driving actionable design recommendations.

2024. Manuscript submitted to ACM

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; *Interaction paradigms*; *HCI theory, concepts and models.*

Additional Key Words and Phrases: Generative AI, design principles, human-centered AI, foundation models

ACM Reference Format:
Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *CHI '24: ACM CHI conference on Human Factors in Computing Systems, May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Generative AI technologies have reached an inflection point in consumer adoption and enterprise value, sparked by technological advancements in machine learning architectures such as GANs [56, 79], VAEs [86], and transformers [38, 170]. Models such as StyleGAN [79], GPT [20, 130, 137, 138], and Codex [28] have demonstrated that powerful generative models can produce works at a human-like level of fidelity. Today, consumer applications such as ChatGPT[1], DreamStudio[2], and DALL-E[3] are making these technologies widely available and setting the bar for people's expectations of what generative AI can do. Startups such as Cohere[4] and Anthropic[5] are reducing the friction of embedding large language models in consumer applications. Enterprises such as IBM, Microsoft, Amazon, and Google are creating platforms for businesses to infuse generative technologies into their products and services.

This commercialization of generative AI technologies is fueled by the ultra-rapid development of large-scale foundation models [19] that reduce the time and costs for developing generative AI systems. However, much attention in machine learning research communities has focused on developing advancements to the *technology*: scaling model parameter counts [91, 158], evaluating model performance [97, 163, 194], tuning models efficiently to perform new tasks [27, 178], and aligning models [132, 196] to reduce their propensity to produce speech that is hateful, abusive, profane, or otherwise toxic [60, 70]. Although these advancements serve to improve the state of the art, they do not recognize an important half of what Ehsan et al. [43] call the "human-AI assemblage" – the *human*.

Generative models have enabled a radically new way for people to interact with computing technologies. People are now able to craft specifications for the kinds of outputs they desire, such as via natural language prompts, and generative models are able to produce outputs that conform to those specifications. Nielsen [127] recently identified this form of interaction as *intent-based outcome specification* and argued that it is the first new UI interaction paradigm in 60 years. This form of interaction is fundamentally different from previous interaction paradigms (e.g. punchcards, command line interfaces, and graphical user interfaces), because it shifts control over how computation is performed away from the user and toward generative AI models. With this shift in control, how are we to design user experiences that help people interact with generative AI applications in effective and safe ways?

Over at least the past four decades, researchers and practitioners within human-computer interaction (HCI) have produced numerous guidelines, principles, practices, and frameworks for the design of effective and safe computing systems. Some guidelines are presented as generally applicable to most kinds of interactive computing systems, such as Nielsen and Molich's heuristics [128] and Shneiderman et al.'s strategies for designing effective human-computer interaction [155]. Other design guidelines are technology-specific, such as Bevan's guidelines for web usability [16] and

[1]ChatGPT. https://chat.openai.com
[2]DreamStudio. https://beta.dreamstudio.ai
[3]DALL-E. https://labs.openai.com
[4]Cohere. http://cohere.ai
[5]Anthropic. http://anthropic.com

various guidelines for AI systems and human-AI interaction [7, 66, 133]. However, none of the guidelines developed within the HCI community have yet addressed the specific nuances present with generative AI.

In this paper, we build upon the HCI tradition of distilling cumulative knowledge into a practical set of principles, specifically for the design of generative AI user experiences (UX). We opted to produce a set of *principles*, rather than *guidelines*, to highlight their fundamental nature to the design of generative AI UX.

Our paper makes the following contributions to the CHI community:

- We **introduce** a set of six principles for the design of generative AI applications (Table 1). Three principles identify new considerations for generative AI, and three offer new interpretations of known issues with AI systems. Each principle is coupled with a set of practical strategies and examples (Appendix A) for how to put the principle into practice – either through the design process itself or through specific UX capabilities – in order to aid design practitioners[6] in applying the principles to real-world UX.
- We establish the **practical value** of the design principles by conducting a rigorous and systematic validation that included multiple rounds of testing, feedback collection, iteration, and use by two generative AI application design teams.
- We **provide insight** on how we made decisions regarding the organization of the principles, navigated issues of conceptual redundancy and overlap, and fostered adoption of the principles within our organization.

2 RELATED WORK

Over the past four decades, the HCI community has conducted numerous examinations of user interface design for computing technologies and distilled best practices into various forms of design guidelines. We identify and review two categories of these guidelines: those that apply to the general UX design of computing systems and those that apply specifically to AI-infused systems.

2.1 Guidelines for human-computer interaction and user interface design

Design guidelines have a rich history in HCI research. We provide a brief historical sketch of human-computer interaction design guidelines and their evolution as a result of new paradigms in user interface technologies.

Licklider and Clark provide an early example of guidelines for the design of console interfaces in, "On-line man-computer communication" [99] (reflecting the gendered language of their time). Although their guidelines are not presented in a modernly-recognizable form[7], their work does introduce ideas about humans' needs when working with computers. Other early forms of human-computer interaction guidelines include those for the Spacelab Experiment Computer Application Software (ECAS) [39], and for the command terminals of battlefield automated systems [156, 157]. Even in these early days of user interface design, it was recognized that, "[l]acking consistent design principles, current practice results in a fragmented and unsystematic approach to system design, especially where the user/operator-system interaction is concerned." [156, p. v].

A significant inflection point in the design of user interfaces for computing technologies came with the rise of personal computing and the concurrent emergence of graphical user interfaces (GUIs) as a new interaction paradigm. New guidelines were needed for the new visual metaphors of windows, icons, menus, and pointers. Apple's Human

[6]We consider a design practitioner to be anyone involved in making deliberate decisions that impact the design of an application, including user researchers, interaction designers, visual designers, product managers, and more.
[7]For example, they provide guidelines on allocating tasks between humans and computers that "exploit the complementation that exists between human capabilities and present computer capabilities" [99, p.114].

Interface Guidelines [8] provides a prominent example of practical guidelines for GUI design. Smith and Mosier [159] developed more formal guidelines for GUIs in which they described six functional areas: data entry, data display, sequence control, user guidance, data transmission, and data protection.

As Grudin described in an influential retrospective analysis [58], the "site" of human-computer interaction began to move away from a terminal in a lab and "reached out" into other contexts such as home and office environments. New methods were required to understand and design for these changing circumstances. With heuristic evaluation [128], Nielsen provided a set of methods for "discount usability engineering" [126] that helped designers more easily assess their interfaces, while Lewis and Wharton's cognitive walkthrough method [93] provided a way for designers to conduct more detailed and tailored analysis.

The next technological inflection point that necessitated a shift in design guidelines occurred with the rise of the Web. Unlike the previous decades, web design involved diverse and competing hardware and software. These complexities led to what Mariage et al. describe as a "jungle of guidelines [intended to] address many different issues" [113, Introduction]. One result was that, out of 11 generic web design guidelines, Cappel and Huang found that a mean of only 5.5 guidelines were followed across 500 companies' websites [25]. Adding to the diversity, technologically-literate advocates emerged for people with disabilities [92, 141], older users [14, 90], and users from diverse cultures [2]. Bevan [16] acknowledged that the available wealth of guidelines addressed different issues in different ways for different constituencies, and that this situation was likely to continue.

The advent and widespread commercial adoption of smartphones, mobile apps, app stores, and mobile web sites necessitated yet another new design language, optimized for smaller screens and touch-based interactions. Some work in this space focused on developing design frameworks and guidelines for mobile apps and workflows (e.g., [61, 78, 109, 124, 136, 151]). Guidelines also emerged covering mobile design for more specialized populations, including older users [5, 30], users of courseware [71], users from diverse cultures [5], users with disabilities [134], and users with diverse literacies [162]. Guidelines were also developed for specific mobile app domains such as health care [4, 75] and finance [3, 64, 118], as well as ethical concerns around privacy [94, 96] and the use of mobile apps for research purposes [115, 142].

2.2 Guidelines for human-AI interaction

Within the past few years, the emergence of AI as a design material [41, 46, 62, 191] has necessitated guidelines that inform its use. A growing body of work within the human-centered AI research community has proposed best practices for human-AI interaction in the form of design guidelines (e.g., [7, 11, 104, 186, 192]), formal studies (e.g., [24, 102]), toolkits (e.g., [110]), and reviews (e.g., [59, 72, 180, 188]).

Some of these guidelines include claims of universal applicability[8] or being of a general nature to AI-infused systems (e.g., [7, 153]). Other guidelines focus on specific types of AI technologies (e.g. text-to-image models [104]), specific domains of use (e.g. creative writing [24]), or specific issues regarding the use of AI, including ethics [11, 59, 69, 72], fairness [110], human rights [50], explainability [119], and user trust [186]. Finally, as more consumer products incorporate AI technologies, industry leaders including Google [133], Microsoft [7, 95] and Apple [9] have developed and published their own guidelines; Wright et al. [188] provide a comparative analysis of these guidelines.

Guidelines that focus on the design of AI systems, and specifically on the ethics of those systems, are critically important. Various attempts have been made to assist design practitioners in the process of operationalizing guidelines

[8]Multiple scholars have critiqued the concept of "universal" design as privileging certain assumed "normal" populations [12, 168].

for AI systems, including guidebooks [133], toolkits [36], and checklists [110]. When design guidelines are successfully applied, they make a positive impact, such as in assisting cross-functional development teams in improving user experiences [95] and addressing ethical challenges [11]. However, several studies have critiqued their comprehensiveness [59], the extent to which they can be operationalized [50], and the lack of consequences when they are not followed [59]. Additionally, Madaio et al. [110] argues that the adoption of an AI ethics process within an organization, "would only happen if leadership changed organizational culture to make AI fairness a priority, similar to priorities and associated organizational changes made by leadership to support security, accessibility, and privacy" [110, p. 8].

Despite the preponderance of guidelines for human-AI interaction and AI ethics, there is a gap in the technologies on which they focus. To date, many of the AI guidelines developed within the HCI community primarily focus on discriminative AI [7, 65, 66, 133], the class of algorithms that identifies boundaries that separate different classes or groups in a data set. These guidelines do not take into account generative AI algorithms that produce artifacts, rather than decision boundaries, as outputs. Because generative AI offers new ways for users to interact with technology, and raises new issues regarding the ethics of AI systems, a new set of design guidelines are needed.

3 WHY GENERATIVE AI NEEDS DESIGN PRINCIPLES

Generative AI technologies have introduced a new paradigm of human-computer interaction, what Nielsen refers to as "intent-based outcome specification" [127]. In this paradigm, users specify *what they want*, often using natural language[9], but not *how it should be produced*. One challenge of this paradigm stems from the distinguishing characteristic of generative AI: it generates artifacts as outputs and those outputs may vary in character or quality, even when a user's input does not change. This characteristic has been described by Weisz et al. [182] as *generative variability*, and it provides what Alvarado and Waern [6] describe as an "algorithmic experience," raising questions on appropriate types of user control, levels of algorithmic transparency, and user awareness of how the algorithms work and how to effectively interact with them.

With generative AI applications, users will need to develop a new set of skills to work with (not against) generative variability by learning how to create specifications that result in artifacts that match their desired intent. One emerging skill revolves around crafting effective natural language prompts, known as in-context learning [20, 40, 189] or prompt engineering [185, 193]. This process is typically informal and relies on trial-and-error [104, 131, 165, 190]. The use of open-ended natural language, rather than a fixed vocabulary of commands, leads to new design challenges. For example, Nielsen argues, "users should not have to wonder whether different words, situations, or actions mean the same thing" [125, p.156]; given the innumerable ways that users can express their intent in a natural language prompt, how can generative AI applications help users achieve desired results? Is it necessarily a "mistake" or "error" when a user's prompt results in an output that they didn't anticipate or like? Does it violate the consistency heuristic when it is difficult for users to achieve replicable results (e.g. [116, 135, 150]), because each click of the "generate" button results in different outputs, even for the same input?

Existing human-AI design guidelines fail to address the unique design challenges of generative AI because they do not cover generative use cases or new considerations stemming from generative variability, and they do not cover new or amplified ethical issues stemming from the models' generative nature. For example, guidelines published by PAIR make recommendations such as, "Design for labelers & labeling" and "Design & evaluate the reward function" [133]. The former of these recommendations will not apply to generative use cases that do not require data labeling, as the

[9]Some user interfaces to generative AI systems allow users to specify their intent via sketches and gestures [32], UI controls [103, 105], video-scanned body movements [174], and even various forms of improv performance [15, 68].

foundation models often used to implement generative capabilities are pre-trained and may not require additional labeled data for tuning. In addition, the latter recommendation is tailored to classification use cases in which false positives and false negatives are important outcome metrics; in generative use cases, these metrics have no meaning.

Guidelines from Amershi et al. [7] may be more readily adapted to generative AI applications, although their coverage of generative-specific considerations is limited and design practitioners may encounter difficulties in making such adaptations. For example, the recommendation to, "Make clear why the system did what it did" is potentially less important when a user's goal is to simply generate a desirable artifact[10]. The recommendation to, "Make clear what the system can do" may be difficult to implement in light of the emergent and unanticipated behaviors of generative foundation models [19], as well as the trial-and-error methods by which users iterate toward a desired outcome [104, 131, 165, 190].

Finally, alongside their tremendous potential to augment people's creative capabilities, generative technologies also introduce new risks and potential user harms. These risks include issues of copyright and intellectual property [47, 107], the circumvention or reverse-engineering of prompts through attacks [35], the production of hateful, toxic, or profane language [60], the disclosure of sensitive or personal information [83], the production of malicious source code [26, 28], and a lack of representation of minority groups due to underrepresentation in the training data [51, 108, 167, 171]. Work by Houde et al. [63] takes concerns such as these to an extreme by envisioning realistic, malicious uses of generative AI technologies. Although it cannot be a designer's responsibility to curb all potentially-harmful usage, existing design guidelines for AI systems fall short in addressing these unique issues stemming from the *generative* nature of generative AI, and AI ethics frameworks are only just starting to appear to provide designers with the language they need to begin discussing these important issues [36, 66, 180].

We therefore conclude that there is a pressing need for a set of general design guidelines that help practitioners develop applications that utilize generative AI technologies in safer and more effective ways – safer because of the new risks introduced by generative AI, and more effective because of the control that users have lost over the computational process. Although recent work has begun to probe at design considerations for generative AI, this work has been limited to specific application domains or technologies. For example, guidelines of various maturity levels exist for GAN-based interfaces [57, 195], image creation [102, 104, 173], prompt engineering [102, 104], virtual reality [169], collaborative storytelling [146], and workflows with co-creative systems [57, 123, 173]. Our work seeks to extend these studies toward principles that can be used *across* generative AI domains and technologies.

4 DESIGN PRINCIPLES FOR GENERATIVE AI APPLICATIONS

We begin by presenting our final set of six design principles and their corresponding strategies in Table 1, along with our overall design framework in Figure 1. We also provide extended descriptions and examples of each principle and strategy in Appendix A. In the rest of this paper, we describe the process we used to develop and validate these principles and strategies.

The principles are generally presented as high-level "design for..." statements that indicate the characteristics that are important to consider when making design decisions. Three principles focus on aspects of existing AI systems that have new interpretations through the lens of generative AI: **Design Responsibly**, **Design for Mental Models**, and

[10]Research by Sun et al. [166] explores the kinds of questions that users have when working with a generative AI system, which include questions about how an artifact was produced. We posit that the utility of a generated artifact need not depend upon the mechanics of *how* that artifact was generated in the same way that a user's trust in a decision recommendation is often predicated on an explanation for how that recommendation was produced (e.g. [10, 98, 172]). Further, some applications of generative AI concern the exploration of a space of multiple possibilities (e.g. [88, 144]), indicating that in some use cases, *how* an artifact was generated may be of lesser importance than the generated artifacts themselves.

Design Responsibly
Ensure the AI system solves real user issues and minimizes user harms
- *Use a human-centered approach**. Design for the user by understanding their needs and pain points, and not for the technology or its capabilities.
- *Identify & resolve value tensions**. Consider and balance different values across people involved in the creation, adoption, and usage of the AI system.
- *Expose or limit emergent behaviors*. Determine whether generative capabilities beyond the intended use case should be surfaced to the user or restricted.
- *Test & monitor for user harms**. Identify relevant user harms (e.g. bias, toxic content, misinformation) and include mechanisms that test and monitor for them.

Design for Generative Variability
Help the user manage the ability of generative models to produce multiple outputs that are distinct and varied
- *Leverage multiple outputs*. Generate multiple outputs that are either hidden or visible to the user in order to increase the chance of producing one that fits their need.
- *Visualize the user's journey*. Show the user the outputs they have created and guide them to new output possibilities.
- *Enable curation & annotation*. Design user-driven or automated mechanisms for organizing, labeling, filtering, and/or sorting outputs.
- *Draw attention to differences or variations across outputs*. Help the user identify how outputs generated from the same prompt differ from each other.

Design for Mental Models
Communicate how to work effectively with the AI system, considering the user's background and goals
- *Orient the user to generative variability*. Help the user understand the AI system's behavior and that it may produce multiple, varied outputs for the same input.
- *Teach effective use*. Help the user learn how to effectively use the AI system by providing explanations of features and examples through in-context mechanisms and documentation.
- *Understand the user's mental model**. Build upon the user's existing mental models and evaluate how they think about your application: its capabilities, limitations, and how to work with it effectively.
- *Teach the AI system about the user*. Capture the user's expectations, behaviors, and preferences to improve the AI system's interactions with them.

Design for Co-Creation
Enable the user to influence the generative process and work collaboratively with the AI system
- *Help the user craft effective outcome specifications*. Assist the user in prompting effectively to produce outputs that fit their needs.
- *Provide generic input parameters*. Let the user control generic aspects of the generative process such as the number of outputs and the random seed used to produce those outputs.
- *Provide controls relevant to the use case & technology*. Let the user control parameters specific to their use case, domain, or the generative AI's model architecture.
- *Support co-editing of generated outputs*. Allow both the user and the AI system to improve generated outputs.

Design for Appropriate Trust & Reliance
Help the user determine when they should or should not rely on the AI system's outputs by teaching them to be skeptical of quality issues, inaccuracies, biases, underrepresentation, and other issues
- *Calibrate trust using explanations*. Be clear and upfront about how well the AI system performs different tasks by explaining its capabilities and limitations.
- *Provide rationales for outputs*. Show the user why a particular output was generated by identifying the source materials used to generate it.
- *Use friction to avoid overreliance*. Encourage the user to review and think critically about outputs by designing mechanisms that slow them down at key decision-making points.
- *Signify the role of the AI*. Determine the role the AI system will take within the user's workflow.

Design for Imperfection
Help the user understand and work with outputs that may not align with their expectations
- *Make uncertainty visible*. Caution the user that outputs may not align with their expectations and identify detectable uncertainties or flaws.
- *Evaluate outputs using domain-specific metrics*. Help the user identify outputs that satisfy measurable quality criteria.
- *Offer ways to improve outputs*. Provide ways for the user to fix flaws and improve output quality, such as editing, regenerating, or providing alternatives.
- *Provide feedback mechanisms*. Collect user feedback to improve the training of the AI system.

Table 1. Design **principles** and *strategies* for generative AI applications. The left column contains principles that offer new interpretations of existing issues in the development of AI applications. The right column contains principles that focus on new issues that stem from generative AI technologies. Strategies that involve following a design process are indicated with an asterisk (*).

Design for Appropriate Trust & Reliance. Three principles identify unique aspects of generative AI UX: **Design for Generative Variability**, **Design for Co-Creation**, and **Design for Imperfection**.

Each design principle is coupled with a set of four design strategies for how to implement that principle. In some cases, implementing the principle involves following a design process; in other cases, it is implemented through the inclusion of specific types of features or functionality.

These principles and strategies can be employed to support two user goals: (1) *optimization*, in which the user seeks to produce an output that satisfies some task-specific criteria; and (2) *exploration*, in which the user uses the generative process to explore a domain, seek inspiration, and discover alternate possibilities in support of their own ideation. The ways each principle and strategy are applied may differ by user goal, and we elaborate on these differences in Section 10.2.

We note that these principles are just that – *principles* – and not hard rules that must be followed in all design processes. Our view is that it is up to design practitioners to exercise their best judgement in deciding whether a principle applies to their particular use case, and whether any particular strategy should (or should not) be applied.

5 METHODOLOGY

Our goal is to produce a set of clear, concise, and relevant design principles that can be readily applied by design practitioners in the design of applications that incorporate generative AI technologies. We aim for the principles to satisfy the following desiderata:

- Provide designers with language to discuss UX issues unique to generative AI applications, motivated by work that provides designers with specialized vocabulary for domains such as video games [23, 81] and IoT [31];
- Provide designers with concrete strategies and examples that are useful for making difficult design decisions, such as those that involve trade-offs between model capabilities and user needs, motivated by work that focuses simultaneously on end-users of systems [55] and on designers as strategic and collaborative end-users of guidelines [82, 87]; and
- Sensitize designers to the possible risks of generative AI applications and their potential to cause a variety of harms (inadvertent or intentional), and outline processes that could be used to avoid or mitigate those harms (e.g. [66, 180]).

We used an iterative process to develop and refine the design principles, inspired by the process used by Amershi et al. [7] in developing their guidelines for human-AI interaction. We crafted an initial set of design principles via a literature search (Section 6), refined those principles via multiple feedback channels (Section 7), conducted a modified heuristic evaluation exercise to assess their clarity and relevance and identify any remaining gaps (Section 8), and finally applied the principles to two generative AI applications under design to demonstrate their applicability to design practice (Section 9).

In each iteration, we engaged in significant discussion and reflection on the feedback gathered from the previous iteration to produce a new version of the design principles and strategies. In some cases, principles or strategies moved to the next iteration unchanged; in many cases, we made organizational and wording changes. We summarize our iterative process and the outcomes of each iteration in Table 2 and we show how the principles evolved over the iterations in Figure 2.

Iteration	Activity	Goal	Key Outcomes
Iteration 1	Literature Review	Identify relevant research and examples of generative AI application design	Observed hierarchy of high-level design principles implemented by specific UX design strategies; identified 7 initial design principles
Iteration 2	Feedback	Collect feedback from conference workshop and designers within our institution	Developed clearer understanding of Exploration vs. Optimization purposes of use
Iteration 3	Modified Heuristic Evaluation	Test principles for clarity, relevance, and coverage by having designers evaluate commercial generative AI applications	Recognized uniqueness of Generative Variability, Co-Creation, and Imperfection to generative AI; re-categorized Exploration and Optimization separately as user goals
Iteration 4	Application	Demonstrate real-world applicability and utility by having two product teams adopt the principles in their own design work	Observed utility of principles across ideation and evaluation design phases; made clarity improvements to strategy descriptions

Table 2. Summary of the iterative process we used to develop the design principles and strategies.

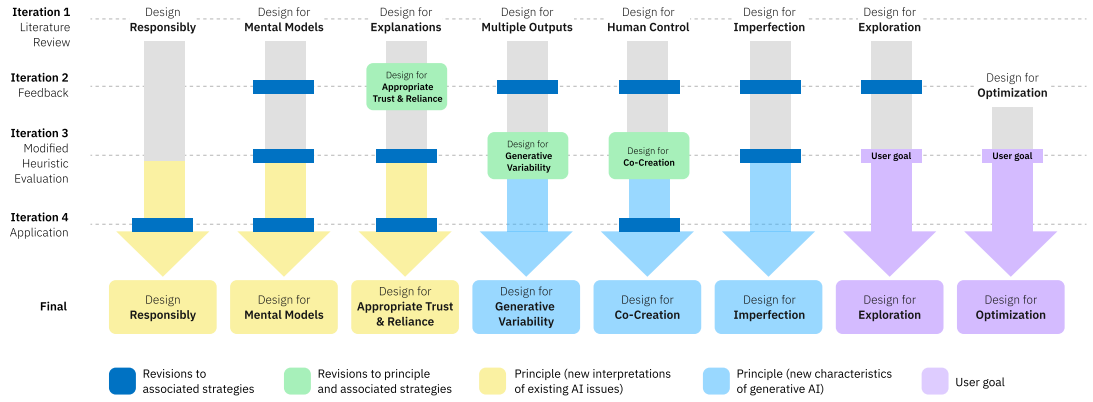

Fig. 2. Evolution of the design principles across four iterations. During Iteration 3, we recognized that some principles offered new interpretations of existing AI system characteristics whereas others identified new characteristics of generative AI.

6 ITERATION 1: CRAFTING INITIAL DESIGN PRINCIPLES

We began our process of identifying design guidelines suitable for generative AI applications by examining recent research in the HCI and AI communities. We conducted a literature review of research studies, guidelines, and analytic frameworks from these communities by searching the ACM Digital Library and Google Scholar for terms including "generative AI, " "design guidelines," and "human-centered AI." These searches identified a set of relevant publications, as well as several recent workshops covering human-AI interaction with generative AI: Human-AI Co-Creation with Generative Models [52, 112, 181], Generative AI and HCI [122], Human-Centered AI [120], and Human-Centered Explainable AI [44]. We then conducted additional searches for terms found within those workshops' proceedings, including "co-creation," "human-AI collaboration," "explainability," and "creative interfaces." Our searches yielded a

representative sample of work that included new advancements and issues in generative AI[11], design guidelines[12], studies of design guideline implementation[13], and studies of human interaction with AI (and generative AI) systems[14]. Finally, to incorporate recent industry developments around generative AI, we also examined a representative set of commercial generative applications (listed in Table 3) to identify common design patterns.

One characteristic that stood out to us in our review was the difference between work that identified important user needs and the specific kinds of UX design that supported those needs. For example, one set of papers examined requirements for explainable AI (XAI) and human-centered explainable AI (HCXAI) through experimental and heuristic methods [43, 98, 166], motivating "explainability" as an important high-level concept. Then, when examining a commercial generative AI system (ChatGPT), we observed how explanations of the system's capabilities and limitations were provided on the home screen. Observations such as these motivated our development of a two-tier principle/strategy structure in which a principle articulates an important characteristic or consideration for a generative AI application and the strategies identify how to implement that principle in the UX.

Our analysis helped us identify several characteristics unique to generative AI that have implications on the user experience: the models' capability of producing multiple outputs [116, 135, 150], the possibility of flaws or imperfections[15] within those outputs [183, 184], and the various ways that people can control or influence those outputs [85, 103, 105]. We also identified how generative AI could enable people to explore a space of possibilities [88] as a byproduct of the generative process. In addition, we identified several existing considerations of AI systems as being particularly important to the generative case, such as using participatory methods [67] to design for real user needs like explainability [44, 166], and understanding the role of the AI in the co-creative process [37, 57, 106, 123, 148, 161].

At this stage, we identified 7 high-level principles and 22 specific strategies for implementing them. Some strategies were related to multiple principles, and at this stage we allowed the overlap; in subsequent iterations, we eliminated these redundancies (we discuss this point further in Section 10.1).

7 ITERATION 2: EXTERNAL AND INTERNAL FEEDBACK

We published the first iteration of the design principles at the Human-AI Co-Creation with Generative Models (HAI-GEN) workshop at IUI [182], attended by approximately 50 researchers from academia and industry. At this workshop, we received informal feedback through discussion sessions and follow-up conversations. We also published this version within our organization as part of a design guide on generative AI, which was viewed by over 1,000 design practitioners. We created an internal discussion channel on this guide to receive additional feedback, including points of confusion and gaps in our framework. Both sources of informal feedback helped us craft the second iteration of the design principles, which introduced the following major changes:

[11]Papers describing advancements and issues in generative AI, including demonstrations of new user interaction capabilities, included Brown et al. [21], Johnson et al. [74], Kaiser et al. [76], Kim et al. [85], Liu and Chilton [103], Louie et al. [105], Metz [117], Perez et al. [135], Ramesh et al. [139], Rombach et al. [140], Ross et al. [143], Sharma [150].

[12]Papers offering design guidelines included Amershi et al. [7], Frijns and Schmidbauer [49], Fukuda-Parr and Gibbons [50], Jobin et al. [72], Liu and Chilton [104], Mohseni et al. [119], PAIR [133], Shneiderman [152], Srivastava et al. [162], Urban Davis et al. [169], Wickramasinghe et al. [186], Wright et al. [188].

[13]Papers examining design guideline implementation included Alnanih and Ormandjieva [4], Li et al. [95], Yildirim et al. [191].

[14]Papers examining human interactions with AI (and generative AI) systems included Deterding et al. [37], Ehsan et al. [43], Gmeiner et al. [54], Grabe et al. [57], Inie et al. [67], Kreminski et al. [88], Liao et al. [98], Lim et al. [101], Liu [102], Louie et al. [105], Lubart [106], Maher [111], Megahed et al. [116], Muller et al. [123], Seeber et al. [148], Spoto and Oleynik [161], Sun et al. [166], Verheijden and Funk [173], Wan et al. [175], Weidinger et al. [180], Weisz et al. [183, 184].

[15]We identified two types of imperfection: (a) flaws present in the model's outputs, such as bugs in source code or hallucinations in Q&A, and (b) misalignments between a user's intent, how that intent is expressed in a prompt, and what the generative model produces as a result. Even if category (a) would disappear (e.g. due to technological improvements), category (b) would remain and imperfection would still be a critical design issue.

- We identified how users' goals in using a generative AI system can differ, leading us to include two task-specific principles: the existing **Design for Exploration** principle, in support of use cases around ideation, exploration, and learning; and a new principle, **Design for Optimization**, in support of use cases for which the production of a singular artifact is desired.
- We recognized that explainability needs for generative AI systems, while important, were not necessarily an "end" in and of themselves. Rather, explainability is one way to **Design for Appropriate Trust & Reliance**, leading us to incorporate existing explainability strategies into this new principle.
- We re-articulated all of the design strategies as rules of action (e.g. a verb followed by 2-6 words), akin to how Amershi et al. phrased their guidelines.
- We identified that five design strategies were about the design process itself rather than specific UX capabilities.

At the end of Iteration 2, we had a set of 8 high-level principles implemented by 29 specific strategies.

8 ITERATION 3: MODIFIED HEURISTIC EVALUATION

Following Iteration 2, we sought to conduct a more rigorous evaluation of the design principles and strategies. Given the potential gap between research literature and real-world practice, we specifically wanted to determine their clarity to our target audience of design practitioners, understand their relevance to commercial generative AI applications, and identify any additional gaps in our framework. In support of these goals, we drew inspiration from Amershi et al. by creating a modified heuristic evaluation exercise.

8.1 Method

Heuristic evaluation is a discount usability method for identifying violations of usability guidelines in a user interface [128]. Amershi et al. [7] developed a *modified* heuristic evaluation in which evaluators reviewed an AI-infused user experience with the purpose of evaluating the heuristics themselves. We similarly developed a modified heuristic evaluation to evaluate our design principles for generative AI applications. We asked evaluators to examine a range of commercial generative AI applications and identify examples that demonstrate the use of the principles and strategies, as well as examples of generative AI-specific design choices that were not covered by the principles and strategies. This exercise helped us evaluate the relevance, clarity, and coverage of the design principles and strategies.

We identified 9 commercial generative AI applications to use in the evaluation, listed in Table 3. We selected these applications due to their popularity in consumer or enterprise markets, their ability to be used within our organization without incurring costs, and the range of use cases and output modalities they supported. We also considered applications that incorporated generative AI features in one of two distinct ways[16]: either as the core user experience or as a component within an existing user experience.

We recruited 18 design practitioners within our organization and outside of our immediate team to perform the modified heuristic evaluation. We sought evaluators with varied design roles and levels of experience to ensure the principles were clear and relevant across different specialties and expertise levels. Of the 18 evaluators, 11 (61.1%) identified as male, 6 (33.3%) identified as female, and 1 preferred not to disclose. The majority of evaluators were User Experience Designers (16, 88.9%), one evaluator was a Design Researcher, and one was a Research Software Developer[17].

[16] Beavers et al. [13] refer to different "altitudes" at which AI capabilities can be embedded in a UX, including "above" (the AI capability is the central focus of the application), "beside" (the AI capability sits in a side panel next to the main UI), and "inside" (the AI capability is embedded within the existing UI).
[17] Despite not having a design role, this evaluator worked on an HCI research team and possessed over 20 years of professional design experience, which we felt sufficiently qualified them to participate in this exercise.

Application	Description	AI Incorporation
ChatGPT	Conversational Q&A	Core (Web app)
Google Bard	Conversational Q&A	Core (Web app)
DALL-E	Text-to-image generator	Core (Web app)
DreamStudio	Text-to-image generator	Core (Web app)
Midjourney	Text-to-image generator	Component (Discord)
Adobe Firefly Generative Fill	Text-to-image generator	Component (Adobe Photoshop)
IBM watsonx.ai Prompt Lab	Prompt playground for large language models	Core (Web app)
Github Copilot	Natural language to source code	Component (Visual Studio Code)
AIVA	Music generation	Core (Web app)

Table 3. Commercial generative AI applications used in our modified heuristic evaluation. AI capabilities were present either as the core user experience or embedded as a component within an existing application.

Four evaluators (22.2%) reported having 1-4 years of experience, four (22.2%) had 5-9 years, two (11.1%) had 10-14 years, two (11.1%) had 15-19 years, and five (27.7%) had 20+ years[18]. Most evaluators had some experience with discount usability testing methods: three evaluators (16.6%) reported low or very low experience, five (27.7%) reported medium experience, and nine (50%) reported high or very high experience. Given that generative AI design is an emerging field, our evaluators tended not to have high levels of experience in this area: eight evaluators (44.4%) reported very low to low experience, eight (44.4%) reported medium experience, and one (5.5%) reported a high level of experience.

Evaluators self-selected an application familiar to them and completed their evaluation individually (as is standard practice [128]) and remotely. As our evaluators were not involved in the design of these products, they were unable to evaluate the process-oriented strategies (all strategies within **Design Responsibly** plus *Evaluate users' mental models*). Thus, these strategies were excluded from Iteration 3, and we made it a point to evaluate them in Iteration 4 (Section 9). Participants recorded their evaluations of all other principles in a Mural[19] template. Two evaluators examined each application and each evaluation took approximately one hour.

We crafted short descriptions[20] for each principle and strategy to orient our evaluators. For each principle, we asked evaluators to begin by capturing examples in the Mural canvas of how their application applied the principle. At this stage, specific strategies in the Mural were covered with an overlay to encourage evaluators to find examples without being biased by our strategies, in hopes that they might identify new ones. After capturing examples, evaluators were instructed to remove the overlay, then label each example with a strategy we provided, "not sure", or a write-in for a new strategy. After finding and labeling examples, evaluators rated the relevance of each design principle and its strategies on a 4-point scale: "Yes, they were clearly relevant," "Yes, they were relevant but I struggled to find examples," "No, they were clearly not relevant," and "Not sure." They also rated the clarity of the principle as a whole on a 5-point scale from "Very unclear" to "Very clear" and provided suggestions for improvement. Finally, after reviewing all of the principles, evaluators were asked to identify any additional design features in their application that were not covered by the principles.

[18] As one evaluator did not complete the demographic survey, percentages do not add up to 100%.
[19] Mural is a collaborative, graphical canvas application. https://mural.co
[20] These descriptions were the first iteration of those shown in Appendix A.

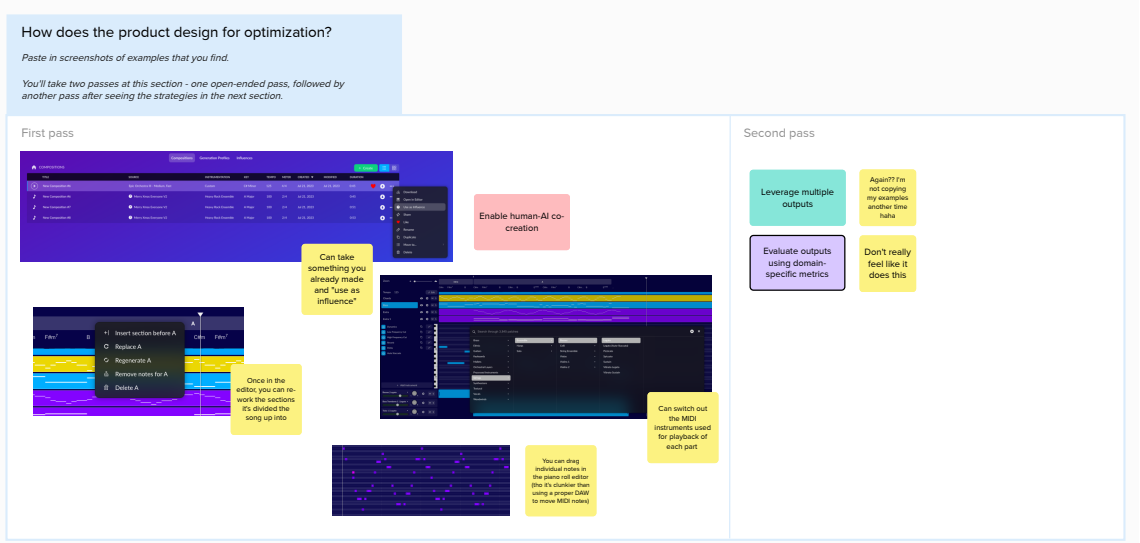

Fig. 3. Portion of an evaluation of AIVA for the principle of **Design for Optimization**.

8.2 Results

Our evaluators produced 18 heuristic evaluation canvases laden with screenshots and sticky notes that identified real-world instances of the principles and strategies. They also left notes about points of difficulty or confusion.

Figure 3 shows a portion of one evaluator's canvas in which they evaluated AIVA for **Design for Optimization**. This example shows how the evaluator found examples of various kinds of controls in the tool, along with feedback on the repetitiveness between *Leverage multiple outputs*, *Show multiple outputs*, and **Design for Multiple Outputs**: "Again?? I'm not copying my examples another time."

To analyze the evaluation data, two authors first individually examined the completed canvases for examples and comments that indicated the relevance, clarity, and coverage of the principles. They also reviewed participants' ratings of relevance and clarity and delved into their examples and comments to understand instances of lower ratings. They then converged with the other authors to review their findings and discuss potential ways to improve the principles and strategies.

8.2.1 Relevance. To assess the relevance of the design principles and strategies to commercial generative AI applications, we counted the number of examples evaluators found. Evaluators identified 286 total examples across all principles; they found a collective average of 11.9 examples for each strategy, and every strategy had at least one example. The wealth of examples found suggests the principles and strategies were relevant to a range of commercial generative AI applications. Evaluators also generally rated each principle as being relevant (Figure 4a).

The relatively lower relevance ratings for **Design for Appropriate Trust & Reliance**, **Design for Human Control**, and **Design for Optimization** stemmed from differences in application domain and output modality. In some cases, we accepted that relevance may vary by use case; in other cases, we addressed issues raised by participants to clarify or expand relevance. For example, the four evaluators who rated **Design for Appropriate Trust & Reliance** as "not relevant" had examined image or music generation applications and felt that overreliance was less of a concern for

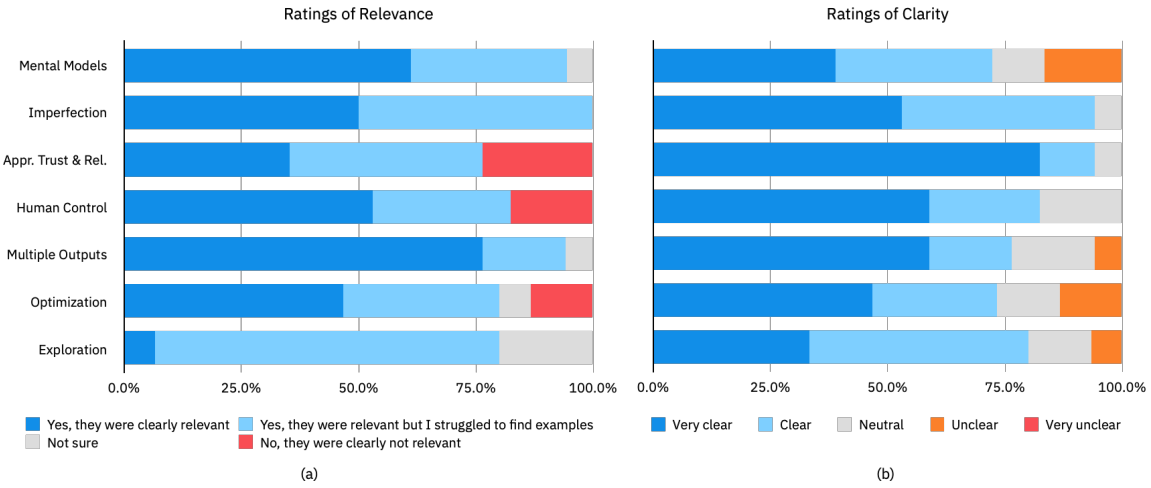

Fig. 4. Evaluators' ratings of the (a) relevance and (b) clarity of each principle and its strategies to their application in the modified heuristic evaluation.

creative applications. In response to this observation, we added examples of risks to be wary of in creative outputs (e.g. quality issues, bias, and underrepresentation [17, 18, 42]) to clarify its relevance to such applications. We made similar modifications to strategies that were too narrowly focused on specific domains or output modalities.

8.2.2 Clarity. To assess the clarity of the design principles and strategies, we identified instances where evaluators noted overlap or redundancy between different principles or strategies, expressed confusion, or interpreted a principle or strategy differently from how we intended. We also asked evaluators to rate the clarity of each principle (Figure 4b), and they were generally rated as being clear.

Participants identified eight overlap issues. Notably, five evaluators found that nearly all strategies in **Design for Exploration** and **Design for Optimization** overlapped in some way with strategies in other principles. This observation led us to reconsider how to incorporate exploration and optimization within our framework. Ultimately, we recognized that exploration and optimization are user goals rather than characteristics of a generative AI application, and hence should be communicated as such (we discuss this point further in Section 10.2). Other overlap issues were reconciled by merging redundant strategies.

We observed 16 instances in which an evaluator's use of a strategy label mismatched our intention for what the strategy represented. We made two major changes in response to these mismatches. First, we reframed **Design for Human Control** as **Design for Co-Creation** in response to frequent misinterpretations of "controls" as affordances unrelated to the generative process (such as Photoshop's editing tools). **Design for Co-Creation** provides greater specificity to generative AI's unique capabilities for human-AI co-creation, which has been examined extensively within HCI communities (e.g., [34, 53, 77, 121, 129]). The second change was to rename **Design for Multiple Outputs** to **Design for Generative Variability** to better characterize its purpose after observing that many evaluators narrowly interpreted this principle as solely being about the *display* of multiple outputs. We made additional wording changes and clarifications to other principles and their associated strategies in response to participants' feedback.

8.2.3 Coverage. Evaluators found new examples that reflected gaps in our framework, resulting in three new strategies: *Teach the AI system about the user*, *Help the user craft effective outcome specifications*, and *Support co-editing of generated outputs*. We included *Teach the AI system about the user* in **Design for Mental Models** as it addresses recent research in Mutual Theory of Mind [33, 177, 187]. We included *Help the user craft effective outcome specifications* and *Support co-editing of generated outputs* in **Design for Co-Creation** as they are most closely related to the co-creative process.

9 ITERATION 4: APPLICATION TO GENERATIVE AI UX DESIGN

Design guidelines can be difficult to put into practice [59, 113, 160, 164, 176, 192], often because they describe goals rather than actions [73]. Our strategies were meant to capture "actions" that practitioners could take to apply the principles to their work. After refining the principles and strategies for relevance and clarity, we evaluated their utility within the design process by conducting structured, exploratory workshops with design practitioners within our organization who work on generative AI applications. Our primary goal was to understand how effectively the design principles could be applied in practice, but we remained open to identifying additional issues regarding relevance, clarity, and coverage.

9.1 Method

We held two workshops with two different teams (Table 4) to evaluate the design principles and strategies in practice. Workshop 1 was held with an internal team comprised of four design practitioners working on the IBM watsonx.ai Prompt Lab[21], a prompt testing environment for large language models. Workshop 2 involved a separate internal team of ten design practitioners in the early, formative stages of designing an internal LLM-based conversational tool that provides UX research support. We selected these two teams as they provided a broader view on the actionability of the design principles in different phases of design: a later, evaluative stage (Workshop 1) and an earlier, ideation phase (Workshop 2).

Workshop 1 Evaluate existing LLM prompt tool		**Workshop 2** Ideate on a future LLM-based conversational tool			
P1-1	Design Lead	P2-1	UX Researcher	P2-6	Design Research Lead
P1-2	Software Designer	P2-2	UX Researcher	P2-7	Design Manager
P1-3	Design Manager	P2-3	UX Researcher	P2-8	UX Researcher
P1-4	UX Researcher	P2-4	Program Manager	P2-9	Technical Program Lead
		P2-5	UX Researcher	P2-10	Design Research Lead

Table 4. Participants in each of the workshops to assess the actionability of the design principles.

Workshops took place remotely via video conferencing and were recorded with participants' consent. Each session lasted 90 minutes and included two moderators and two note-takers. One moderator began each workshop by presenting an overview of the design principles and strategies. To minimize the time required of participants, we split each session into two groups and assigned three principles to each group. Each break-out group contained one moderator and one note-taker.

[21]Watsonx.ai. https://watsonx.ai

For each principle, participants were first asked to identify ways they were already "designing for" or considering the principle. Next, they identified relevant strategies that they had not yet considered and brainstormed ways to leverage them to improve their product. This brainstorming session produced new design ideas that team members shared and discussed with each other. At the end of the session, participants reflected on the actionability of the principles within their design process. The moderators and note-takers of each workshop reviewed participants' design ideas and recording transcripts to identify insights on the usefulness of the principles and recommendations for improvement.

9.2 Results

9.2.1 Applicability to practice. Participants in Workshop 1 brainstormed a total of 46 design ideas and participants in Workshop 2 brainstormed 56 design ideas. These design ideas included feature requirements, new affordances, design processes to try, and questions to consider when making design decisions. Groups generated between 5 and 14 ideas per principle, and participants generated multiple, varied ideas for all of the principles. For example, when considering the strategy, *Help the user craft effective outcome specifications*, P1-3 thought of an idea to "provide different 'effects' that bake in some prompt content, (e.g. in the style of a famous author)," and P1-4 proposed a system to "reward 'best in class' prompt authors and celebrate and share" their work.

When asked about the actionability of their brainstormed ideas, P1-1 responded, "I definitely think a lot of these could go on a future roadmap." P2-1 commented that the workshop, "made it clear crucial blind spots that could put the [application] idea at risk if not addressed," and that it helped their team, "quickly generate new requirements." Participants' breadth of design ideas and comments on workshop outcomes indicate that practitioners are able to leverage the principles and strategies to inform useful, actionable design improvements.

Participants also shared ideas on how to improve the actionability of the principles. As their understanding of the principles was limited to the brief overview provided at the start of the workshop, they felt that having more details and resources to learn about the principles and strategies would make them easier to apply. P1-3 commented, "having some examples of these concepts out in the world or in other tools might be a useful way to get a grasp of how the concept works." In support of this need, we include a library of examples in Appendix A that provide richer detail on how each strategy has been applied within existing applications.

Participants also shared insights on how the principles would be incorporated into their design process. P1-4 commented that involving more roles in a workshop, such as developers and product managers, would add value. P2-8 felt that **Design for Imperfection** "can't be applied unless user research is done" due to a lack of understanding of the user's expectations for model outputs. P2-6 also spoke about the value of user research since "the idea or solution... may look differently depending on the user." These comments indicate that a baseline understanding of users is needed to identify concrete design ideas from the principles and strategies, in line with our recommendation to *Use a human-centered approach*.

The outcomes from these workshops demonstrated that the design principles can be applied in both an early ideation stage and a later evaluation stage to drive actionable design ideas.

9.2.2 Relevance and clarity improvements. Workshop participants also provided feedback on the relevance and clarity of the design principles. This feedback primarily resulted in minor wording changes to the process-oriented strategies that we were unable to evaluate in Iteration 3; no major organizational changes were made as a result of this feedback. After incorporating this feedback, we produced our final set of design principles and strategies (Table 1).

9.3 Final clarity evaluation

To determine whether the wording changes we made after Iterations 3 & 4 impacted the clarity of the final design principles and strategies, we ran a follow-up survey with the evaluators from Iteration 3. Fourteen of 18 evaluators responded to our survey (77.7% response rate).

The clarity ratings collected in Iteration 3 were at the higher end of the 5-point scale (M (SD) = 4.29 (0.90) of 5). The changes we made in Iteration 4 made a small but significant improvement to clarity (M (SD) = 4.58 (0.58) of 5), $F(1, 184) = 5.88$, $p = .02$, $\eta_p^2 = .03$ (small).

10 DISCUSSION

We identified a set of six principles important to the design of generative AI applications, along with a companion set of 24 strategies for implementing those principles within a user experience. These principles were developed iteratively using a combination of critical conceptual analyses (to ensure scientific validity) and empirical work (to ensure real-world utility).

We collected formal feedback on the principles from a 18 design practitioners who collectively evaluated them against 9 commercial applications. We then collected feedback from 12 design practitioners on two design teams who applied them to both the formative and evaluative stages of product design. We found that the principles helped design practitioners generate useful and actionable design improvements and were applicable to a range of generative AI applications, including those that generate different types of media (e.g. text, images, music).

We discuss two issues that kept surfacing throughout our development process that required us to think deeply about where to "draw the lines," either between different principles and their strategies when we identified overlap or redundancy, or between what we later identified as a difference between user goals and characteristics of generative AI. We also discuss our strategies for putting the principles into action within our organization, as well as limitations and opportunities for future work.

10.1 Guideline organization

Early in our first iteration, we observed a hierarchical relationship emerge between high-level design principles that identified unique or differentiated aspects of generative AI and lower-level strategies for implementing those principles in a user experience. However, the relationships between which strategies applied to which principles were not always clear, as some strategies could be used to support multiple principles. For example, in Iteration 1, the strategy *Visualizing differences* was included in **Design for Multiple Outputs**, as it could help users understand the differences amongst those outputs (especially for use cases where those differences might be subtle). But it was also included in **Design for Imperfection**, as it could help users more easily identify problematic outputs. Another example from Iteration 1 was the use of a *Sandbox / Playground Environment*, which supported **Design for Imperfection** by not tainting an artifact-under-creation with potentially-problematic generated content (e.g. source code with bugs or text containing factual errors). But it was also included in **Design for Exploration**, as a sandbox provides a separate space for users to explore new candidates without interfering with their main working environment.

As we worked through subsequent iterations, we wrestled with whether we should continue to allow strategies to overlap between principles or aim for a clean separation. During Iteration 2, with the delineation between **Design for Exploration** and **Design for Optimization**, even more redundancy was introduced as many strategies support both

kinds of uses. At this point, we even considered completely decoupling the strategies from the principles and providing an indication on each strategy (such as a tag) for which principle(s) it supported.

We ultimately decided to maintain the nesting of strategies within principles and aim for establishing clean boundaries. We made this decision because the amount of overlap diminished when we refined the principles during Iteration 3 and separated out the user goals of optimization and exploration. Our evaluators also experienced frustration when they were unable to differentiate between strategies, indicating a need to eliminate overlaps. However, we note that a single UX feature may be used to implement more than one principle or strategy (see Appendix A for examples); therefore, we only sought to reduce conceptual overlaps between the principles and strategies themselves, as opposed to overlaps when a specific UX capability addresses multiple principles or strategies.

10.2 User goals versus design principles

In reviewing the *Library of Mixed Initiative Creative Interfaces* [161], we realized that generative capabilities are sometimes an end in themselves, but other times are a means to achieving another goal. We identified these two different purposes of use as optimization and exploration, respectively:

- In optimization use cases, the process of generating artifacts is an end: users use the generative capability to produce one or more artifacts that satisfy their needs, such as a source code function that implements a desired operation, a molecule that possesses specific properties, or an image that depicts a desired scene or character. We labeled this class of usage as "optimization" in recognition that the generative AI model may not produce a flawless or "perfect" output, and some amount of refinement (either by the user or the AI) may be required before it is satisfactory.
- In exploratory use cases, the process of generating artifacts is a means to an end: the purpose is not to generate the artifact, but to use the generated artifacts in order to learn about a domain (e.g. programming [100] or medicine [89]) or be inspired by seeing new or different possibilities (e.g. brainstorming [145] or pre-writing [175]). Few other types of AI technology support this kind of usage, where the emphasis is on assisting people in conducting a thought process.

During Iteration 3, we ultimately decided to remove **Design for Exploration** and **Design for Optimization** as core design principles because of the strong degree of overlap between their strategies and the strategies of other principles. In fact, we could not even clearly delineate different generative AI applications as supporting exploratory versus optimization usage, because many applications supported both, and users might even alternate between the two kinds of usage when using the application. For example, work by Weisz et al. [184] shows how software engineers used generative technologies not only to produce source code translations (optimization), but also to improve their own knowledge of programming (exploration), within the same overall task context. Hence, we drew a line between user goals and design principles (depicted in Figure 1).

We assert that each principle broadly supports both user goals, but the extent of their support does differ by goal. **Design for Imperfection** is strongly aligned with optimization use cases, as the reason why optimization is even necessary is because of the imperfect outputs produced by generative models. Concurrently, **Design for Generative Variability** is strongly aligned with exploration use cases, as generative variability is a key enabler of exploration. However, we note that **Design for Imperfection** can also support exploratory use by embracing unexpected "imperfections" that arise from discrepancies between a user's intent and the model's output. In addition,

Design for Generative Variability can also support optimization use cases by helping users narrow down on an option that fits their needs from a wide field.

Another principle that has a high degree of affinity to optimization is **Design for Appropriate Trust & Reliance**, especially when generated outputs are used within high-stakes domains (e.g. code, customer service). Trust and reliance may be of a lesser concern for exploration use cases, although users should still be wary of bias, underrepresentation, and other harms that may occur.

Finally, **Design for Co-Creation** can be applied to both exploration and optimization tasks, but the strategies of *Help the user craft effective outcome specifications* and *Support co-editing of generated outputs* may be more important when users need to optimize generated outputs to fit certain criteria.

We conclude that the principles and strategies form a toolbox that design practitioners can use holistically or selectively as they craft user experiences for generative AI applications. Design practitioners know their users and their needs best – exemplified by *Use a human-centered approach* – and it is our hope that we have provided useful vocabulary for them to understand and design for the new and different kinds of uses that generative AI systems offer.

10.3 Adoption within our organization

As discussed in Section 2.1, the HCI community has produced a prodigious number of design guidelines throughout its history. But, as noted by both Soni et al. [160] and Stark et al. [164], our community struggles with bridging the gap between the development of scientifically-grounded guidelines and real-world design practice. We developed our design principles specifically to provide practical and actionable support to design practitioners. Therefore, we undertook a number of efforts to promote their adoption within our organization.

(1) **Actionable activities**. To bridge the gap between theory and practice, we developed activities for designers to apply the principles and strategies to their own work. Chief among them is a heuristic evaluation that uses the principles as heuristics for designers to evaluate the user experience of generative AI applications. We created and disseminated a self-contained Mural template that guides designers through this evaluation to identify new ideas and opportunities for design improvement. We also developed workshop activities for identifying applications of generative AI that drive user value and evaluating a user's mental model of an AI system.

(2) **Progressive detail**. When we initially developed the principles and strategies, we wrote about them extensively in a comprehensive guide that provided foundational knowledge and case studies on generative AI, which we shared with our internal design community. We received feedback that the level of detail was informative but too lengthy for busy designers. In response, we developed two condensed presentations: 1) paragraph-length descriptions for each principle and strategy (shown in Appendix A) which were included in the generative AI heuristic evaluation template, and 2) one-sentence descriptions of each principle and strategy (shown in Table 1) which were published on an internal website for the design of AI applications.

(3) **Hands-on outreach**. We conducted outreach activities to raise awareness of the principles within our organization. Some of these efforts targeted a general design audience, such as creating a discussion group[22] for generative AI design and presenting the principles at internal seminars. Other outreach targeted designers on key product teams. As one example, we held a workshop attended by 62 people at an internal design event to teach designers how to conduct a heuristic evaluation of generative AI applications. Instead of using a sample application, we evaluated a product recently released by our organization and invited the product's design team

[22]This group grew to over 1,200 members over the course of 9 months.

to participate. In an hour-long session, participants identified 10 usability issues and 6 new feature ideas, which we discussed in detail with the product team in follow-up meetings. They reported our findings to be useful and included several recommendations in their roadmap.

(4) **Executive sponsorship**. In addition to bottom-up dissemination, we also worked with key executives in our design organization to encourage relevant product teams to adopt the principles (as recommended by Madaio et al. [110]). Through this effort, we introduced the principles to 10 product teams who were in the process of learning about generative AI and identifying opportunities for incorporating it into their product.

Akin to Yildirim et al.'s observations on how their guidebook improved AI literacy within their organization and helped designers establish credibility and advocate for user needs [192], we found our materials had a similar impact. The executive sponsorship of our work and the adoption of the principles by numerous product teams speak not just to their practical utility, but also for the great need to equip design practitioners and enable them to "have a seat at the table" in the creation of generative AI applications.

10.4 Limitations and future work

The field of generative AI is undergoing rapid innovation, both in the pace of technological development and in how those technologies are being brought to the market. We view our principles as beginning a discussion on how to design effective and safe generative AI applications. As the pace of innovation continues and new generative AI applications are developed, we anticipate new challenges to be uncovered, necessitating new sets of guidelines, tools, best practices, design patterns, and evaluative methods.

One challenge we encountered with the modified heuristic evaluation was in its use to evaluate *the design principles themselves* through the process of evaluating a generative AI application. Not all of our evaluators understood this distinction, and as a result, we sometimes received feedback about shortcomings of the applications that was less relevant to our goal of improving the principles. We recommend providing stronger introductory examples that focus on how they help evaluate the principles rather than the products.

Another limitation of the modified heuristic evaluation was our focus on evaluating commercially-available generative AI applications. There are also many experimental applications in this space, but we did not examine them. Our restriction to commercial applications excluded other ways of interacting with generative AI applications, such as through narrative [24], lyric and other poetic forms [147], and movement [174].

Finally, our design guidelines are entirely focused on helping design practitioners to develop the user experience for a generative AI application. But, UX design is only one portion of the AI development lifecycle, which includes other phases such as model selection, model tuning, prompt engineering, deployment & monitoring, and more. As decisions made during those phases will ultimately impact the user experience, we believe design practitioners ought to have their inputs considered. However, the design principles do not currently help them understand, for example, how to determine which generative model should be used to implement a Q&A use case and when that model's performance is "good enough," or how to hide a generative model's inference latency. In addition, organizational policies may be created that govern the uses of generative AI. We believe there is room for expansion to identify how designers can participate in these kinds of technical and policy decisions that have an impact on the user experience.

11 CONCLUSION

We developed a set of six principles for the design of applications that incorporate generative AI technologies. Three principles – **Design Responsibly**, **Design for Mental Models**, and **Design for Appropriate Trust & Reliance** – offer new interpretations of known issues with the design of AI systems when viewed through the lens of generative AI. Three principles – **Design for Generative Variability**, **Design for Co-Creation**, and **Design for Imperfection** – identify issues that are unique to generative AI applications. Each principle is coupled with a set of strategies for how to implement it within a user experience, either through the inclusion of specific types of UX features or by following a specific design process. We developed the principles and strategies using an iterative process that involved reviewing relevant literature in human-AI collaboration and co-creation, collecting feedback from design practitioners, and validating the principles against real-world generative AI applications. We also demonstrated the value and applicability of the principles by applying them in the design process of two generative AI applications. As generative AI technologies are rapidly being incorporated into existing applications, and entirely new products are being created with these technologies, we see significant value in principles that aid design practitioners in harnessing these technologies for the benefit of their users in safe and effective ways.

ACKNOWLEDGMENTS

We thank everyone at IBM who provided valuable feedback and guidance in the development of these design principles.

REFERENCES

[1] Adobe. 2023. Adobe Releases New Firefly Generative AI Models and Web App; Integrates Firefly Into Creative Cloud and Adobe Express. https://news.adobe.com/news/news-details/2023/Adobe-Releases-New-Firefly-Generative-AI-Models-and-Web-App-Integrates-Firefly-Into-Creative-Cloud-and-Adobe-Express/default.aspx.

[2] Rukshan Alexander, David Murray, and Nik Thompson. 2017. Cross-cultural web design guidelines. In *Proceedings of the 14th International Web for All Conference*. 1–4.

[3] Radwan Ali, Mike Gallivan, and Seema Sangari. 2019. A Study of mobile apps in the banking Industry. *International Journal of Digital Society (IJDS)* 10, 3 (2019), 1524–1533.

[4] Reem Alnanih and Olga Ormandjieva. 2016. Mapping HCI principles to design quality of mobile user interfaces in healthcare applications. *Procedia Computer Science* 94 (2016), 75–82.

[5] A Alsswey, IN Umar, and H Al-Samarraie. 2018. Towards mobile design guidelines-based cultural values for elderly Arabic users. *Journal of Fundamental and Applied Sciences* 10, 2S (2018), 964–977.

[6] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[8] Apple Computer, Inc. 1987. *Apple human interface guidelines: The Apple desktop interface.* Addison Wesley Publishing Company.

[9] Apple, Inc. 2022. Human Interface Guidelines for Machine Learning. Retrieved 14-Aug-2023 from https://developers.apple.com/design/human-interface-guidelines/technologies/machine-learning/introduction

[10] Maryam Ashoori and Justin D Weisz. 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675* (2019).

[11] Nagadivya Balasubramaniam, Marjo Kauppinen, Sari Kujala, and Kari Hiekkanen. 2020. Ethical guidelines for solving ethical issues and developing AI systems. In *Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21*. Springer, 331–346.

[12] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[13] Kurtis Beavers, Shawndell Elfring, Hannah Reed, and Rachel Shepard. 2023. *UX: Designing for Copilot | DIS214H.* Microsoft Developer. https://youtu.be/WiCVEMH4HTI

[14] Shirley Ann Becker. 2004. A study of web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11, 4 (2004), 387–406.

[15] Kirsty A Beilharz, Joanne Jakovich, and Sam Ferguson. 2006. Hyper-shaku (Border-crossing): Towards the Multi-modal Gesture-controlled Hyper-Instrument.. In *NIME*. 352–357.

[16] Nigel Bevan. 2005. Guidelines and standards for web usability. In *Proceedings of HCI International*, Vol. 2005. Citeseer, 10.

[17] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 396–410.

[18] IBM AI Ethics Board. 2023. Foundation models: Opportunities, risks and mitigations. Retrieved 05-Sep-2023 from https://www.ibm.com/downloads/cas/E5KE5KRZ

[19] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[22] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[23] Paul Cairns, Christopher Power, Mark Barlet, and Greg Haynes. 2019. Future design of accessibility in games: A design vocabulary. *International Journal of Human-Computer Studies* 131 (2019), 64–71.

[24] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[25] James J Cappel and Zhenyu Huang. 2007. A usability analysis of company websites. *Journal of Computer Information Systems* 48, 1 (2007), 117–123.

[26] PV Charan, Hrushikesh Chunduri, P Mohan Anand, and Sandeep K Shukla. 2023. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads. *arXiv preprint arXiv:2305.15336* (2023).

[27] Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. Parameter-Efficient Fine-Tuning Design Spaces. *arXiv preprint arXiv:2301.01821* (2023).

[28] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[29] Vijil Chenthamarakshan, Payel Das, Samuel Hoffman, Hendrik Strobelt, Inkit Padhi, Kar Wai Lim, Benjamin Hoover, Matteo Manica, Jannis Born, Teodoro Laino, et al. 2020. CogMol: Target-specific and selective drug design for COVID-19 using deep generative models. *Advances in Neural Information Processing Systems* 33 (2020), 4320–4332.

[30] Kulsiri Chirayus and Aziz Nanthaamornphong. 2020. Cognitive mobile design guidelines for the elderly: a preliminary study. In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 673–678.

[31] Yaliang Chuang, Lin-Lin Chen, and Yoga Liu. 2018. Design vocabulary for human–IoT systems communication. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[32] John Joon Young Chung, Minsuk Chang, and Eytan Adar. 2021. Gestural Inputs as Control Interaction for Generative Human-AI Co-Creation. In *Workshops at the International Conference on Intelligent User Interfaces (IUI)*.

[33] Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological medicine* 50, 7 (2020), 1057–1061.

[34] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 196–207.

[35] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *arXiv preprint arXiv:2307.08715* (2023).

[36] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.

[37] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 628–635.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[39] DW Dodson and NL Shields Jr. 1978. Development of user guidelines for ECAS display design, volume 1. (1978). https://ntrs.nasa.gov/api/citations/19790007504/downloads/19790007504.pdf

[40] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[41] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[42] Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating Algorithmic Bias: From Neutrality to Justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 691–704.

[43] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[44] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI conference on human factors in computing systems extended abstracts*. 1–7.

[45] Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. *arXiv preprint arXiv:2306.15087* (2023).

[46] KJ Kevin Feng, Maxwell James Coppock, and David W McDonald. 2023. How Do UX Practitioners Communicate AI as a Design Material? Artifacts, Conceptions, and Propositions. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2263–2280.

[47] Giorgio Franceschelli and Mirco Musolesi. 2022. Copyright in generative deep learning. *Data & Policy* 4 (2022), e17.

[48] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[49] Helena Anna Frijns and Christina Schmidbauer. 2021. Design Guidelines for Collaborative Industrial Robot User Interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*. Springer, 407–427.

[50] Sakiko Fukuda-Parr and Elizabeth Gibbons. 2021. Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy* 12 (2021), 32–44.

[51] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6957–6966.

[52] Werner Geyer, Lydia B Chilton, Justin D Weisz, and Mary Lou Maher. 2021. HAI-GEN 2021: 2nd Workshop on Human-AI Co-Creation with Generative Models. In *26th International Conference on Intelligent User Interfaces-Companion*. 15–17.

[53] Vlad Petre Glăveanu. 2014. *Distributed creativity: What is it?* Springer.

[54] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[55] Chao Gong, Yue Qiu, and Bin Zhao. 2018. Establishment of Design Strategies and Design Models of Human Computer Interaction Interface Based on User Experience. In *Design, User Experience, and Usability: Theory and Practice: 7th International Conference, DUXU 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 7*. Springer, 60–76.

[56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[57] Imke Grabe, Miguel González-Duque, Sebastian Risi, and Jichen Zhu. 2022. Towards a Framework for Human-AI Interaction Patterns in Co-Creative GAN Applications. *Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland* (2022).

[58] Jonathan Grudin. 1990. The computer reaches out: The historical continuity of interface design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 261–268.

[59] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines* 30, 1 (2020), 99–120.

[60] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509* (2022).

[61] Hartmut Hoehle, Ruba Aljafari, and Viswanath Venkatesh. 2016. Leveraging Microsoft's mobile usability guidelines: Conceptualizing and developing scales for mobile application usability. *International Journal of Human-Computer Studies* 89 (2016), 35–53.

[62] Lars Erik Holmquist. 2017. Intelligence on tap: artificial intelligence as a new design material. *interactions* 24, 4 (2017), 28–33.

[63] Stephanie Houde, Vera Liao, Jacquelyn Martino, Michael Muller, David Piorkowski, John Richards, Justin Weisz, and Yunfeng Zhang. 2020. Business (mis) use cases of generative ai. *arXiv preprint arXiv:2003.07679* (2020).

[64] Johannes Huebner, Remo Manuel Frey, Christian Ammendola, Elgar Fleisch, and Alexander Ilic. 2018. What people like in mobile finance apps: An analysis of user reviews. In *Proceedings of the 17th international conference on mobile and ubiquitous multimedia*. 293–304.

[65] IBM. 2022. Design for AI. Retrieved 14-Aug-2023 from https://www.ibm.com/design/ai/

[66] IBM. 2022. What is AI ethics? Retrieved 14-Aug-2023 from https://www.ibm.com/topics/ai-ethics

[67] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[68] Mikhail Jacob, Alexander Zook, and Brian Magerko. 2013. Viewpoints AI: Procedurally Representing and Reasoning about Gestures.. In *DiGRA conference*.

[69] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 310–323.

[70] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. *arXiv preprint arXiv:2307.04657* (2023).

[71] Jiyou Jia and Bilan Zhang. 2018. Design guidelines for mobile MOOC learning—an empirical study. In *Blended Learning. Enhancing Learning Success: 11th International Conference, ICBL 2018, Osaka, Japan, July 31-August 2, 2018, Proceedings 11*. Springer, 347–356.

[72] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[73] Jeff Johnson. 2020. *Designing with the mind in mind: simple guide to understanding user interface design guidelines*. Morgan Kaufmann.

[74] Tristan E Johnson, Youngmin Lee, Miyoung Lee, Debra L O'Connor, Mohammed K Khalil, and Xiaoxia Huang. 2007. Measuring sharedness of team-related knowledge: Design and validation of a shared mental model instrument. *Human Resource Development International* 10, 4 (2007), 437–454.

[75] Nick Jones and Matthew Moffitt. 2016. Ethical guidelines for mobile app development within health and mental health fields. *Professional Psychology: Research and Practice* 47, 2 (2016), 155.

[76] Benjamin Kaiser, Akos Csiszar, and Alexander Verl. 2018. Generative models for direct generation of cnc toolpaths. In *2018 25th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*. IEEE, 1–6.

[77] Anna Kantosalo and Tapio Takala. 2020. Five C's for Human-Computer Co-Creativity-An Update on Classical Creativity Perspectives.. In *ICCC*. 17–24.

[78] Amy K Karlson, Shamsi T Iqbal, Brian Meyers, Gonzalo Ramos, Kathy Lee, and John C Tang. 2010. Mobile taskflow in context: a screenshot study of smartphone usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009–2018.

[79] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[80] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1723–1734.

[81] Kamran Khowaja and Siti Salwah Salim. 2020. A framework to design vocabulary-based serious games for children with autism spectrum disorder (ASD). *Universal Access in the Information Society* 19, 4 (2020), 739–781.

[82] Huhn Kim. 2010. Effective organization of design guidelines reflecting designer's design strategies. *International Journal of Industrial Ergonomics* 40, 6 (2010), 669–688.

[83] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881* (2023).

[84] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[85] Tae Soo Kim, Arghya Sarkar, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. LMCanvas: Object-Oriented Interaction to Personalize Large Language Model-Powered Writing Environments. *arXiv preprint arXiv:2303.15125* (2023).

[86] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[87] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. 2020. Social acceptability in HCI: A survey of methods, measures, and design strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–19.

[88] Max Kreminski, Isaac Karth, Michael Mateas, and Noah Wardrip-Fruin. 2022. Evaluating Mixed-Initiative Creative Interfaces via Expressive Range Coverage Analysis.. In *IUI Workshops*. 34–45.

[89] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[90] Sri Kurniawan and Panayiotis Zaphiris. 2005. derived web design guidelines for older people. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*. 129–135.

[91] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* (2020).

[92] Barbara Leporini and Fabio Paternò. 2008. Applying web usability criteria for vision-impaired users: does it really improve task performance? *Intl. Journal of Human–Computer Interaction* 24, 1 (2008), 17–47.

[93] Clayton Lewis and Cathleen Wharton. 1997. Cognitive walkthroughs. In *Handbook of human-computer interaction*. Elsevier, 717–732.

[94] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. 2022. Understanding challenges for developers to create accurate privacy nutrition labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[95] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2022. 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] Yucheng Li, Deyuan Chen, Tianshi Li, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. 2022. Understanding iOS privacy nutrition labels: An exploratory large-scale analysis of app store data. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[97] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[98] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[99] Joseph Carl Robnett Licklider and Welden E Clark. 1962. On-line man-computer communication. In *Proceedings of the May 1-3, 1962, spring joint computer conference*. 113–128.

[100] Mark Liffiton, Brad Sheese, Jaromir Savelka, and Paul Denny. 2023. CodeHelp: Using Large Language Models with Guardrails for Scalable Support in Programming Classes. *arXiv preprint arXiv:2308.06921* (2023).

[101] Weng Marc Lim, Asanka Gunasekara, Jessica Leigh Pallant, Jason Ian Pallant, and Ekaterina Pechenkina. 2023. Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education* 21, 2 (2023), 100790.

[102] Vivian Liu. 2023. Beyond Text-to-Image: Multimodal Prompts to Explore Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.

[103] Vivian Liu and Lydia B Chilton. 2021. Neurosymbolic Generation of 3D Animal Shapes through Semantic Controls.. In *IUI Workshops*.

[104] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[105] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[106] Todd Lubart. 2005. How can computers be partners in the creative process: classification and commentary on the special issue. *International Journal of Human-Computer Studies* 63, 4-5 (2005), 365–369.

[107] Nicola Lucchi. 2023. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation* (2023), 1–23.

[108] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408* (2023).

[109] I Lupanda and JT Janse van Rensburg. 2021. Design Guidelines for Mobile Applications. In *International Conference Interfaces & Human Computer Interaction*. 92–99.

[110] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[111] Mary Lou Maher. 2012. Computational and collective creativity: Who's being creative?. In *ICCC*. Citeseer, 67–71.

[112] Mary Lou Maher, Justin D Weisz, Lydia B Chilton, Werner Geyer, and Hendrik Strobelt. 2023. HAI-GEN 2023: 4th Workshop on Human-AI Co-Creation with Generative Models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 190–192.

[113] Céline Mariage, Jean Vanderdonckt, and Costin Pribeanu. 2011. State of the Art of Web Usability Guidelines. In *Handbook of human factors in Web design*, Kim-Phuong L Vu and Robert W Proctor (Eds.). Crc Press.

[114] Christopher McComb, Peter Boatwright, and Jonathan Cagan. 2023. FOCUS AND MODALITY: DEFINING A ROADMAP TO FUTURE AI-HUMAN TEAMING IN DESIGN. *Proceedings of the Design Society* 3 (2023), 1905–1914.

[115] Donald McMillan, Alistair Morrison, and Matthew Chalmers. 2013. Categorised ethical guidelines for large scale mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1853–1862.

[116] Fadel M Megahed, Ying-Ju Chen, Joshua A Ferris, Sven Knoth, and L Allison Jones-Farmer. 2023. How generative ai models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study. *Quality Engineering* (2023), 1–29.

[117] Cade Metz. 2022. Meet GPT-3. It Has Learned to Code (and Blog and Argue). (Published 2020). https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html

[118] Lalit Mohan, Neeraj Mathur, and Y Raghu Reddy. 2015. Mobile App Usability Index (MAUI) for improving mobile banking adoption. In *2015 International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*. IEEE, 313–320.

[119] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[120] Michael Muller, Plamen Agelov, Hal Daume, Q Vera Liao, Nuria Oliver, David Piorkowski, et al. 2022. HCAI@NeurIPS 2022, Human Centered AI. In *Annual Conference on Neural Information Processing Systems*.

[121] Michael Muller, Heloisa Candello, and Justin Weisz. 2023. Interactional Co-Creativity of Human and AI in Analogy-Based Design. In *International Conference on Computational Creativity*.

[122] Michael Muller, Lydia B Chilton, Anna Kantosalo, Charles Patrick Martin, and Greg Walsh. 2022. GenAICHI: Generative AI and HCI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[123] Michael Muller, Justin D. Weisz, and Werner Geyer. 2020. Mixed initiative generative AI interfaces: An analytic framework for generative AI applications. ICCC 2020 Workshop, The Future of Co-Creative Systems. https://computationalcreativity.net/workshops/cocreative-iccc20/papers/Future_of_co-creative_systems_185.pdf

[124] Stacey F Nagata. 2003. Multitasking and interruptions during mobile web tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. SAGE Publications Sage CA: Los Angeles, CA, 1341–1345.

[125] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 152–158.

[126] Jakob Nielsen. 1995. Scenarios in discount usability engineering. In *Scenario-Based Design: Envisioning work and technology in system development*. 59–83.

[127] Jakob Nielsen. 2023. AI: First New UI Paradigm in 60 Years. *Nielsen Norman Group* (18 06 2023). https://www.nngroup.com/articles/ai-paradigm/

[128] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[129] Hugo Gonçalo Oliveira, Raquel Hervás, Alberto Díaz, and Pablo Gervás. 2014. Adapting a Generic Platform for Poetry Generation to Produce Spanish Poems.. In *ICCC*. 63–71.

[130] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[131] Jonas Oppenlaender. 2022. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988* 2 (2022).

[132] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[133] Google PAIR. 2021. *People+AI Guidebook*. Google. https://pair.withgoogle.com/guidebook/

[134] Kyudong Park, Taedong Goh, and Hyo-Jeong So. 2014. Toward accessible mobile application design: developing mobile application accessibility guidelines for people with visual impairment. *HCI Korea* (2014), 31–38.

[135] Alejandro Perez, Iaroslav Elistratov, Fynn Schmitt-Ulms, Ege Demir, Sadhana Lolla, Elaheh Ahmadi, and Alexander Amini. 2023. Risk-Aware Image Generation by Estimating and Propagating Uncertainty. (2023).

[136] Aryo Pinandito, Hanifah Muslimah Az-zahra, Lutfi Fanani, and Anggi Valeria Putri. 2017. Analysis of web content delivery effectiveness and efficiency in responsive web design using material design guidelines and User Centered Design. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*. IEEE, 435–441.

[137] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[138] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[139] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[140] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[141] Dagfinn Rømen and Dag Svanæs. 2012. Validating WCAG versions 1.0 and 2.0 through usability testing with disabled users. *Universal Access in the Information Society* 11 (2012), 375–385.

[142] John Rooksby, Parvin Asadzadeh, Alistair Morrison, Claire McCallum, Cindy Gray, and Matthew Chalmers. 2016. Implementing ethics for a mobile app deployment. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. 406–415.

[143] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *28th International Conference on Intelligent User Interfaces*.

[144] Mattias Rost and Sebastian Andreasson. 2023. Stable Walk: An interactive environment for exploring Stable Diffusion outputs. 3124 (2023). http://ceur-ws.org/Vol-3359/paper14.pdf

[145] Vildan Salikutluk, Dorothea Koert, and Frank Jäkel. 2023. Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 153–167.

[146] Jose Ma Santiago III, Richard Lance Parayno, Jordan Aiko Deja, and Briane Paul V Samson. 2023. Rolling the Dice: Imagining Generative AI as a Dungeons & Dragons Storytelling Companion. *arXiv preprint arXiv:2304.01860* (2023).

[147] Regina Schober. 2022. Passing the Turing test? AI generated poetry and posthuman creativity. *Artificial Intelligence and Human Enhancement: Affirmative and Critical Approaches in the Humanities* 21 (2022), 151.

[148] Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. 2020. Machines as teammates: A research agenda on AI in team collaboration. *Information & management* 57, 2 (2020), 103174.

[149] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. *arXiv preprint arXiv:2212.08061* (2022).

[150] Yashraj Shashikant Sharma. 2023. *Generating Wildfire Risk Maps for Critical Infrastructure Systems Using Integrated Generative AI and Simulation Techniques Under Information Uncertainty*. Ph. D. Dissertation. State University of New York at Buffalo.

[151] Maria Shitkova, Justus Holler, Tobias Heide, Nico Clever, and Jörg Becker. 2015. Towards usability guidelines for mobile websites and applications. *Wirtschaftsinformatik Proceedings* (2015). https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1106&context=wi2015

[152] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[153] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[154] Ben Shneiderman and Michael Muller. 2023. On AI Anthropomorphism. *Human-Centered AI (Medium)* (10 April 2023). https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae

[155] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the user interface: strategies for effective human-computer interaction.* Pearson.

[156] RC Sidorsky, RN Parrish, JL Gates, SJ Munger, and SYNECTICS CORP FAIRFAX VA. 1984. Design guidelines for user transactions with battlefield automated systems: Prototype for a handbook. *ARI Researoh Product* (1984), 84–08.

[157] Raymond C Sidorsky and Robert N Parrish. 1980. Guidelines and criteria for human-computer interface design of battlefield automated systems. In *Proceedings of the Human Factors Society Annual Meeting*, Vol. 24. SAGE Publications Sage CA: Los Angeles, CA, 98–102.

[158] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. 2023. The effectiveness of MAE pre-pretraining for billion-scale pretraining. *arXiv preprint arXiv:2303.13496* (2023).

[159] Sidney L Smith and Jane N Mosier. 1986. *Guidelines for designing user interface software.* Citeseer.

[160] Nikita Soni, Aishat Aloba, Kristen S Morga, Pamela J Wisniewski, and Lisa Anthony. 2019. A framework of touchscreen interaction design recommendations for children (tidrc) characterizing the gap between research evidence and design practice. In *Proceedings of the 18th ACM international conference on interaction design and children.* 419–431.

[161] Angie Spoto and Natalia Oleynik. 2017. Library of Mixed-Initiative Creative Interfaces. Retrieved 14-Aug-2023 from http://mici.codingconduct.cc/

[162] Ayushi Srivastava, Shivani Kapania, Anupriya Tuli, and Pushpendra Singh. 2021. Actionable UI Design Guidelines for Smartphone Applications Inclusive of Low-Literate Users. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.

[163] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[164] Luke Stark, Jen King, Xinru Page, Airi Lampinen, Jessica Vitak, Pamela Wisniewski, Tara Whalen, and Nathaniel Good. 2016. Bridging the gap between privacy by design and privacy in practice. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 3415–3422.

[165] Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics* 29, 1 (2022), 1146–1156.

[166] Jiao Sun, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces.* 212–228.

[167] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[168] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63.

[169] Josh Urban Davis, Fraser Anderson, Merten Stroetzel, Tovi Grossman, and George Fitzmaurice. 2021. Designing co-creative ai for virtual environments. In *Creativity and Cognition.* 1–11.

[170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[171] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Shomir Wilson, et al. 2023. Nationality Bias in Text Generation. *arXiv preprint arXiv:2302.02463* (2023).

[172] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[173] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–8.

[174] Benedikte Wallace, Charles P Martin, Jim Tørresen, and Kristian Nymoen. 2021. Learning embodied sound-motion mappings: Evaluating AI-generated dance improvisation. In *Creativity and Cognition.* 1–9.

[175] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2023. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *arXiv preprint arXiv:2307.10811* (2023).

[176] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–16.

[177] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–14.

[178] Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. 2023. Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861* (2023).

[179] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[180] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[181] Justin D Weisz, Mary Lou Maher, Hendrik Strobelt, Lydia B Chilton, David Bau, and Werner Geyer. 2022. HAI-GEN 2022: 3rd Workshop on Human-AI Co-Creation with Generative Models. In *27th International Conference on Intelligent User Interfaces*. 4–6.

[182] Justin D Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward general design principles for generative AI applications. In *Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE co-located with the ACM International Conference on Intelligent User Interfaces*, Vol. 3124. CEUR. http://ceur-ws.org/Vol-3359/paper14.pdf

[183] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces*. 402–412.

[184] Justin D Weisz, Michael Muller, Steven I Ross, Fernando Martinez, Stephanie Houde, Mayank Agarwal, Kartik Talamadupula, and John T Richards. 2022. Better together? an evaluation of ai-supported code translation. In *27th International Conference on Intelligent User Interfaces*. 369–391.

[185] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[186] Chathurika S Wickramasinghe, Daniel L Marino, Javier Grandio, and Milos Manic. 2020. Trustworthy AI development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, 130–136.

[187] Alan FT Winfield. 2018. Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI* 5 (2018), 75.

[188] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. 2020. A comparative analysis of industry human-AI interaction guidelines. *arXiv preprint arXiv:2010.11761* (2020).

[189] Sang Michael Xie and Sewon Min. 2022. How does in-context learning work? A framework for understanding the differences from traditional supervised learning. http://ai.stanford.edu/blog/understanding-incontext/

[190] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. *arXiv preprint arXiv:2309.03409* (2023).

[191] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[192] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[193] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[194] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

[195] Shuoyang Zheng. 2023. StyleGAN-Canvas: Augmenting StyleGAN3 for Real-Time Human-AI Co-Creation. (2023).

[196] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A EXTENDED DESCRIPTIONS AND EXAMPLES

We provide extended descriptions for each design principle and strategy to offer deeper insight into their meaning and application, written in second-person for an audience of design practitioners. We also provide examples for each design strategy to illustrate how they have been used in a realistic context. Each example was drawn either from a commercial generative AI application or an experimental generative AI system. Most examples were identified in the modified heuristic evaluation of Iteration 3; new examples were found for the three new strategies added after that iteration. For process-related strategies (denoted with an asterisk (*)), we discuss how the process would theoretically be applied as we did not have visibility into the actual design process for the generative AI applications we examined.

We make reference to the following commercial systems in our examples:

- Adobe Firefly: https://firefly.adobe.com
- Adobe Photoshop: https://www.adobe.com/products/photoshop.html
- AIVA: https://www.aiva.ai/
- ChatGPT: https://chat.openai.com
- DALL-E: https://labs.openai.com
- DreamStudio (powered by Stable Diffusion): https://dreamstudio.ai
- Github Copilot: https://github.com/features/copilot/

- Google Bard: https://bard.google.com
- Midjourney: https://midjourney.com

We also note that a single type of UX feature or functionality can be used to implement more than one design strategy. We have included similar or duplicate examples below to illustrate this point.

A.1 Design responsibly

The most important principle to follow when designing generative AI systems is to design responsibly. The use of all AI systems, including those that incorporate generative capabilities, may unfortunately lead to diverse forms of harms, especially for people in vulnerable situations. As designers, it is imperative that we adopt a socio-technical perspective toward designing responsibly: when technologists recommend new technical mechanisms to incorporate into a generative AI system, we should question how those mechanisms will improve the user's experience, provide them with new capabilities, or address their pain points.

A.1.1 Use a human-centered approach.* Technosolutionism is the idea that technology will solve all of our (human) problems, and it should be avoided at all costs. Human-centered approaches can determine whether the use of generative AI is appropriate; use cases that apply these technologies for their own sake may fail to deliver real user value.

Example: Human-centered approaches such as design thinking and participatory design allow you to observe users' workflows and pain points to ensure proposed uses of generative AI are aligned with users' actual needs. For example, involving stakeholders in the co-design of prototypes can serve as a probe for discussions around user value and technological feasibility.

A.1.2 Identify and resolve value tensions.* Often, there are multiple stakeholders involved in the creation of a generative AI application, including the end users, those who design and build the application (e.g. designers, developers, product managers), and those who make purchasing or licensing decisions (e.g. CIOs, CEOs). When these stakeholders' values are not aligned, it results in a value tension. Addressing these tensions is important for creating feasible and valuable products that meet end users' needs.

Example: Value Sensitive Design (VSD) [48] is a method that can help designers identify who the important stakeholders are and navigate the value tensions that exist across them.

A.1.3 Expose or limit emergent behaviors.* Generative AI can exhibit emergent behaviors — the ability to perform tasks beyond ones they were trained for. These emergent behaviors can be a delighter, such as when a conversational Q&A system answers an out-of-domain question, or they can be a risk, such as when that output is toxic or aggressive. As a designer, you should carefully consider the trade-offs between optimizing your user experience for a well-defined set of capabilities versus providing a more open-ended experience that may surface potentially risky emergent behaviors.

Example: Conversational interfaces that enable open-ended interactions will allow such emergent behaviors to surface. For example, a user may discover that ChatGPT can perform sentiment analysis, a task that it (likely) wasn't explicitly trained to do. By contrast, graphical user interfaces (GUIs), such as AIVA, can place limits on the ways a user can interact with the underlying generative model by only exposing selected functionality.

A.1.4 Test & monitor for user harms.* Generative models may produce a variety of harmful outputs, such as language that is hateful, abusive, profane, or otherwise toxic. They may also harm users by failing to produce outputs that provide a fair or accurate representation of diversity. It is imperative to work closely with technologists to understand and

evaluate the potential risks stemming from the use of a generative model in an application. It is also imperative to assume that user harms will occur and develop reporting and escalation mechanisms for when they do.

Example: One way to test for harms is by benchmarking models on known data sets of hate speech [60] and bias [45, 149, 171]. After deploying an application, harms can be flagged through mechanisms that allow users to report problematic model outputs.

A.2 Design for mental models

A mental model is a simplified representation of the world that people use to process new information and make predictions [80]. It is their own understanding of how something works and how their actions affect it. Generative AI poses new challenges to users, and designers must carefully consider how to impart useful mental models to their users to help them understand how a system works and how their actions affect that system. Also consider the user's background and goals and how to help the AI form a "mental model" of the user.

A.2.1 Orient the user to generative variability. Help the user understand the AI system's behavior, and that it may produce multiple, varied outputs that may not be reproducible, even when given the same input. This behavior will be unexpected for novice users because it is fundamentally different from traditional AI systems that always give the same outcome for the same input.

Example: Google Bard provides answers in the form of multiple drafts, indicating that it came up with multiple, varied answers for the same question.

A.2.2 Teach effective use. Users need to understand how to work effectively with a generative AI application to accomplish their goals. Mechanisms such as tutorials, examples, explanations, and social transparency [43] (i.e. showing other users' inputs and outputs) can help users form mental models for how to effectively use an application.

Example: DALL-E provides curated examples of generated outputs and the prompts used to generate them. Adobe Photoshop provides pop-ups and tooltips to introduce the user to its Generative Fill feature.

A.2.3 Understand the user's mental model.* Conduct evaluations, such as interviews, to determine whether a user has formed a useful mental model of a generative AI application. One prompt that can be useful to ask is how they think the application provides a certain capability, which forces the user to articulate their theory of how the system works. The goal is not for the user to possess an accurate model, but rather, one that is useful for working effectively with the system. Furthermore, understanding the user's mental model can also help you leverage their existing knowledge of similar applications to inform your design decisions.

Example: In evaluating a Q&A application, you might ask the user, "how did the system answer your question about who the current President is?" Answers such as, "it looked it up on the web" might indicate a need to educate users about hallucination issues. Users' existing mental models of other applications can also be useful to understand. For example, Github Copilot builds on users' mental models by following the same interaction pattern as its existing code completion features, which are familiar to many developers, hence easing their learning curve.

A.2.4 Teach the AI system about the user. LLMs are adept at tailoring their language to a target audience. Designers can induce these models to produce personalized responses to users – in essence, teaching the model about the user – by including additional prompt text such as, "explain like I'm five" or "please give me a detailed, technical answer." Capturing the user's expectations, behaviors, and preferences can improve the AI's interactions with them. Users can

also provide information about their background in UI outside of their prompt, which the application can then opaquely incorporate back into the prompt.

Example: ChatGPT provides a form for "Custom Instructions" in which users provide answers to questions such as, "Where are you based?", "What do you do for work?", and "What subjects can you talk about for hours?" In this way, users teach ChatGPT about themselves in order to receive more personalized responses.

A.3 Design for appropriate trust & reliance

Trustworthy generative AI applications are those that produce high-quality, useful, and (where applicable) factual outputs that are faithful to a source of truth. Calibrating users' trust is crucial for establishing appropriate reliance: teaching users to scrutinize a model's outputs for quality issues, inaccuracies, biases, underrepresentation, and other issues to determine whether they are acceptable (e.g. because they achieve a certain level of quality or veracity) or if they should be modified or rejected.

A.3.1 Calibrate trust using explanations. Be clear and upfront about what the application can and cannot do by explaining its capabilities and limitations. Teach users to be skeptical of potentially imperfect model outputs and help them understand when they can trust the system.

Example: ChatGPT explains its capabilities (e.g. "answer questions, help you learn, write code, brainstorm together") and limitations (e.g. "ChatGPT may give you inaccurate information. It's not intended to give advice.") directly on its introduction screen. Google Bard provides a notice below the prompt field that states, "Bard may display inaccurate or offensive information that doesn't represent Google's views."

A.3.2 Provide rationales for outputs. Show the user why a particular output was generated by showing the model's "chain of thought" [179] or identifying the source materials used to generate it. For example, identifying source documents for answers expected to be factually correct or revealing the image sets that a text-to-image model was trained on can help users calibrate their trust.

Example: Google Bard provides a list of sources it used to produce answers to questions. Adobe discloses that its Generative Fill feature was trained on "stock imagery, openly licensed work, and public domain content where the copyright has expired" [1].

A.3.3 Use friction to avoid overreliance. Encourage the user to review and think critically about the generative model's outputs by introducing mechanisms that slow them down at key decision-making points. These mechanisms are known as cognitive forcing functions [22]. Examples include offering multiple AI-generated options for the user to select from, highlighting uncertainty, or requiring the user to create content or render a decision before showing the AI's output.

Example: Google Bard displays multiple drafts for the user to review, which can encourage them to slow down and consider which drafts may be of lower or higher quality.

A.3.4 Signify the role of the AI. AI systems may act in different roles, such as "tools," "partners," "analytics," or "coaches" [114]. These roles shape users' expectations of how the AI system fits into their workflow, such as the extent to which it takes initiative (e.g. by acting proactively vs. reactively) and agency (e.g. by directly manipulating an artifact vs. making suggestions or recommendations). The role that is signified shapes users' perceptions of the system: research by Kim et al. [84] shows that AI "tools" are viewed as less genuine and caring than AI "mediators" or AI "assistants." Anthropomorphic signifiers such as giving a human name to an AI, representing the AI with a human-like avatar, having the AI refer to itself using first-person pronouns, or showing an animated typing bubble to hide inference latency

may all give users the false impression that they are interacting with a human. Although there is no clear consensus on the extent to which AI-infused user experiences should favor or discourage anthropomorphism [154], as a designer, it is crucial to clearly signify to the user when they are interacting with an AI system and what content was AI-generated.

Example: Github Copilot's tagline is "Your AI pair programmer", which elicits the role of a partner. Copilot fulfills this role by proactively making suggestions as the user writes code. It also possesses a limited form of agency by making autocompletion suggestions directly in the user's code editor, although it requires the user to explicitly accept or reject those suggestions (e.g. by pressing tab or escape).

A.4 Design for generative variability

One distinguishing characteristic of generative AI systems is that they can produce multiple outputs that vary in character or quality, even when the user's input does not change. This characteristic raises important design considerations: to what extent should multiple outputs be visible to users, and how might we help users organize and select amongst varied outputs?

A.4.1 Leverage multiple outputs. Take advantage of multiple outputs to help users produce the one that fits their needs. Multiple outputs can be exposed to the user or remain under the hood and instead allow the model to select the best option(s) to surface.

Example: DreamStudio, DALL-E, and Midjourney all generate multiple distinct outputs for a given prompt; for example, DreamStudio produces four images by default and can be configured to produce up to 10. ChatGPT allows the user to regenerate a response to see more options.

A.4.2 Visualize the user's journey. Users may not be able to reproduce prior outputs. Capturing and displaying the user's history of outputs, along with the input parameters used (e.g. prompts, controls) can help them track their work. Also consider ways to guide the user to new output possibilities that they have not explored.

Example: DreamStudio, DALL-E, and Midjourney all show a history of the user's inputs and resulting image outputs. Research prototypes extend the idea of "visualizing the user's journey" even further using visualization techniques to show unexplored parts of an output space [88] (using dots overlaid on 2D histograms) or of a parameter configuration space [144] (by showing configurations a user has and has not yet tried in a grid, with untried configurations rendered as placeholders).

A.4.3 Enable curation & annotation. When a generative model produces multiple outputs, users may need to curate or annotate them. Curation may include collecting, filtering, or organizing outputs (possibly from the generation history). Annotation may include the ability to tag artifacts (e.g. "I like this picture") or make notes within an artifact (e.g. "this line of code looks suspicious").

Example: DALL-E allows the user to mark images as favorites and store them within groups called collections. Users may create and name multiple public or private collections to organize their work.

A.4.4 Draw attention to differences or variations across outputs. A generative model can sometimes produce a set of similar outputs that are difficult to tell apart. When outputs are similar, tools that aid users in identifying the similarities and differences between multiple outputs can be useful.

Example: DreamStudio, DALL-E, and Midjourney all display multiple outputs in a grid-like fashion to allow the user to identify differences, but fine-grained differences between outputs are not explicitly highlighted. A prototype source

code translation interface by Weisz et al. [183] visualizes the differences across multiple generated code translations through granular highlights, as well as interactively through a list of alternate translations.

A.5 Design for co-creation

Generative AI offers new co-creative capabilities. Help the user create outputs that meet their needs by providing controls that enable them to influence the generative process and work collaboratively with the AI.

A.5.1 Help the user craft effective outcome specifications. Generative AI has introduced a new interaction paradigm, intent-based outcome specification, in which users specify what they want but not how it should be produced. Orient the user to this new paradigm and assist them in prompting effectively to produce outputs that fit their needs.

Example: Google Bard sends out a newsletter that includes a section called "Prompt Engineering 101," which features tips and examples to help users improve their prompt writing. The IBM watsonx.ai Prompt Lab documentation includes a set of tips and examples to help the user understand how to improve their prompts.

A.5.2 Provide generic input parameters. Let the user control generic aspects of the generative process such as the number of outputs and the random seed used to produce those outputs. Generic controls apply across most use cases, independent of which model is used.

Example: DreamStudio provides a slider for users to indicate the number of images they want to produce for a given prompt, along with an input field for random seed.

A.5.3 Provide controls relevant to the use case and technology. Let the user control parameters specific to their use case, domain, or model architecture. Some model architectures provide specialized means of control, such as semantic sliders for latent space models [103] or decoding strategy and temperature for LLMs. Other models have controls specific to the domain of the application (e.g. code, music, art).

Example: AIVA allows the user to customize domain-specific characteristics of the musical compositions it generates, such as the type of ensemble and emotion.

A.5.4 Support co-editing of generated outputs. Allow both the user and the AI system to improve generated outputs. Although the user should retain agency and decision-making authority, the AI can co-create with the user to improve outputs, such as by providing suggestions, organizing outputs, or editing an artifact.

Example: Adobe Photoshop exposes generative AI capabilities within the same design surface as its other image editing tools, enabling both the user and the generative AI model to co-edit an image.

A.6 Design for imperfection

Users must understand that generative model outputs may be imperfect according to objective metrics (e.g. untruthful or misleading answers, violations of prompt specifications) or subjective metrics (e.g. the user doesn't like the output). Provide transparency by identifying or highlighting possible imperfections, and help the user understand and work with outputs that may not align with their expectations.

A.6.1 Make uncertainty visible. Caution the user that outputs may not align with their expectations and identify detectable uncertainties or flaws. Provide disclaimers about the potential for imperfection, show the model's confidence level when possible, and utilize "I don't know" responses when confidence is low.

Example: Google Bard's interface states, "Bard may display inaccurate info, including about people, so double-check its responses." This disclaimer alerts the user to the possibility of uncertainties or imperfections in its outputs. A prototype source code translation interface proposed by Weisz et al. [183] makes the generative model's uncertainty visible to the user by highlighting source code tokens based on the degree to which the underlying model is confident that they were correctly translated. These highlights help guide the user's attention toward places that require their review.

A.6.2 Evaluate outputs using domain-specific metrics. In some cases, the quality of a generative model's outputs can be assessed with measurable criteria. For example, answers to customer queries can be evaluated for faithfulness to source documents, and design mock-ups can be evaluated for adherence to a UI style guide.

Example: Molecular candidates generated by CogMol [29], a prototype generative application for drug design, are evaluated with a molecular simulator to compute domain-specific attributes such as molecular weight, water solubility, and toxicity.

A.6.3 Offer ways to improve outputs. Provide ways for the user to fix imperfections and improve output quality, such as editing tools, an option to regenerate, or providing alternative outputs to select from.

Example: DALL-E and DreamStudio allow users to refine outputs by erasing and regenerating parts of an image (inpainting) or generating new parts of the image beyond its boundaries (outpainting). Google Bard offers options for the user to modify outputs to be shorter, longer, simpler, more casual, or more professional.

A.6.4 Provide feedback mechanisms. Allow users to provide feedback on the quality of a model's output. Feedback may be implicit (e.g. the user makes edits to a generated artifact indicating potential issues) or it may be explicit (e.g. rating the quality with a thumbs up / thumbs down).

Example: ChatGPT offers an option for the user to provide a thumbs up or thumbs down rating for its responses, along with open-ended textual feedback.