

Next Steps for Human-Centered Generative AI: A Technical Perspective

XIANG ‘ANTHONY’ CHEN, UCLA HCI Research,
JEFF BURKE, UCLA REMAP,
RUOFEI DU, Google Research,
MATTHEW K. HONG, Toyota Research Institute,
JENNIFER JACOBS, UCSB,
PHILIPPE LABAN, Salesforce Research,
DINGZEYU LI, Adobe Research,
NANYUN PENG, Computer Science Department, UCLA,
KARL D.D. WILLIS, Autodesk Research,
CHIEN-SHENG WU, Salesforce Research,
BOLEI ZHOU, Computer Science Department, UCLA,

Through iterative, cross-disciplinary discussions, we define and propose next-steps for Human-centered Generative AI (HGAI) from a technical perspective. We contribute a roadmap that lays out future directions of Generative AI spanning three levels: aligning with human values; accommodating humans’ expression of intents; and augmenting humans’ abilities in a collaborative workflow. This roadmap intends to draw interdisciplinary research teams to a comprehensive list of emergent ideas in HGAI, identifying their interested topics while maintaining a coherent big picture of the future work landscape.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**.

Additional Key Words and Phrases: Generative AI, Human-Centered Design

ACM Reference Format:

Xiang ‘Anthony’ Chen, Jeff Burke, Ruofei Du, Matthew K. Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl D.D. Willis, Chien-Sheng Wu, and Bolei Zhou. 2018. Next Steps for Human-Centered Generative AI: A Technical Perspective. In . ACM, New York, NY, USA, 34 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

The recent development of Generative AI—ranging from large language models (LLMs) [25, 140] to visual generation techniques [94, 109, 149]—promises to revolutionize how humans work in a wide range of tasks [103]. Meanwhile, various research communities, will soon, if not already, be working on topics related to Generative AI. HCI, in particular, as an interdisciplinary field, has the opportunity to serve as a nexus that connects multiple disciplines related to Generative AI.

To support such emergent research across disciplines, this paper proposes Human-centered Generative AI (HGAI, pronounced ‘H’-/gai/) as an overarching topic and lays out specific next steps for achieving HGAI. Our focus is on identifying joint HGAI research opportunities across related technical disciplines, mainly including technical HCI research [69], machine learning, natural

and note-taker who focused on prompting the participant to elaborate, clarify, and broaden their ideas. We intentionally avoided delving into each idea as we focused on breadth in this iteration while leaving deeper discussions later. Each 1:1 discussion lasted between 45 minutes to an hour: all but one discussion was conducted in-person.

After the discussions, the first author summarized the notes of each discussion into a list of research agenda topics with brief descriptions. Further, we aggregated participants' proposed definitions of HGAI and divided them into three levels of interpretations as detailed later in §3.

2.2 Iteration #2: Paired Discussion

Next, we conducted five discussions that each involved the first author as the moderator and two other participants. In all but one discussion, the two participants were split between academia and industry. In all discussions, the two participants had different areas of expertise. The main purpose was to identify interdisciplinary HGAI research opportunities, such as common problems shared by multiple disciplines and innovative system designs by combining multiple disciplinary elements. Each discussion revolved around three research agenda topics selected based on the previous discussions: we first selected topics both mentioned by the two participants, which resulted in either one or two topics; then we selected the remaining one or two topics amongst the ones with the most extensive discussions in the previous iteration (measured by the amount of notes). For each topic, after briefly describing what was discussed before, we asked the participants to think further about prior work related to the topic, the gap in said topic, and specific research activities to pursue the topic. Each discussion was conducted remotely and lasted for about 45 minutes.

After the discussions, the first author followed the Affinity Diagram approach to organize notes from the previous two discussions into tree-like structures: each research agenda topic was the mid-level node, whose low-level nodes consisted of prior work or specific research activities for future work, and each top-level node was a theme to connect multiple related topics.

2.3 Iteration #3: Virtual Walk-the-Wall Discussion

Finally, we asked each participant to walk through the Affinity Diagram laid out on a shared document. Participants could add to the low-level nodes, suggest or edit research agenda topics at the mid-level, or add or re-organize the top-level themes. We tracked the changes made by each participant and the first author facilitated ad hoc discussions whenever there were conflicted changes or disagreements amongst participants. This walk-the-wall discussion took place asynchronously over a period of 15 days, after which the first author finalized the additions and changes to create a clean version of the Affinity Diagram.

Below we present our finalized collective ideas, starting with our definition of HGAI across three levels, followed by detailed discussions of next-steps within each level.

3 DEFINING HUMAN-CENTERED GENERATIVE AI (HGAI)

We start with defining key terminologies below.

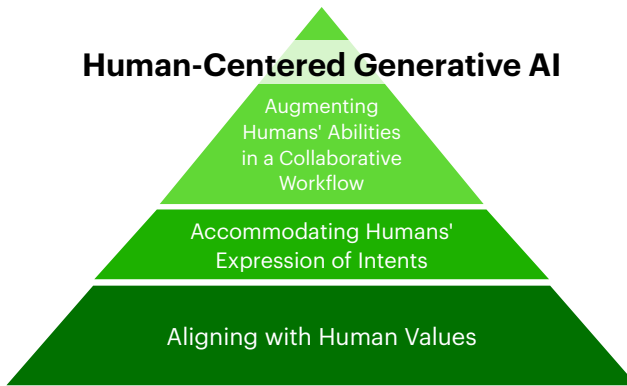
What do we mean by “Generative AI”? Commonly, Generative AI is defined as AI models that can generate new data instances, which contrasts with Discriminative AI that aims at distinguishing between different kinds of data instances [3]. Although the notion of Generative AI as defined above can be fairly broad and can date back to early work (e.g., Topology Optimization [112] proposed in the early 90's), the majority of our discussions were concerned with the recent developments of data-driven Generative AI, e.g., LLMs and text-to-image generations.

What do we mean by “Human” in HGAI? There are various stakeholders involved in the ecosystem of Generative AI, including

- (1) people whose data is used for model training (e.g., artists' and designers' work),
- (2) people who label and moderate the data (e.g., to filter out toxic contents [5]),
- (3) people who develop Generative AI models (e.g., academic professors/students and employees in tech companies who build LLMs),
- (4) people who develop systems that use Generative AI models (e.g., game development platforms that use LLMs to enable conversational characters),
- (5) end-users of Generative AI and its applications,
- (6) and finally, people who are impacted (in)directly by various (un)intended consequences of Generative AI (e.g., teachers grading AI-generated essays).

Our discussion of HGAI primarily focuses on the betterment of individuals who own the data and the end-users of Generative AI, although some of our next-step ideas and calls-for-actions do speak to the other stakeholders as well.

What do we mean by “Human-centered Generative AI”? We propose the following definition and then differentiate it from related concepts in prior work.



Definition

Human-centered Generative AI (HGAI) should achieve three levels of human-centered objectives: (i) Aligning with human values; (ii) Accommodating humans' expression of intents; and (iii) Augmenting humans' abilities in a collaborative workflow.

Fig. 2. Our definition of Human-centered Generative AI (HGAI) across three levels.

Related concepts in prior work. Human-centeredness is not unique or specific to AI or Generative AI. In [32], Chancellor surveyed and identified several prior framings and definitions, e.g., distinguishing from work that mainly focused on the technical aspect [64], that “must take account of varied social units that structure work and information” [77], and that asked what should be, rather than could be, produced [57], and that related to the “social-technical gap” [13] in CSCW literature. Building off of such historical contexts, Chancellor then proposed a renewed definition of human-centeredness in machine learning as following a set of practices to achieve balances between technical innovation and human and social concerns. The above evolution of human-centeredness is in line with Level 1 in our definition centered on human values and purposes.

Another recent paper by Capel and Bereton [29] surveyed how human-centeredness has been interpreted in various subfields, e.g., Explainable and Interpretable AI, Human-Centered Approaches

to Design and Evaluate AI, Humans Teaming with AI, and Ethical AI, based on which they proposed a new definition: “Human-Centered Artificial Intelligence utilizes data to empower and enable its human users, while revealing its underlying values, biases, limitations, and the ethics of its data gathering and algorithms to foster ethical, interactive, and contestable use.” This definition echos both Level 1 and 3 in our definition where we further include Level 2 that is unique to Generative AI as humans need to express and realize their intents to control what contents will be generated. Related to our Level 1 definition of HGAI, a recent talk by Pascale Fung [28] laid out different types of harms caused by LLMs ranked by severity, ranging from offensive and biased language, to deepfake and discriminatory generation, to law-breaking privacy violation and misinformation, and to life-threatening acts such as medical misdiagnosis and terrorism.

On the industry side, OpenAI has been conducting alignment research [10], which is defined as “engineering a scalable training signal for very smart AI systems that is aligned with human intent”, including training AI systems using human feedback, to assist human evaluation, and to perform alignment research (*e.g.*, generating explanations of LLMs [20]). Such effort is mostly related to our Level 2 definition of HGAI.

Below we provide an overview of each level of HGAI (Figure 2), each of which necessitates and builds off of previous level(s).

3.1 Aligning with Human Values

The foundational objective of HGAI is the alignment with human values. We consider values as the fundamental beliefs that define the ethics of Generative AI. Admittedly, there is no easy way to define a one-size-fits-all value system for the entire humanity¹; instead, human values often vary across cultures and regions and should be carefully considered with respect to the specific stakeholders of Generative AI.

Alignment with human values has been widely discussed in value-sensitive design [55] and, more recently, in addressing “black box” AI’s violation of human values [39, 129]. To align Discriminative AI with human values, one example is preventing racial biases when performing facial recognition [30]; in a similar vein, Generative AI that aligns with human values should not generate racially-biased images of human faces when given certain text prompts. Note that the need for alignment does not mean there exist some universal human values to be aligned with. Generative AI’s behaviors need to consider values of specific population groups involved in the model’s development and usage and acknowledge that a solution without trade-offs might not exist. For example, Generative AI might align well with small business owners by helping them costlessly create artworks or slogans for advertisement; yet, in the mean time, such generated contents might have displaced or violated artists’ rights to profit from their work, thus misaligning with artists’ values.

To ensure HGAI’s alignment with human values, simply training a larger model is not enough [104]; we argue that we should follow a human-centered process, not only for designing systems that utilize Generative AI, but further for creating such Generative AI models in the first place. Human-centered design is a well-established body of methods to ensure that a system will actually benefit the stakeholders it intends to serve. However, besides benefiting its intended users (*e.g.*, the aforementioned small business owners), Generative AI should also prevent causing harm to affected people (*e.g.*, artists whose work has been incorporated into the model). Thus, a human-centered process for Generative AI should be extended from design activities to the model training and development stages, particularly by involving people who might be negatively impacted by the resultant model. In §4, we discuss various HGAI next-steps throughout this process.

¹Although there have been multiple ongoing efforts of unification, *e.g.*, the Blueprint for an AI Bill of Rights [4] and PCAST Working Group on Generative AI [11].

3.2 Accommodating Humans' Expression of Intent

Unlike Discriminative AI where the input is some existing information (*e.g.*, an image for object recognition or some text for summarization), input to Generative AI is much more open-ended.

Foremost, HGAI should offer an expressive medium through which users can freely and effectively convey their intents of generating certain contents. Existing approaches, such as text prompts, might be a convenient shortcut to convey intent; however, it remains limited in many scenarios where text alone either cannot clearly represent a user's intent or does not allow a user to iteratively refine their intent expression.

Further, HGAI should ensure that the generative process follows a user's expressed intent. Since the generative model is often trained on massive amounts of data on the Internet, it could produce uncontrollable behaviors that deviate from a user's intent (*e.g.*, hallucinated contents [71]). As such, HGAI should provide explicit control mechanisms for a user to steer the generative process or allow a user to "talk back" to the Generative AI by editing an imperfect result.

Meanwhile, enabling user control of Generative AI is a balance act. We want to provide the appropriate amount of control to the end-users: too much control and the tool is less accessible, whereas too little control makes the model output unsteerable by the user. Finding the right balance of control in each application setting is important, as demonstrated in some recent work on controlling Generative Adversarial Networks (GANs) [45].

Note that it is not always the case that humans have clear and strong intents that can be articulated in any form (text prompts or otherwise). Sometimes implicit control mechanisms (*e.g.*, learning preferences from past interactions) are important too. HGAI should explore the diversity of control mechanisms that range from implicit to explicit, text to rich media formats, which adapt to human intents in-the-moment because intents might be uncertain, constantly-evolving, and perhaps best supported in a mixed-initiative manner [67].

Finally, it is important to realize that not all human intents are benign (*e.g.*, using LLMs to fabricate false news); therefore, HGAI should foremost align with human values (Level 1) before committing to realizing human intents (Level 2).

3.3 Augmenting Humans' Abilities in a Collaborative Workflow

Generative AI that can accommodate users' intent expression should then aim to augment humans' abilities in achieving some overall goals. Despite the promises of Generative AI, there often remains a gap between what the AI model can generate and how such AI can actually benefit a domain user's work. For example, consider OpenAI's Codex—an LLM capable of generating functional code snippets: such a model alone might not be beneficial to a programmer who works in conventional integrated development environments (IDEs). In contrast, GitHub's Copilot—an LLM with similar code generation capabilities—is fully integrated with programmers' IDEs. A human-centered example of such integration is allowing programmers to control AI-generated completions of their in-progress programs—one line *vs.* multiple lines of code—a feature that is considerate of programmers' work practices and goes beyond generating code alone.

To further close the gap, HGAI should divide the labor in ways match what the human and AI each does best. For example, consider video production. Perhaps the human is best as the director who asks GPT to write the script and keep the scene settings and back stories consistent, which is often challenging for human screenwriters. As another example, consider novice users interacting with Generative AI for visual design tasks. As these users are probably unfamiliar with the best choice of terminology in a text prompt, HGAI can start with guiding the user to formulate the scope of what visual contents they want to create, then incrementally brainstorm examples to populate a set of candidates, and then help the user continuously narrow down and refine their choices.

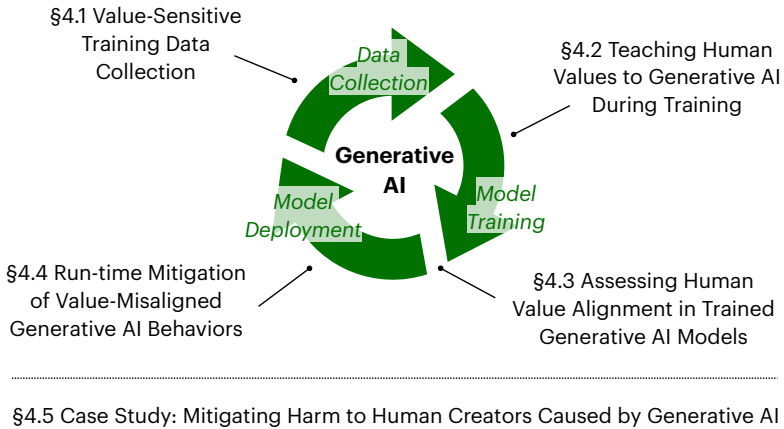


Fig. 3. Overview of HGAI Level 1: next-steps in aligning with human values.

4 NEXT-STEPS FOR HGAI: ALIGNING WITH HUMAN VALUES

As shown in Figure 3, this section lays out next-steps for HGAI throughout the Generative AI lifecycle, from collecting training data, to training models, to assessing trained models, and to run-time mechanisms. Further, we zero in on a case study of mitigating Generative AI’s harm to human creators.

4.1 Value-Sensitive Training Data Collection

Since modern AI is mostly data-driven, a main culprit of ethical issues is the training data. For Discriminative AI, some training data might cause AI to aim at the wrong target [102], such as using “income” to determine a credit score because there is no other better attributes, *e.g.*, “credit worthiness”. To mitigate such limitations engendered in training data, one approach is simply making the process of collection and the composition of training data as transparent as possible, *e.g.*, via a specification document like a datasheet [58]. For HGAI, the next-steps should foster a value-sensitive training data collection process.

4.1.1 Preventing training data from giving rise to biases. Since Generative AI aims to learn the distribution of a certain domain to generate new data instances, it is even more critical to ensure that the training dataset is unbiased. The ever-increasing size of required dataset and its breadth goes well beyond a traditional (non-generative) learning problem, making it even harder to achieve this. Meanwhile, we should recognize that sometimes there is no universally unbiased dataset or solution; therefore, when determining biases, it is important to contextualize a dataset by which population group its resultant model aims to serve. Thus the next-steps of HGAI should aim at the following:

- Expanding the well-established user-centered [76] and task-centered [82] design process to encompass the early stage of data collection to ensure no misrepresentation of the affected user populations will propagate to downstream models and systems. Fei-Fei Li mentioned that, before the start of a recent project in her lab to benchmark robotic tasks, their team conducted a large-scale user study to identify the winning task most beneficial to the target users, which became the focus of the project [6]. Based on such user-centered practices, the next-step is to formulate and evaluate a generalizable protocol to guide the training data collection process prior to developing Generative AI.

- Rather than suppressing potential biases so the model would never learn such behaviors, another direction for next steps is to develop Generative AI that is aware of biases already existing in the real world. Consider how a text-to-image AI might learn biases and generate predominantly male images for *CEO*. Since such biases stem from the gender imbalance in the executive world, it might be worth for Generative AI to learn about such phenomena, so that the model not only will prevent generating biased images but can further explain to the user how it unbiases the generation or even let the user control such processes.
- As a biased Generative AI model's output might permeate into the real world (e.g., people using ChatGPT to write various sorts of documents), it will soon become imperative to stop such biased AI-generated data from being used to train future models. According to Veselovsky *et al.*, 33-46% of crowd workers are estimated to use LLMs in completing their tasks [131]. One next-step is to incorporate ongoing efforts that detect AI-generated contents (e.g., [97]) into the screening of training data.

4.1.2 Preventing creators' data from being used for training Generative AI. In a recent discussion, Pamela Samuelson described latest legal cases and the challenge of determining what constitutes infringement in the context of AI-generated contents [41]. Similar to how the General Data Protection Regulation (GDPR) influenced a wide range of changes in how technology handles user data (e.g., cookies on websites), we can anticipate changes in Generative AI systems as new advances on the legal front take place, such as adding Adversarial noises [116] or "certified" watermarks [16] to prevent unobstructed usage of artists' work. Next-steps for HGAI are as follows.

- One grand challenge is enabling creators to protect their works in public domains from being used for training Generative AI, which requires the entire industry and research community to establish new protocols of data collection. For example, Adobe Firefly was trained with data either out of copyright, licensed for training, or in the Adobe Stock library. Similar to how open source licenses enable and regulate how one's code can be used, we need to develop similar mechanisms for writers, artists and designers to specify permitted usage of their work. For example, one possible mechanism is that artists who opt-in to contribute their work for training a model can have access to that model's generated contents for their future projects.
- Alternatively, Generative AI developers should allow creators to audit existing training data with tools to help them identify whether their work has been inadvertently included. For example, "Have I Been Trained"² made an important step to help artists detect whether their work is in public datasets like LAION-5B³.
- User interfaces of Generative AI should inform end-users whether the underlying model has been trained indiscriminately on public domain data or restricted to a (much) smaller set of examples with creators' permission. As the latter model is likely to suffer from inferior generative quality, end-users should be informed of such trade-offs, allowing them to select and compare different Generative AI models' performance.

4.2 Teaching Human Values to Generative AI During Training

Beyond learning from labeled data, prior work has discussed and demonstrated interactive machine learning approaches [53] of teaching Discriminative AI models [122, 148]. However, teaching machines beyond labels remains a rather under-investigated problem and below we discuss some starting points of teaching human values to Generative AI.

²Have I Been Trained: <https://haveibeentrained.com>

³LAION-5B: <https://laion.ai/blog/laion-5b>

4.2.1 Defining value-sensitive metrics and reward functions. Researchers have found that western, educated, industrialized, rich, and democratic (WEIRD) populations have been dominating the participant groups in behavior science [65] and HCI [86]. A similar issue also exists in some Generative AI models that perform unequally. For example, it has been shown that using ChatGPT as a Question-Answer tool works to various degree [126]: the success rate seems higher for users in developed countries than those in developing countries, *e.g.*, asking for a recipe for making Western vs. non-Western cuisines. Such inequality in performance causes a viscous cycle: populations who benefit less from Generative AI will become less engaged and contribute less, resulting in future models to under-serve them even more severely. Although it is possible to mitigate such inequality via fine-tuning large base models with imitation data, a recent paper has found such approaches still fall short in closing the gap of what is unsupported by the base models [62]. Some next-steps for HGAI are as follows.

- We should study how Generative AI developers currently are aware of and how they overcome the performance inequality issues. For example, one recently-proposed approach is Constitutional AI [15]—the development of AI that complies with explicitly written human rights and ethical principles, such as privacy, transparency, accountability, and non-discrimination. It would be interesting to learn how such an approach works in practice amongst Generative AI developers.
- Rather than assuming there is an invisible, one-size-fits-all objective function to define human values, we should introduce population-specific metrics in the loop of training Generative AI models. Via a collaboration between disciplines, we can approach this goal from both a software engineering perspective (how to efficiently build multiple generative models tailored to specific populations?) and a machine learning methodological perspective (is it possible to incorporate multiple population-specific objectives into the development of one model?)
- One opportunity is to work with social scientists to develop an ontology to better define, categorize, and quantify ethics and value related issues in Generative AI. Such an ontology will inform HGAI research with a comprehensive and hierarchical view, based on which we can better conduct systematic studies, targeted data collections, and develop mitigation methods to address ethical issues in Generative AI. Using ontology as a tool, AI researchers can better identify more subtle fairness issues [24, 120] and more accurately measure the extent to which the models are biased.

4.2.2 Adding controls to Generative AI. Knowing what biases already exist in Generative AI, we can devise targeted controllable generation methods to mitigate such biases, *e.g.*, by inducing negative biases and positive biases for another demographic, or by equalizing biases between demographics [119]. As another example, using a constrained decoding technique, it is possible to limit the generation of Ad Hominem language that targets some features of a person’s character instead of the position the person is maintaining [120]. Some next-steps for HGAI include the following.

- Currently, such controls are administered as part of the model training process and future work can develop new interfaces to let end-users interactively manipulate such controls. For example, we can enable end-users to interactively reduce biases in images created by Generated Adversarial Networks, *e.g.*, by selecting additional images to balance the proportion between genders or adjusting weights assigned to images [52].

4.3 Assessing Human Value Alignment in Trained Generative AI Models

Despite extensive testing in the lab, Generative AI deployed to the real world still likely to behave suboptimally and unexpectedly; as such, next-steps for HGAI can perform risk assessment or auditing of models post-training.

4.3.1 Risk assessment of Generative AI models. After a Generative AI model is developed, there remains important work of risk assessment for each application that utilizes the model, trying to foresee its ethics-related impact. Performing such risk assessments could be resource-demanding and there are no standardized approaches at present.

- One next-step could be research and studies of risk assessment methods. For example, low-risk applications can employ check-lists to observe how much the generated contents violate some pre-defined rules. High-risk applications may require costly virtual sandboxing experiments to contain possible riskier actions before public release, such as generative approaches to perform medical diagnoses or controlling field robots.
- Inspired by some recent work on generative agents for simulating human behavior [106], another next-step is to develop toolkits that support custom simulative experiments with generative AI in virtual environments to elicit possible problematic behaviors.

4.3.2 Auditing Generative AI models. As Generative AI is frequently updated with significant changes, it is important to perform repeated timely assessment. One existing solution is integrating assessment capabilities with the model development, such as a tool suite built on top of TensorFlow Model Analysis that can be used to compute and visualize commonly-identified fairness metrics for classification models, *e.g.*, false positive and negative rates [1].

One related approach is auditing algorithms—“a method of repeatedly querying an algorithm and observing its output in order to draw conclusions about the algorithm’s opaque inner workings and possible external impact” [93]. One recent example of this approach is polling different demographic groups to measure how language models underrepresent or misalign with them [113]. Generative AI presents unique challenges to perform such audits due to the sheer amount and variety of generated contents. Further, it is often unclear how to track or assess whether certain changes in the Generative AI result in better or worse contents, such as writing or artwork where the assessment can be subjective and unscalable if requiring human involvement.

To make it possible to audit generative AI, there are several next-steps:

- Collecting datasets of changing generated contents due to model updates and performing analyses to identify unexpected changes, thus the need to perform audits. In some domains, *e.g.*, news, some past benchmarks might become part of the future training set, thus it is important to prevent such overlaps along the time axis.
- Conducting studies to understand how Generative AI developers and end-users currently are aware of and cope with model updates and changes. Such work requires longitudinal studies or building online community support for users to report unexpected changes over time.
- Developing toolkits—both developers and end-users facing—to support auditing (*e.g.*, the AuditNLG library [2]), including curating a set of benchmarks, defining criteria, reporting audit outcomes, and troubleshooting unexpected changes (*e.g.*, certain types of generated contents start to show biases). Such tools can build on recent work that shows the possibilities of using language model automate evaluations over time to track changing behaviors [107].

4.4 Run-time Mitigation of Value-Misaligned Generative AI Behaviors

Even after a model is deployed to an application and in the end-users’ hands, there are still opportunities to put more guard rails at run-time on Generative AI’s behaviors. Admittedly, these

solutions simply mitigate inappropriate model behaviors but do not fundamentally correct such models that violate human values (e.g., trained with biased data).

4.4.1 Informing users of possible unethical behavior. Model card—a document that details the model’s intended use, the data it was trained on, its accuracy, and potential biases—has been a popular approach to inform the public of an AI model’s performance and potential limitations, and to increase transparency and accountability in AI development. To support the creation of model cards, there have been toolkits integrated with the AI development pipeline, such as Google’s Model Card Toolkit [7]. To further broaden engagement in the community, the Model Card Authoring Toolkit provides a tool that helps members of a community to review and choose from a range of machine learning models based on their shared values, by providing assistance in understanding, navigating, and evaluating the models [117].

- For Generative AI, at present, there is a lack of discussion about how model cards should be different than Discriminative AI’s. Thus one next-step is to perform a bottom-up study of developers’ existing approaches of creating Generative AI model cards, engage other stakeholders (e.g., content creators and end-users) to elicit their feedback on such model cards, and formulate metrics and standards to guide the best practices. One unique challenge to tackle is that documenting Generative AI’s problematic output: how to provide a set of samples with sufficient coverage without overwhelming the readers?

4.4.2 Explaining the probabilistic nature of Generative AI. Spearheaded by DARPA’s initiative [63], a plethora of research has thrived on eXplainable AI (XAI), most of which assumes the context of Discriminative AI that outputs decisions rather than contents. As such, there has been scarce discussion of how to define and enable explainability for Generative AI. One recent work focused on code generation and took a scenario-based approach to elicit developers’ needs for explanation when using Generative AI in various programming scenarios: natural language to code, code translation, and code auto-completion [127]. Some next-steps for HGAI are as follows:

- Conducting formative studies to understand end-users’ needs for explaining Generative AI: when they need explanations and how they act on explanations? For examples, one hypothesis might be the need for explanation when certain prompts do not result in desired generated contents; a successful explanation, in turn, should enable a user to improve their prompts, *i.e.*, higher satisfaction with the new generated contents.
- Studying whether and how existing XAI techniques apply to Generative AI and identifying the gap. Based on numerous taxonomies of XAI [124], we can attempt to draw analogies to the generative domains, e.g., how counterfactual techniques [43] can be redefined in the scenario of explaining text-to-image prompts.
- Complementary to the above approach that starts from XAI literature, we can also explore techniques that explain the output of Generative AI via participatory design and technology probe [70], spanning multiple generated modalities e.g., text, image, and audio.
- Implementing and evaluating explanation techniques in the contexts of representative application scenarios, e.g., programming, writing, and visual design. Similar to how plug-ins enable ChatGPT to access specific domain information, future Generative AI interfaces can provide explanation plug-ins to promote a wide range of available techniques at end-users’ disposal.

4.4.3 Detecting and disabling inappropriate generated contents. Given how multiple generative AI models might generate the same types of inappropriate contents (e.g., images that contain gender stereotypes), one next-step for HGAI is to enable users to detect such inappropriate contents, specifically:

- For models that support such detection, we can employ existing classifiers and leverage LLMs to self-detect consistency to human value (e.g., whether the just-generated text contains toxic contents [134]).
- For integrating such detection into the user interface, we should carefully study the impact of detection models' performance: in particular, false negatives might cause users to inadvertently use inappropriate generated contents in downstream tasks.
- Another important challenge is to ensure that such detection models align with end-users' values, such as allowing them to program-by-example and specify what contents should be considered inappropriate.
- One next-step following detection is allowing users to disable inappropriate or undesirable content generation, such as disabling the chatbot from talking about politics [8] or disabling age-inappropriate elements when generating stories for children.

4.4.4 Augmenting input and filtering output. Given how it is unrealistic to change or even just fine-tune a model at run-time, next-steps for HGAI can instead focus on augmenting the input and filtering output to achieve value-aligned generated contents.

- Analogous to how data augmentation [130] can artificially increase the size and diversity of a dataset, one next-step is to develop techniques to augment a user's input to Generative AI (e.g., text prompts) to preempt the generation of biased contents. For example, by inferring from a user's prompt that race could be a latent biased variable, we can append additional terms that request more racially-diverse output. Then, as the model returns a large number of results, we can present a subset of randomly-sampled results to the user.
- Another type of output filtering might also mitigate the issue of appropriating other creators' work. Similar to how generating pseudocode can inform a user to implement certain program without directly using others' code, one next-step could be generating 'pseudo artwork' or 'pseudo writing' that represents a creator's style without appropriating their work. There are two key considerations here to pursue such techniques: (i) keeping creators in the loop—we should learn from creators what constitute a good piece of 'pseudo work'; and (ii) keeping users engaged—a system should provide sufficient tool support for a user to create their own versions of an artist's or a writer's work by following and mimicking their 'pseudo work'.
- Meanwhile, it is also necessary to address the trade-offs of the above approaches, such as increased system latency. The system should also make it transparent in how it augments the user's input or filters the model's output, and further let the user have control over such mechanisms, such as setting the number of requested samples to manage latency.

4.5 Case Study: Mitigating Harm to Human Creators Caused by Generative AI

To end our discussion of HGAI Level 1, we focus on a case study to address one of the most concerning issues of Generative AI: how it causes harm to human creators, e.g., artists and designers. It has been widely recognized that creators' work in public domains were being scrapped for training Generative AI, causing various kinds of harm to both individuals and the community as a whole [9]. Although our case study here mainly focuses on artists and designers to allow for a deep discussion, other creators are also affected in similar ways, e.g., software engineers whose open-sourced code was scrapped by Generative AI and used inappropriately elsewhere.

4.5.1 Attributing elements of generated contents to the work of creators'. As mentioned above, HGAI can learn from practices in open-source software communities. However, unlike source code that structurally follows certain programming languages, other generative domains, such as painting

and music composition, make it much harder to enable content attribution. As such, some next-steps for HGAI include:

- Analogous to how NLP models summarize text, we can attempt to develop both abstractive and extractive attribution models: the former aims to provide a high-level description of how the generated contents overall can be attributed to certain creators' work while the latter highlight specific elements and map them to certain creators' example work to indicate attribution;
- From a user interface perspective, we should also study the ambiguity and scalability of the above approach. (i) Ambiguity: how to visualize AI-generated contents being partially influenced by a creator's style or a combination of multiple creators'? (ii) Scalability: suppose AI-generated contents mimic a large number of creators' work, e.g., a 'remix' on Spotify that merge numerous musicians' work, is it still useful to show attribution and how to avoid overwhelming end-users?
- One natural next-step built on the two above is allowing end-users to remove certain elements from AI's generated contents to avoid imitating other creators' work. We can provide explicit controls for end-users to limit what AI can generate or provide tool support to let them develop their own style in lieu of some creators'.

4.5.2 Involving creators in the process of developing Generative AI systems. Despite the harm that has already been done, it seems highly likely that Generative AI will continue to play a major role in the art and design community. As such, some next-steps for HGAI should aim to allow for symbiotic co-existence between Generative AI and human creators.

- Going beyond how traditional user-centered design methods ensure a system design provides values to end-users, we should employ similar participatory design methods that also involve creators so that a Generative AI system can both provide values to end-users while minimizing harm done to the creators.
- Revisiting value-sensitive design [55] as the values and incentives of the HCI and AI community are probably very different from the values of professional artists. Building less harmful artist-oriented AI technologies requires broadening or redefining our value sets within the HGAI community.
- If the intended end-user of a Generative AI system is the creator, we should explore interface and interaction designs that go beyond prompting. Most generative models take text or RGB pixels as input, likely due to technical convenience. However, artists might possess a much larger set of creative methods, e.g., brush strokes, vocals, camera framing, and poetry, and there needs to be a deeper understanding in how Generative AI can support such idiosyncratic creation process.

4.5.3 Providing an educational platform for creators to explore usages of Generative AI. On the positive side, AI holds the promise to help designers and artists to automate the tedious and repetitive parts of their job. For example, well before robust background removal AI, background or "green screen" removal involves lots of human labor from video artists and editors. Later, as the technology matured, artists and editors who embraced AI tools and learned their pros and cons became more efficient at their work.

- As recent generative AI has brought forth lots of new capabilities, creators might find it challenging to catch up with the fast paces. Thus one next-step for HGAI is to provide an educational platform for creators to demystify Generative AI and explore how to best integrate it into their work.

- Exploring the optimal way to collaborate with Generative AI should not be left to each individual human creator; rather, we should study the best practices (e.g., from artists who are also familiar with Generative AI) and provide tool support for the less tech-savvy creators, e.g., tracking and comparing how their work evolves in the course of invoking Generative AI's assistance.

4.5.4 Ensuring Generative AI developers follow guidelines aimed at protecting designers/artists. Further, even if we provide guidelines for human-centered methods of AI development for art and design, it remains unclear whether and how researchers, independent developers, and the open source AI community would follow these guidelines. Some next-steps for HGAI include—

- Starting with our own academic communities, we can attempt to set up norms and guidelines that prescribe ways in which researchers might consider using (or not using) training data that contains creators' work. We should discuss the degree to which efforts to assess risk and impact of using Generative AI should be documented in research publications or judged in peer review, similar to how NeurIPS 2020 started asking authors to include a section that discusses the broader impact of their work. Perhaps beyond just a statement alone, research that claims to develop Generative AI tools to serve creators should be judged by whether and how there are partnerships with communities of artists as a component of the project.
- Once the above guidelines mature in the research communities, we should aim to extend them to the developer communities. Given how it has become so much easier for individual or hobbyist programmers to fine-tune Generative AI and build their own applications, we should conduct studies to understand their current practices and whether and how developers would follow such guidelines.
- Building on the aforementioned educational platforms, we can extend the scope to build and study an online community that allows creators and developers to better communicate their work, respectively. Guidelines for development can be embodied in such creator-developer communication. Developers can get inspirations from creators what Generative AI tools are interesting to build, creators can guide developers to collect training data, and the tools can be developed and tested via a closed-loop collaboration between the two groups.

While the proposed efforts to prescribe guidelines and approaches for reducing harm are important, we need to acknowledge the reality that Generative AI has already caused harm to many professional artists and quite likely presents an existential threat to entire artistic professions, such as illustration and graphic design. Thus it is likely that creators in the art and design community might have already formed a reasonable skepticism when some Generative AI tools promise to benefit individual artists. Overcoming such established aversion is an indispensable part of the next-steps for HGAI. Further, learning from some recent reflection on empathic approaches in accessibility research [18], it is important not to assume or oversimplify the need or accomplishment of engaging with creators. Following best practices that promote benefits while reducing harm to creators should be implemented and assessed as rigorously as the research or development of models. For example, for academic authors, an impact statement is perhaps more appropriately included in the limitation section, acknowledging what has actually been done to mitigate harm, whether such measures are evaluated, and limitations of the effect.

5 NEXT-STEPS FOR HGAI: ACCOMMODATING HUMANS' EXPRESSION OF INTENTS

Perhaps the expression of intent is Generative AI's most distinguishing factor from Discriminative AI. The problem is that intent is ambiguous. Some might argue that the popular approach of text-to-generated-contents already works quite well as it successfully mimics how humans universally communicate intents with each other using language. However, even humans' expression and

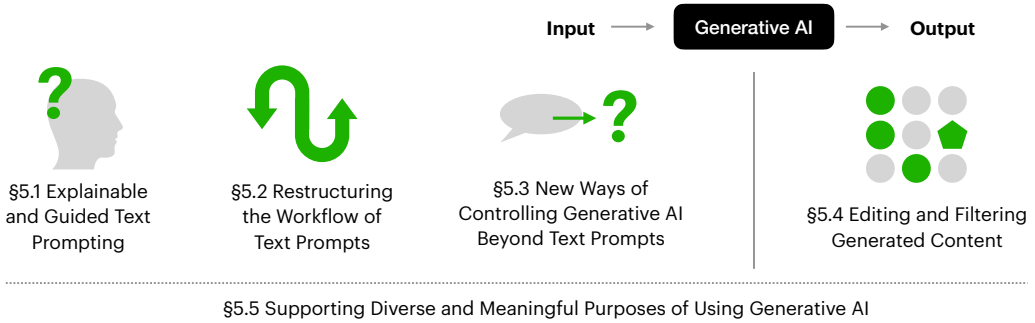


Fig. 4. Overview of HGAI Level 2: next-steps in accommodating human's expression of intents.

understanding of each other's intents is inherently ambiguous and often not perfect—how can we expect to do it better with Generative AI?

One obvious solution for improvement is providing better media for intent expression, such as combining multiple modalities: text prompt, sketch, and gesture. However, the ambiguity of intent could be much more fundamental in that the human might not really know what they want (Generative AI) to create in the first place. As such, it might be useful to maintain a continuous conversational between users and Generative AI, rather than expecting to arrive at ideal results in one-shot attempts. Some intents are implicit, assumed by the human but often unspoken. For example, a user generating a furniture with a computer-aided design (CAD) tool (e.g., [37, 114]) might look good on the screen yet they also implicitly expect it to look equally good when manufactured and placed in the intended environment without realizing that the Generative AI does not know about the manufacturing process or what the environment is like.

To address the inherent challenge of ambiguity in intent expression, our next-steps for HGAI span both input and output of Generative AI (Figure 4) with an emphasis on rethinking the dominant use of text prompts.

5.1 Explainable and Guided Text Prompting

Given the overwhelming popularity of text prompting in numerous Generative AI scenarios, some future efforts on HGAI should be devoted to better supporting such input with explanation and guidance, which can be helpful for both end-users and developers.

5.1.1 Enabling end-users to understand and manipulate input/output relationship. End-users often do not know how good is the text prompt they use in getting Generative AI to produce the result they are looking for. Similar to how adversarial examples lead to unexpected errors in Discriminative AI [59], the analogous issue exists in Generative AI when the changes a user makes in the text prompt fails to produce the changes they expect to see in the generated contents or, worse, produces undesirable changes. The opaque or unexplainable relationship between text prompt input and generated output often leads end-users to run unguided trials that drains their time and wastes computing resources. As a result, users might have a hard time establishing trust and willingness to accepted generated contents [26]. Some next-steps for HGAI include—

- More studies should be conducted to understand how humans use prompts to interact with Generative AI, such as when writing text [44] or conversing with a chatbot [143]. Such studies should aim to provide concrete evidence that complements the currently anecdotal understanding of prompting and to further connect with Generative AI researchers and developers to inform their model-building work.

- Generative AI systems should provide tutorials and examples that educate end-users about the non-deterministic behavior of Generative AI and manage their expectation, thus maintaining a reasonable level of user’s trust in the model while preventing unguided prompt engineering.
- Developing feedforward techniques [17] to visualize what certain edits in a text prompt might lead to changes in the generated contents. Although there have been studies on the effects and trade-offs of such feedforward controls (using pre-generated examples) [45], it remains unclear how to implement such techniques at interactive speed without requiring pre-generated examples.
- Similar to feedforward, borrowing the autocompletion approach in text entry [36], we can develop techniques to suggest words or phrases following a user’s partial text prompt and optionally show what contents will be generated if provided with the completed text prompt. Importantly, the user should be able to scroll through multiple autocompletion candidates to explore which one best fits their intent.
- For explaining Discriminative AI, past work has employed attention models and saliency maps (e.g., [123]) to indicate which parts of the input is “responsible” for certain output. Analogously, we can develop and integrate similar approaches for Generative AI, e.g., the Cross Attention approach [128], allowing the user to explore and understand how their text prompt is associated with the generated output. Further, we can develop techniques for the user to directly manipulate the output (e.g., certain parts of a generated image) and see, inversely, how the input text prompt changes.
- Enabling controlled generation (discussed later in this section) can also contribute to the user’s understanding as it allows a user to manipulate specific generation-controlling modules and see its causal effect on the generated results.

5.1.2 Enabling developers to discover and troubleshoot problematic input/output relationships. Beyond obtaining a satisfactory result, it is also important for a Generative AI model to respond appropriately to certain changes in input. Similar to the general concept of sensitivity analysis, some recent work has found suboptimal or problematic input-output response relationship in Generative AI, e.g., the orders in which training samples are provided result in drastically different performance [90] and over half the tokens in a prompt can be removed while maintaining or even improving model performance [141]. Building off of these findings, one next-step for HGAI is—

- Developing tool support for sensitivity analyses of Generative AI that enables model or application developers to surface problematic responses that might otherwise go unnoticed. Such tools can provide a useful dashboard that monitors and visualizes model responses given a set of benchmark input perturbation.

5.2 Restructuring the Workflow of Text Prompts

Going beyond the monolithic “prompt-revise-repeat” cycle, HGAI research should propose alternate workflows that allow a user to break down a complex generative task or to efficiently iterate on suboptimal generated contents. Here we focus on discussing one specific alternate workflow—the insertion-based control approach.

5.2.1 Iterative insertion-based control of Generative AI. Rather than a one-shot text prompt input, a new workflow can start with something simple, which then allows for iterative insertion of additional input to extend or refine generated contents. InsNet is one technical solution for such insertion-based control where the method generates sentences in random orders by inserting tokens to existing partial contexts [89]. Another approach employs smaller models to control larger

models where users can, for example, stipulate that they want three specific words or a specific style to appear in the generated text [46, 92, 145]. Next-steps for HGAI along this direction include:

- Building off of recent work on chaining prompts [139] and models [49], we can develop more varieties of mixed-initiative workflow for natural language generation. For example, a user writing a story can start with a single word, *e.g.*, “flower”. Next, the system prompts the user to describe how they feel about flowers, *e.g.*, “I love flowers”, which then allows the natural language generation methods to insert new elements to create longer and more sentences. At each turn, the user can also insert their own elements before AI takes over. In the meantime, the user interface displays the history of iterations showing how the text “evolves”, thus allowing the user to roll back if the expansion has taken an undesirable direction.
- In the image generation domain, we can develop a workflow where a user starts with a simple text prompt and Generative AI returns an image. Next, the system generates a textual description based on the image and extract words associated with visual elements of the image. The user can then edit the generated image by manipulating the corresponding words in the textual description. The LUCSS demo⁴ showcases a proof-of-concept prototype related to this approach, where a user can colorize an initially black-and-white generated image via manipulating color attributes in the textual description [150].

5.3 New Ways of Controlling Generative AI Beyond Text Prompts

Currently, the interaction design of prompting mainly assumes that humans initiate and AI reacts, while ignoring other types of possibilities (*e.g.*, human-machine co-creativity and mixed initiatives), as theoretically constructed as a $2 \times 2 \times 2$ design space in [85]. Although text prompts leverage humans’ familiarity with using language to express intents, in the meantime, it also limits other forms of expression that also exist in human-human communication, such as visual language, gestures, and facial expression. For example, research in psychology [91] has found that gesture is more than an auxiliary aid to speech but rather an integral part of human communication and that it is closely linked to thought, working together with speech to develop meanings. An even more fundamental issue is, regardless of what forms of expressions are available, users themselves might not always know their intent (*i.e.*, what exactly they want to create). It is possible that a user’s intent evolves and clarifies itself as they iteratively attempt to express it to Generative AI and to revise the input based on the generated contents. Current text prompting interfaces have not been specifically designed to support such intent exploration—a successfully “engineered” prompt often looks remotely like humans’ natural language expression and is challenging for non-expert users to come up with [143].

5.3.1 Spatial and gestural input to control Generative AI. In contrast to the “1D” text prompt, input to Generative AI can leverage an additional degree of freedom and we should explore how 2D or 3D techniques can allow users to express their intents in various domains of creation. Specific next-steps for HGAI include:

- Existing techniques, *e.g.*, ControlNet [146] allows a user to condition image generation with additional images, *e.g.*, one with edge detection to provide a skeletal “template” for the model to fill in generated details. Building on and going beyond this approach, we can study what kinds of templates a user would create to express their intents to the model, which will likely lead to new features currently unsupported by Generative AI models. For example, in the medical domain, a pathologist specifying the generation of synthetic histopathological images

⁴LUCSS: Language-based User-customized Colorization of Scene Sketches (LUCSS): <https://youtu.be/IsBdrXtU0MI>

of tumor cells might provide very different kinds of sketches than an architect outlining a new concept of office buildings.

- Expanding input to 3D, we can expect crosspollination with gestures and augmented or virtual reality (AR/VR). Beyond using generated contents to construct the AR/VR world [68], another challenge and opportunity is enabling users to create generated 3D objects (*e.g.*, using Shape-E [12]) such as furniture or art installations. For furniture design, one direction is to integrate some existing techniques (*e.g.*, freehand gestures [66] and 2D+3D sketching [14]) with text prompting and the latest Generative AI models.
- Even in the natural language domain, *e.g.*, writing, prior work has demonstrated the possibility of using a brush-like input to directly manipulate key attributes of the story, such as the fortune of the protagonist character [40]. One next-step is to explore how users can control other parameters in text generation using 2D input techniques, including the usage of a dashboard that presents comprehensive key parameters of generated text using data visualization techniques [96].

5.3.2 Controlling Generative AI with implicit intents. Some intents are implicit and naturally unspoken, assumed to be understood by others where appropriate actions should take place accordingly. Examples include contexts, activities, and personal history. Building on recent developments such as zero-shot multimodal reasoning [144], next-steps in HGAI should aim to incorporate these implicit intents:

- Leveraging a large body of work on context-aware computing [115] and activity recognition [35], we can incorporate additional information as representations of a user’s intents. Such an approach can be useful when employing Generative AI to automate physical tasks (*e.g.*, via controlling a robot or an Internet-of-Things). For example, as a user wakes up and walks towards the kitchen, the time and location contexts can inform the Generative AI model to reason that the user might want to make coffee and proactively generate next-step actions to start the coffee machine.
- Some user input or edits might represent implicit intents that Generative AI needs to incorporate when creating certain contents. For example, consider a furniture designer using a CAD tool to put together the tabletop and three legs, which carries the implicit intent that these legs need to maintain contact with the tabletop. As such, Generative AI that morphs the shape of the tabletop should also reposition the legs to maintain contact.
- Personal history (*e.g.*, daily routines) can also be helpful, such as in the above automating coffee-making example. Although generative conversational agents like ChatGPT or Bing Chat do retain certain history, much more work needs to focus on how domain users utilize history in their work, such as programmers debugging a large codebase or doctors examining a patient, which should, in turn, inform the development of new Generative AI models that can leverage such historical data to improve generated contents, whether it is a code snippet or a summary of patient history.

5.4 Editing and Filtering Generated Content

Whereas the above discussion in this section focuses on reinventing the input techniques, below we switch gears to consider end-user editing and filtering of Generative AI’s output. When a user is unsatisfied with the generated contents, rather than restarting the whole generation process to obtain a new sample, it would be more efficient for the user to specify what is not right and directly edit the generated contents.

5.4.1 Semantic control of generated contents. Prior work has demonstrated “smart” edits where a user’s drawing or erasing of the generated design translates into new constraints that steer the AI to generate a new version while addressing the user’s intent [37]. A series of similar approaches have emerged when working with GANs, from providing sliders to adjust various attributes factorized from the latent space such as pose and texture [118] to allowing for directly dragging the generated image [105, 133]. Building upon these methods, some next-steps for HGAI are—

- It is important to provide such semantic controls with continuous feedback at interactive speed, which might require new interpolative and approximative techniques beyond just generating whole brand-new contents every time a user dials the knob. We can conduct studies and analyze the “knob-dialing” behavior of end-users when provided with semantic controls and understand what is the minimum amount of feedforward information we can provide to inform users’ control without adding latency to the system.
- A more fundamental approach is to change what AI generates: not pixels or tokens but semantic controls, which can offer users not just a static result but also a tool to access a large space of alternative contents. Recent work that simplifies an image into highly abstract yet representative sketches [132] shows promises of providing such semantic controls as the outcome of the generative process. Relatedly, Videomap lets a user perform video editing by navigating on a 2D view of the latent space [84]. Generating controls beyond texts or pixels requires both technical breakthroughs as well as studies to verify that such controls would allow a user to realize their intents without too much cognitive load.
- Given how generated contents are likely to contain imperfect or flawed elements (*e.g.*, blurry faces in generated images), we should develop techniques for users to fix such issues as directly and quickly as possible. Prior work has demonstrated some user-driven steps for fixing entanglement issues in GAN [52], such as specifying regions on the generated images that should be disentangled. Similarly, we can employ other techniques to support direct fixes, such as using image inpainting techniques [142] to redraw blurry faces.
- In the natural language domain, we can explore techniques to present generated output (*e.g.*, a chatbot’s response) in a more editable way to collect users’ immediate feedback that informs the next iteration. For example, consider summarizing a news article. The output summary can come with “+/-” buttons to adjust its length or allows a user to highlight portion of the text as “important/unimportant” so the model can create a new summary accordingly. A similar approach has been demonstrated in a text reader where a user’s highlights in an article can steer the summary to put more emphasis on the highlighted texts [38].

5.4.2 Filtering generated contents. Alternative to directly editing the generated content, a user can express their intents via selecting which types of results they prefer over the others, which sends a signal to the model for generating more relevant contents in future iterations. Such approaches have demonstrated expressive power when the user is faced with a large number of data points, such as document collection [42] and GAN editing directions [51]. Specific next-steps for HGAI include—

- Developing methods to represent and extract user intents from the selection they perform, which the Generative AI model then incorporates as a signal into the subsequent generation of new contents.
- Conducting studies of such a user-driven coarse-to-fine selection of generated contents to compare key metrics (*e.g.*, content quality, cognitive load, user satisfaction) with conventional approaches that involve repeated generation with prompt tweaking.

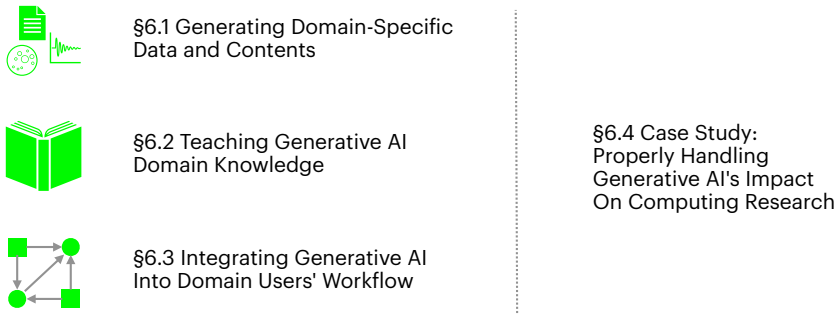


Fig. 5. Overview of HGAI Level 3: next-steps in augmenting humans' abilities in a collaborative workflow.

- Assessing the potential risk of generating a wide range of unexpected content for users to choose from, some of which might fall way out of distribution and only serve as noises and some might even include inappropriate elements such as toxic language or disinformation.

5.5 Supporting Diverse and Meaningful Purposes of Using Generative AI

Finally, human intents of using Generative AI would also be influenced by what Generative AI is capable of. In contrast to AI-generated contents as commodity (*e.g.*, generating an attention-grabbing video optimized for virality rather than deep meaning), one next-step for HGAI is—

- Exploring designs that support more diverse and meaningful reasons for humans to use AI-generated contents (*e.g.*, generating a video to tell the life story of a family member). Consider using Generative AI for communication, such as parents telling bedtime stories to children. To support such communicative purposes, the focus of Generative AI should go beyond producing a stereotypical story and aim to support conveying cultural meanings, cultivating familial relationship, or even allowing parents to teach children a specific lesson metaphorically via storytelling.

6 NEXT STEPS FOR HGAI: AUGMENTING HUMANS' ABILITIES IN A COLLABORATIVE WORKFLOW

Generative AI is more than a computational model or a tool; it will significantly change how professionals are able to work. For example, filmmakers might use Generative AI in the video editor to create a shot they forget to take. With such a capability, filmmakers no longer need to take footages exhaustively and worry about missing some shots. Further, Generative AI will change how people perceive their profession just like how algorithmic automation has been changing work in many fields [125]. For example, an illustrator might rethink what it means to create a piece of work given how Generative AI can automate partially or even mostly what they do. Generative AI will blur the boundaries between professions. One example is the blending of art and engineering: people who are familiar with the inner-working of Generative AI and good at “prompt engineering” can explore the creation of artworks whereas artists can tap into the engineering world to harness the power of industry-scale models via fine-tuning.

This section's discussion of HGAI next-steps (Figure 5) covers domain-specific data and content generation, teaching domain knowledge to Generative AI, and integrating Generative AI into domain users' workflow. We further discuss implication of Generative AI on computing research.

6.1 Generating Domain-Specific Data and Contents

One important way for Generative AI to augment domain users' abilities is to help them overcome the hurdles of acquiring data, which could be costly and time-consuming due to data scarcity and limited resources.

6.1.1 Generating 3D objects and scenes. Although there exist solutions to generate 3D objects and scenes (e.g., [72, 138]), there remain gaps in making such generation useful for specific domain users. We discuss a few exemplar cases below and their next-steps for HGAI.

- Consider generating traffic data for training self-driving systems. Current Generative AI cannot generate safety-critical traffic scenarios because the model learns to fit a data distribution and sampling from such a model can result in the most probable synthetic sample with limited training value. In many applications, self-driving systems care about corner cases or low-tail samples, e.g., accident-prone traffic scenarios. One important next-step is to enable Generative AI to “extrapolate” and synthesize less frequent but safety-critical scenarios. The same concept can benefit LLMs and text-to-image models to provide even more creative output, instead of giving an average answer based on its large pool of training data.
- Consider generating 3D objects for digital design and fabrication. Currently, Generative AI can only create static 3D models (i.e., point clouds or meshes) that cannot be easily modified to fit domain users' different needs. One next-step is to generate machine code (e.g., G-code) that drives a fabrication machine to create an object so that domain users can modify the code for custom designs. One such example is generating 3D printed hair [80]: rather than generating the geometry of hair, it is more customizable to generate G-code where a user can directly manipulate key parameters, such as length, thickness, and curliness.
- Consider generating architectural designs such as floorplan layout and furniture arrangement. Beyond the current approaches that focus on generating static plans [56, 100], one next-step is also generating a simulation [106] of how people interact with each other within the space to better inform architects and designers to further iterate on their work.

6.1.2 Generating medical data. Data has always been both the fuel and the bottleneck of medical AI development due to the data scarcity of certain diseases as well as the high cost of collection, processing, and labeling. Multiple opportunities and challenges exist for the future for HGAI:

- Given the recent development of synthetic medical data generation (e.g., in histopathology [48] and radiology [31]), it is time to study the effects of using such synthetic data in downstream tasks, especially on medical AI models' performance on out-of-distribution datasets as well as how doctors and patients perceive such models knowing the training data is “not real”. Further, to address accountability issues, it is important to allow doctors to trace an AI error to potentially problematic synthetic data and verify its factual correctness [98].
- Besides medical data of different pathologies, it is also possible to use Generative AI to create “synthetic” control patients in clinical trials [6]. One related challenge is providing tool support for experimenters to carefully control Generative AI when manipulating parameters of “synthetic” control patients and to generate reports with transparency to fully inform policymakers the limitations and potential risks.

6.1.3 Generating contents across traditional boundaries of formats. For art and design, one opportunity and need is to support media objects and systems that can naturally cross traditional boundaries (physical/digital, 2D/3D, interactive/static, raster/vector, authored/generated). Specific next-steps for HGAI are as follows:

- As mentioned in the previous section, image generation should consider providing users with editable contents beyond static pixels, *e.g.*, vector graphics or design tool files (*e.g.*, .ps and .ai). One additional benefit of this approach is that users can perform edits on these files, which offers process-oriented information to instruct AI how to generate next-iteration contents to better meet the user’s needs.
- An even further goal is for AI to define and generate an “invariant representation” of an object that can be malleably converted into a wide range of formats so that users can readily import the generated contents into their work using different tools. Consider multimedia industry creating a new character that spans comic books, animations, movies, theme parks, and video games. Generative AI will create the “core” of the character, which is then expanded to different media. Any future updates or additions to the character will also be automatically and consistently reflected in individual types of media. The benefits of this approach are that (i) Generative AI can obtain and take in feedback from heterogenous types of users (*e.g.*, comic book readers, movie viewers, and theme park goers) and that (ii) Generative AI can serve as the nexus connecting different departments within a company or industry to co-develop the character.

6.2 Teaching Generative AI Domain Knowledge

Current Generative AI models indiscriminately scrap and learn from data on the Internet without explicit recognition of domain knowledge. As a result, some types of generative contents are problematic. While synthesizing texts from a book without knowing the subject matter might still produce sensible writing [101], synthesizing pixels from images of human hands does not work as well due to the ignorance of anatomy [33]. There are numerous other cases where Generative AI’s lack of domain knowledge will cause performance issues, *e.g.*, generating drug designs, protein structures, molecular models, building codes, and industrial manufacturing equipment.

One popular approach to overcome the knowledge gap is in-context learning [95] where an end-user provides an example (*e.g.*, an existing story) and asks Generative AI to produce something similar (*e.g.*, a new story of the same genre). However, this approach likely will not work across many other domains, *e.g.*, generating a W2 tax form that requires much more knowledge available in a single example. Alternatively, retrievable augmented generative models [147] can adapt to highly specialized domains. We can still train general-purpose large models but have small, domain-specific knowledge bases to retrieve from and use the retrieved results to augment the black-box general language models to quickly adapt to new domains.

Below we discuss next-steps to teach Generative AI two types of domain knowledge: “what” and “how”.

6.2.1 Teaching Generative AI “what-knowledge” by concepts. One common way to represent knowledge is hierarchical concepts, such as the description and organization of medical conditions [22]. Concepts can be thought of as “what-knowledge” as it informs us what makes an object what it is and different from the others, *e.g.*, defining a human hand by its constituent parts as well as their spatial relationship. Currently, Generative AI is oblivious of concepts, *e.g.*, able to generate an image of a bike but unaware of the correspondence between different parts of the image to specific parts of the bike. Research on Discriminative AI has realized the importance of teaching concepts to a model [27, 34, 74, 79], mainly for interpretability purposes and to ensure that AI is right for the right reason [111]. Meanwhile, concept teaching in Generative AI remains a nascent research topic with some next-steps as follows.

- Similar to how Concept Activation Vector [74] uses positive and negative examples to represent a concept, we can develop Generative AI models that can follow user-specified concepts.

GANravel is a tool that employs this approach, letting users select example images to unbias GAN’s image generation [52]. Examples can more effectively and intuitively represent a user’s domain knowledge where textual expression falls short (*e.g.*, due to inherent vagueness or under-defined terms), such as pathologists describing pathognomonic—visual features of certain types of tumor cell.

- Rather than having multiple models learning different modalities of data, we should develop Generative AI that learns symbolic concepts (*e.g.*, door), which can be represented equivalently in different types of generative contents (*e.g.*, an abstract icon, a photo-realistic image, the sound of a door opening, mechanical behavior of door knob and hinges). In this way, Generative AI can produce comprehensive contents necessary for the user’s task, such as a video showing a person opening a door. To achieve this, one challenge is the need for pairs of data (contents and the constituent concepts) and one solution would be using CLIP [108] to construct a concept dataset from images to texts.
- The recent development on image segmentation [75] shows promises in “dissecting” static images into components, which can support concept learning. However, we should aim further to establish hierarchical relationships between parts. In so doing, the model can learn if different parts are functionally similar, *e.g.*, scissors and a knife both have blades. Building on the “library learning” approach in program synthesis (*i.e.*, discovering library components or subroutines in a program with semantic meaning) [50, 137], we can explore Generative AI that learns to generate instructions of creating certain contents (*e.g.*, G-code) and to extract concepts represented by components within such instructions.

6.2.2 Teaching Generative AI “how-knowledge” by examples and demonstration. Complementary to concepts as “what-knowledge”, teaching Generative AI how to create certain contents serves as “how-knowledge” that can bridge the gap between a generic model and users’ domain-specific needs. Some next-steps for HGAI are as follows.

- Taking the programming-by-demonstration approach, we can allow domain experts to perform a creative task for Generative AI to learn and imitate. In the Discriminative AI domain, past work has demonstrated users’ teaching an object recognizer in real-time [148]. For Generative AI, for example, an artist drawing caricature in some unique styles can demonstrate the key steps they follow; Generative AI, in turn, can learn to perform these steps, each of which would allow the artist to tweak, adjust, or innovate in their familiar ways. The grounded generation approach [83] provides a nice starting point that can allow artists to work on semantically separated elements in the generated image either by adjusting the prompt or by direct edits.
- On the other hand, some *a priori* domain knowledge should be incorporated during model training, rather than having to be demonstrated by each user when interacting with the model. For example, for scene generation, the model should ideally learn to provide camera controls (*e.g.*, aperture, focal length) as they are well-known parameters a human photographer or cinematographer would want to control.

6.3 Integrating Generative AI Into Domain Users’ Workflow

Perhaps the greatest challenge and opportunity in this HGAI level is integrating Generative AI appropriately into a domain user’s workflow based on a solid understanding of how they work, what is the best role for Generative AI, how to augment the user along each step, and what task-related contexts can further inform Generative AI.

6.3.1 Understanding humans' mental model of collaborating with Generative AI. In terms of mental model, is collaborating with Generative AI similar in some way to collaborating with other types of computing systems or collaborating with humans? As suggested by prior work on chatbot [73], a more fundamental understanding of Generative AI users' mental model can influence user experiences and inform design decisions: whether we should fit Generative AI in the old ways of work or it is worth defining a new eco-system of work unique to Generative AI. One lesson from an analogous domain is the use of freehand gestures. When camera-based tracking became widely available (e.g., Microsoft Kinect), some applications simply mapped freehand gestures to GUI buttons (old ways) rather than exploring more natural and expressive input superior to button pushing [135]. Next-steps for HGAI should carefully consider users' mental model when employing Generative AI, such as in code generation as an example—

- For programmers, simply integrating prompt-based code generation into their IDEs might be insufficient to fully realize Generative AI's potential in augmenting their programming abilities. More nuanced designs should consider a broad spectrum of issues: how long the generated code should be, when to trigger single vs. multi-line code, whether to provide single or multiple suggestions, how much latency is acceptable, which information to condition the model on (all files open in the IDE vs. the single file in focus), how to communicate to the programmer what information is being “read” by the model, how to allow for a model not to have access to sensitive files, and how to onboard users to learn all functionalities of the tool.

Based on a well-understood users' mental model, one central question to answer is where to find the best place to use Generative AI in a user's work, which we discuss next.

6.3.2 Finding the right places for Generative AI. Although Generative AI promises to provide on-demand contents throughout a user's workflow, it remains unclear where a user should employ Generative AI, how, and how much is the utility of incorporating generated contents compared to conventional approaches. For example, image generation sounds useful for visual designers but sometimes retrieving contents from stock images (e.g., a photo of a McDonald's restaurant) is already fairly convenient and will have no quality issues that some Generative AI models suffer from at times. For such tasks, Generative AI can replace but will do no better than conventional approaches. On the other hand, if the contents needed cannot be easily found (i.e., out of distribution, such as an underwater McDonald's restaurant), then Generative AI will play an indispensable role and save much efforts (e.g., searching for and Photoshopping the non-existing image). To find the right places for Generative AI, some next-steps are:

- Prior work on human-AI collaboration surfaced two “camps” of involving AI in human's work: a top-down approach where AI reports findings to a domain user (e.g., via a hierarchical organization of diagnostic evidence [60]) and a bottom-up approach where AI acts as a “copilot” to assist individual steps performed by a human (e.g., recommending where to examine on a medical image [61]). Analogously, we can instrument Generative AI either in a top-down or bottom-up manner to assist creators' work and study their reaction and preference between these two canonical approaches.
- Meanwhile, it is also worth studying unique issues of Generative AI—the gap between what a model is expected to do to help domain users and what domain users actually find useful for their work. For example, to find out whether LLMs can help screenwriters with their scripts, we need to understand how screenwriters go through different stages of writing scripts and then identify at which stages they are most likely to use LLMs and how. One recent project has studied and developed tools for 3D designers to use text-to-image generation in their

workflow [87]. Importantly, surfacing unexpected usages as well as non-usages will inform the development of next-generation Generative AI models and the eco-system of tools.

- There is the expectation that Generative AI should magically act as the genie that grants users' wishes of specific contents they are unable to create on their own. However, for creators that pursue original works, another valuable use of Generative AI is to support early-stage exploration. Thus creativity support tools powered by Generative AI should focus the interaction design on encouraging back-and-forth iterations, presenting diverse somewhat-optimal contents (rather than narrowly-defined optimal ones), and tutorials for making things based on which the creator can extrapolate and expand on their own.
- Collecting training data that not only includes the final outcome of a creator's work but also intermediate data that documents the process, *e.g.*, different versions of a drawing from rough outlines to sketches and to a version with fine details. Such a dataset would allow us to benchmark Generative AI's performance at different steps and recommend when a user should involve AI. Collectively, datasets like this across various domains provide evidence for developing a theory of what human and Generative AI are good at, respectively.

6.3.3 Step-by-step generation grounded on specific domain knowledge. Although aiming at the same final goal, human and AI might take very different approaches. In early research of medical AI, Blois found that human doctors' diagnosis often follows a funnel-like process [21]: starting with broad hypotheses, then running tests to gradually narrow down possibilities, and finally confirming the most probable cause of the observed symptoms. In contrast, most medical AI models only did best towards the end of the funnel (*i.e.*, telling whether the patient has disease X) but not so well at the triage step at the beginning.

Similarly, in most human creative process, whether it is writing a story or painting a picture, the creators would develop a domain-specific step-by-step approach, which is rarely reflected in Generative AI that achieves the same content creation. Generative AI is generally unaware of any intermediate steps and only aims for the tokens or pixels in the final result. Although it is possible to use Generative AI to simulate the step-by-step approach, *i.e.*, generating intermediate artifacts and using them as input for further generation. There is no guarantee that the result will be superior to the one-shot approach. Some next-steps for HGAI are as follows.

- Studying whether and how end-users simply utilize Generative AI to obtain the final result or there exist attempts to perform a step-by-step workflow by generating intermediate results to build on. A systematic study (*e.g.*, using technology probe) can surface a design space of using Generative AI throughout the entire creative process.
- Focused on specific domains, *e.g.*, architects designing buildings, conducting studies to understand human creators' step-by-step workflow in their existing practices, based on which we can assess whether a Generative AI module can support each step.
- Evaluating the one-shot vs. step-by-step approaches, comparing the qualities of generated contents as well as end-users' agency, workload, and satisfaction. One hypothesis is that one-shot generation is better for the initial exploration step whereas step-by-step generation is better when the creator has identified a specific direction and wants to incorporate their own creative elements into the generated contents.

6.3.4 Imbuing Generative AI with task-related contexts. As we are expected to invoke Generative AI frequently throughout our workflow, it is important that Generative AI should obtain as much contextual information as possible. To achieve this, some next-steps for HGAI include—

- Study what task-related contexts domain users make use of in their work and whether such information can be used by Generative AI. For example, for creators in theatre, given a

spatial configuration of speakers, Generative AI can suggest optimal acoustic effects for best experiences; for artists exploring multiple displays, Generative AI can propose uncommon sequences of visuals. Another example is involving intelligent tool support (not Generative AI per se) in designers' ideation process where they would draw on materials to construct a mood board [78]. A current Generative AI model might assume designers can change their process and use text prompting to help their ideation; yet the above project demonstrates how embedding support into their familiar workflow ("designer-led") based on task-related contexts is a more pragmatic approach.

- To obtain relevant task-related contexts, one promising approach is integrating Generative AI with AR systems equipped with sensors that can recognize real world objects and scenes. Imagine a user asking Generative AI to come up with a recipe based on what ingredients they have. Integrated with AR and sensors, the system can map each generated step onto specific ingredients and track the cooking progress, which is a much more immersive and natural experience than just relying on generated text recipes.
- Another open challenge and opportunity is how Generative AI can help multiple creators collaborate (e.g., human-human ideation [121]), which requires an understanding of creators' dialog with each other and what contents AI should generate that can catalyze collaboration. One analogous project focused on human-human communication where the system retrieves relevant images by inferring when a user might find it useful to have such visual information to illustrate their speech in a video conference with others [88].

6.4 Case Study: Properly Handling Generative AI's Impact On Computing Research

At the end of this section, we dedicate some discussion to how we can augment computing researchers' abilities by properly handling Generative AI's impact on how research is conducted.

6.4.1 Handling changes in conducting HCI research due to Generative AI. Beyond the obvious (and non-HCI-exclusive) use of LLM to assist writing, there are other emerging changes in how we conduct HCI research that need to be handled properly as next-steps:

- Some qualitative coding software already introduced LLM-based code generation (cf. a discussion [47] on LLM for thematic analyses [23]), which trades off original interpretation with convenience. Qualitative research is subjective, as one projects their identity into the interpretation of the data; using AI loses one's position and subjectivity. The HCI community should develop guidelines, reporting requirements, and reviewing criteria of research that analyzes qualitative data using LLMs.
- When deep learning first became a popular tool, gesture recognition algorithms found a hard time to claim a contribution given how deep learning models could often achieve a higher performance than many handcrafted algorithms. Learning from this historical lesson, as a next-step, technical HCI research [69] should define new agenda, focusing on inventing new interactive systems that catalyze Generative AI to support users to achieve more than what a vanilla Generative AI model can offer.
- On the other hand, using mature and high-performing off-the-shelf tools (be it deep learning or Generative AI) might soon be disregarded as contribution. To handle such changes, the HCI community needs to renew our definition of what constitutes an artifact contribution [136], addressing possible reviewer questions like "when is using LLM in building a system considered a contribution?"
- While a plethora of HCI research will soon flourish by building useful tools based on Generative AI, the community should ensure equal, if not more, emphasis on tools that prevent or mitigate harm done by Generative AI, from preventing programmers from over-relying on

LLM-based code generation [54] to reducing the chance of training models on artists' work (e.g., by adding adversarial noises [116]).

- The HCI community can promote tool building that supports researchers to properly and productively use Generative AI, such as following Soylent's approach [19] to let LLM automate the tedious and non-intellectual parts of paper writing. Another direction is enabling junior researchers to exchange ideas with Generative AI and obtain quick feedback to their work.

6.4.2 Cross-disciplinary collaboration (HCI + Generative X) will become a necessity. Consider the emergence of LLMs, most notably OpenAI's ChatGPT, which has democratized Generative AI to the vast public where models are no longer evaluated by benchmarks but directly judged by end-users. Therefore, there is an opportunity to introduce human subject evaluation methods to NLP research. On the other hand, HCI research is no longer constrained by a lack of NLP expertise because industry-scale models are now easily available and can be fine-tuned to fit specific use cases. Given how NLP and HCI develop the need for each other, one natural next-step is promoting the norm of collaboration across the two fields and further across HCI and other "Generative X" domains.

7 CONCLUSION

We describe the landscape of next-steps for HGAI, focusing on a technical perspective. Specifically,

- We define the term "Human-centered Generative AI" (HGAI) from a technical perspective.
- We followed a structured process to iteratively formulate next steps for HGAI.
- Our proposed next steps cross disciplinary boundaries and draw on insights from both academic and industrial research.

We hope these next-steps can serve as starting points for researchers across disciplines to collaborate and pursue specific ideas while staying informed of the big picture. As Generative AI continues to develop at unprecedented speed and scale, we believe that taking a human-centered approach early on can have a significant long-term impact on the future of human-AI symbiosis.

REFERENCES

- [1] 2019. Fairness Indicators: Scalable Infrastructure for Fair ML Systems. <https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html>
- [2] 2023. AuditNLG: Auditing Generative AI Language Modeling for Trustworthiness. <https://github.com/salesforce/AuditNLG> original-date: 2023-04-26T16:24:57Z.
- [3] 2023. Background: What is a Generative Model? | Machine Learning. <https://developers.google.com/machine-learning/gan/generative>
- [4] 2023. Blueprint for an AI Bill of Rights | OSTP. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
- [5] 2023. Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- [6] 2023. *Generative AI: Perspectives from Stanford HAI*. Technical Report. <https://hai.stanford.edu/generative-ai-perspectives-stanford-hai>
- [7] 2023. Model Card Toolkit. <https://github.com/tensorflow/model-card-toolkit> original-date: 2020-07-24T16:48:58Z.
- [8] 2023. NeMo Guardrails. <https://github.com/NVIDIA/NeMo-Guardrails> original-date: 2023-04-18T12:32:47Z.
- [9] 2023. New EPIC Report Sheds Light on Generative A.I. Harms. <https://epic.org/new-epic-report-sheds-light-on-generative-a-i-harms/>
- [10] 2023. Our approach to alignment research. <https://openai.com/blog/our-approach-to-alignment-research>
- [11] 2023. PCAST Working Group on Generative AI Invites Public Input | PCAST. <https://www.whitehouse.gov/pcast/briefing-room/2023/05/13/pcast-working-group-on-generative-ai-invites-public-input/>
- [12] 2023. Shap-E. <https://github.com/openai/shap-e> original-date: 2023-04-19T18:54:32Z.
- [13] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human-Computer Interaction* 15, 2-3 (Sept. 2000), 179–203. https://doi.org/10.1207/S15327051HCI1523_5 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI1523_5.

- [14] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. 2018. SymbiosisSketch: Combining 2D & 3D Sketching for Designing Detailed 3D Objects in Situ. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3173574.3173759>
- [15] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).
- [16] Arpit Bansal, Ping-Yeh Chiang, Michael J Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, and Tom Goldstein. 2022. Certified Neural Network Watermarks with Randomized Smoothing. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 1450–1465. <https://proceedings.mlr.press/v162/bansal22a.html>
- [17] Olivier Bau and Wendy E. Mackay. 2008. OctoPocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology (UIST '08)*. Association for Computing Machinery, New York, NY, USA, 37–46. <https://doi.org/10.1145/1449715.1449724>
- [18] Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the "Other". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300528>
- [19] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology (UIST '10)*. Association for Computing Machinery, New York, NY, USA, 313–322. <https://doi.org/10.1145/1866029.1866078>
- [20] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>
- [21] Marsden S. Blois. 1980. Clinical Judgment and Computers. *New England Journal of Medicine* 303, 4 (1980), 192–197. <https://doi.org/10.1056/NEJM198007243030405> arXiv:<https://doi.org/10.1056/NEJM198007243030405> PMID: 7383090.
- [22] Marsden S. Blois. 1984. *Information and medicine: the nature of medical descriptions*. University of California Press, Berkeley.
- [23] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, Washington, DC, US, 57–71. <https://doi.org/10.1037/13620-004>
- [24] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1664–1674. <https://doi.org/10.18653/v1/D19-1176>
- [25] 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. <https://doi.org/10.5555/3495724.3495883>
- [26] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. 2023. Perceived Trustworthiness of Natural Language Generators. <https://doi.org/10.1145/3597512.3599715> arXiv:2305.18176 [cs].
- [27] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300234>
- [28] CAiRE. 2023. ChatGPT: What It Can and Cannot Do by Prof. Pascale Fung. <https://www.youtube.com/watch?v=ORoTJZcLXek>
- [29] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. <https://doi.org/10.1145/3544548.3580959>
- [30] Davide Castelvecchi. 2020. Is Facial Recognition Too Biased to be Let Loose? *Nature* 587, 7834 (Nov. 2020), 347–349. <https://doi.org/10.1038/d41586-020-03186-4>
- [31] Pierre Chambon, Christian Bluthgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. 2022. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. <https://doi.org/10.48550/arXiv.2211.12737> arXiv:2211.12737 [cs].

- [32] Stevie Chancellor. 2023. Toward Practices for Human-Centered Machine Learning. *Commun. ACM* 66, 3 (Feb. 2023), 78–85. <https://doi.org/10.1145/3530987>
- [33] Kyle Chayka. 2023. The Uncanny Failure of A.I.-Generated Hands. *The New Yorker* (March 2023). <https://www.newyorker.com/culture/rabbit-holes/the-uncanny-failures-of-ai-generated-hands> Section: rabbit holes.
- [34] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html>
- [35] Liming Chen, Jesse Hoey, Chris D. Nugent, Diane J. Cook, and Zhiwen Yu. 2012. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (Nov. 2012), 790–808. <https://doi.org/10.1109/TSMCC.2012.2198883> Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).
- [36] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. 2287–2295. <https://doi.org/10.1145/3292500.3330723>
- [37] Xiang 'Anthony' Chen, Ye Tao, Guanyun Wang, Runchang Kang, Tovi Grossman, Stelian Coros, and Scott E. Hudson. 2018. Forte: User-Driven Generative Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. 1–12. <https://doi.org/10.1145/3173574.3174070>
- [38] Xiang 'Anthony' Chen, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. Marvista: A Human-AI Collaborative Reading Tool. *arXiv preprint arXiv:2207.08401* (2022).
- [39] Brian Christian. 2021. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.
- [40] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. <https://doi.org/10.1145/3491102.3501819>
- [41] CITRIS. 2023. Generative AI Meets Copyright - Pamela Samuelson. <https://www.youtube.com/watch?v=6sDGlrVO6mo>
- [42] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 2017. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *ACM SIGIR Forum* 51, 2 (Aug. 2017), 148–159. <https://doi.org/10.1145/3130348.3130362>
- [43] Xinyue Dai, Mark T. Keane, Laurence Shalloo, Elodie Ruelle, and Ruth M.J. Byrne. 2022. Counterfactual Explanations for Prediction and Diagnosis in XAI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 215–226. <https://doi.org/10.1145/3514094.3534144>
- [44] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3544548.3580969>
- [45] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3491102.3502141>
- [46] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. <https://doi.org/10.48550/arXiv.1912.02164> arXiv:1912.02164 [cs].
- [47] Stefano De Paoli. 2023. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. <https://doi.org/10.48550/arXiv.2305.13014> arXiv:2305.13014 [cs].
- [48] Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. 2023. A Large-scale Synthetic Pathological Dataset for Deep Learning-enabled Segmentation of Breast Cancer. *Scientific Data* 10, 1 (April 2023), 231. <https://doi.org/10.1038/s41597-023-02125-y> Number: 1 Publisher: Nature Publishing Group.
- [49] Ruofei Du, Na Li, Jing Jin, Michelle Carney, Scott Miles, Maria Kleiner, Xiuxiu Yuan, Yinda Zhang, Anuva Kulkarni, Xingyu Liu, Ahmed Sabie, Sergio Orts-Escolano, Abhishek Kar, Ping Yu, Ram Iyengar, Adarsh Kowdle, and Alex Olwal. 2023. Rapsai: Accelerating Machine Learning Prototyping of Multimedia Applications Through Visual Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 23 pages. <https://doi.org/10.1145/3544548.3581338>
- [50] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2021. DreamCoder: bootstrapping inductive program synthesis with

- wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (PLDI 2021)*. Association for Computing Machinery, New York, NY, USA, 835–850. <https://doi.org/10.1145/3453483.3454080>
- [51] Noyan Evirgen and Xiang 'Anthony' Chen. 2022. GANzilla: User-Driven Direction Discovery in Generative Adversarial Networks. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3526113.3545638>
- [52] Noyan Evirgen and Xiang 'Anthony' Chen. 2023. GANravel: User-Driven Direction Disentanglement in Generative Adversarial Networks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 19, 15 pages. <https://doi.org/10.1145/3544548.3581226>
- [53] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. <https://doi.org/10.1145/604045.604056>
- [54] James Finnie-Ansley, Paul Denny, Brett A. Becker, Andrew Luxton-Reilly, and James Prather. 2022. The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. In *Proceedings of the 24th Australasian Computing Education Conference (ACE '22)*. Association for Computing Machinery, New York, NY, USA, 10–19. <https://doi.org/10.1145/3511861.3511863>
- [55] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.
- [56] Theodoros Galanos, Antonios Liapis, and Georgios N. Yannakakis. 2023. Architext: Language-Driven Generative Architecture Design. <https://doi.org/10.48550/arXiv.2303.07519> arXiv:2303.07519 [cs].
- [57] Susan Gasson. 2003. Human-centered vs. user-centered approaches to information system design. *Journal of Information Technology Theory and Application (JITTA)* 5, 2 (2003), 5.
- [58] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. <https://doi.org/10.48550/arXiv.1803.09010> arXiv:1803.09010 [cs].
- [59] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. <https://doi.org/10.48550/arXiv.1412.6572> arXiv:1412.6572 [cs, stat].
- [60] Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Zesheng Chen, Shuo Ni, Chunxu Yang, Wenzhong Yan, Xinhai Robert Zhang, Yang Li, Mohammad Haeri, and Xiang 'Anthony' Chen. 2022. Improving Workflow Integration with XPath: Design and Evaluation of a Human-AI Diagnosis System in Pathology. *ACM Trans. Comput.-Hum. Interact.* (Dec. 2022). <https://doi.org/10.1145/3577011> Place: New York, NY, USA Publisher: Association for Computing Machinery.
- [61] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. 2023. Augmenting Pathologists with NaviPath: Design and Evaluation of a Human-AI Collaborative Navigation System. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3544548.3580694> event-place: Hamburg, Germany.
- [62] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The False Promise of Imitating Proprietary LLMs. <https://doi.org/10.48550/arXiv.2305.15717> arXiv:2305.15717 [cs].
- [63] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (June 2019), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850> Number: 2.
- [64] Bo Hedberg and Enid Mumford. 1975. The design of computer systems: Man's vision of man as an integral part of the system design process. *Human choice and computers* 31 (1975), 59.
- [65] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (July 2010), 29–29. <https://doi.org/10.1038/466029a> Number: 7302 Publisher: Nature Publishing Group.
- [66] Christian Holz and Andrew Wilson. 2011. Data miming: inferring spatial object descriptions from human gesture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 811–820. <https://doi.org/10.1145/1978942.1979060>
- [67] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [68] Yongquan Hu, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, and Aaron Quigley. 2023. Exploring the Design Space of Employing AI-Generated Content for Augmented Reality Display. <https://doi.org/10.48550/arXiv.2303.16593> arXiv:2303.16593 [cs].
- [69] Scott E. Hudson and Jennifer Mankoff. 2014. Concepts, Values, and Methods for Technical Human-Computer Interaction Research. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). Springer, New York, NY, 69–93. https://doi.org/10.1007/978-1-4939-0378-8_4

- [70] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, Ft. Lauderdale, Florida, USA, 17–24. <https://doi.org/10.1145/642611.642616>
- [71] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (March 2023), 248:1–248:38. <https://doi.org/10.1145/3571730>
- [72] Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. <https://doi.org/10.48550/arXiv.2305.02463> arXiv:2305.02463 [cs].
- [73] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 163:1–163:26. <https://doi.org/10.1145/3415234>
- [74] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html> ISSN: 2640-3498.
- [75] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. <https://doi.org/10.48550/arXiv.2304.02643> arXiv:2304.02643 [cs].
- [76] Rob Kling. 1977. The Organizational Context of User-Centered Software Designs. *MIS Quarterly* 1, 4 (1977), 41–52. <https://doi.org/10.2307/249021> Publisher: Management Information Systems Research Center, University of Minnesota.
- [77] Rob Kling and Susan Leigh Star. 1998. Human centered systems in the perspective of organizational and social informatics. *Acm Sigcas Computers and Society* 28, 1 (1998), 22–29. Publisher: ACM New York, NY, USA.
- [78] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E. Mackay. 2020. ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 45:1–45:27. <https://doi.org/10.1145/3392850>
- [79] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 5338–5348. <https://proceedings.mlr.press/v119/koh20a.html> ISSN: 2640-3498.
- [80] Gierad Laput, Xiang 'Anthony' Chen, and Chris Harrison. 2015. 3D Printed Hair: Fused Deposition Modeling of Soft Strands, Fibers, and Bristles. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 593–597. <https://doi.org/10.1145/2807442.2807484>
- [81] Clayton Lewis. 1982. *Using the "thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY.
- [82] Clayton Lewis and John Rieman. 1993. Task-centered user interface design. *A practical introduction* (1993).
- [83] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Guiding Text-to-Image Diffusion Model Towards Grounded Generation. <https://doi.org/10.48550/arXiv.2301.05221> arXiv:2301.05221 [cs].
- [84] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. 2022. VideoMap: Video Editing in Latent Space. <https://doi.org/10.48550/arXiv.2211.12492> arXiv:2211.12492 [cs].
- [85] Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, and Mark Riedl. 2023. Beyond Prompts: Exploring the Design Space of Mixed-Initiative Co-Creativity Systems. <https://doi.org/10.48550/arXiv.2305.07465> arXiv:2305.07465 [cs].
- [86] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3411764.3445488>
- [87] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2022. 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. <https://doi.org/10.48550/arXiv.2210.11603> arXiv:2210.11603 [cs].
- [88] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang 'Anthony' Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3544548.3581566> event-place: Hamburg, Germany.
- [89] Sidi Lu, Tao Meng, and Nanyun Peng. 2022. InsNet: An Efficient, Flexible, and Performant Insertion-based Text Generation Model. <https://doi.org/10.48550/arXiv.2102.11008> arXiv:2102.11008 [cs].
- [90] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. <https://doi.org/10.48550/arXiv.2104.08786>

- arXiv:2104.08786 [cs].
- [91] David McNeill. 2008. *Gesture and Thought*. In *Gesture and Thought*. University of Chicago Press. <https://doi.org/10.7208/9780226514642>
- [92] Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable Text Generation with Neurally-Decomposed Oracle. <https://doi.org/10.48550/arXiv.2205.14219> arXiv:2205.14219 [cs].
- [93] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction* 14, 4 (Nov. 2021), 272–344. <https://doi.org/10.1561/1100000083> Publisher: Now Publishers, Inc..
- [94] Rada Mihalcea and Chee Wee Leong. 2008. Toward Communicating Simple Sentences Using Pictorial Representations. *Machine Translation* 22, 3 (2008), 153–173.
- [95] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? <https://doi.org/10.48550/arXiv.2202.12837> arXiv:2202.12837 [cs].
- [96] Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. <https://doi.org/10.48550/arXiv.2304.01964> arXiv:2304.01964 [cs].
- [97] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. <https://doi.org/10.48550/arXiv.2301.11305> arXiv:2301.11305 [cs].
- [98] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. 2021. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. <https://doi.org/10.48550/arXiv.2010.10042> arXiv:2010.10042 [cs].
- [99] Florian Floyd Mueller, Pedro Lopes, Paul Strohmeier, Wendy Ju, Caitlyn Seim, Martin Weigel, Suranga Nanayakkara, Marianna Obrist, Zhuying Li, Joseph Delfa, Jun Nishida, Elizabeth M. Gerber, Dag Svanaes, Jonathan Grudin, Stefan Greuter, Kai Kunze, Thomas Erickson, Steven Greenspan, Masahiko Inami, Joe Marshall, Harald Reiterer, Katrin Wolf, Jochen Meyer, Thecla Schiphorst, Dakuo Wang, and Pattie Maes. 2020. Next Steps for Human-Computer Integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376242>
- [100] Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. 2021. House-GAN++: Generative Adversarial Layout Refinement Networks. <https://doi.org/10.48550/arXiv.2103.02574> arXiv:2103.02574 [cs].
- [101] Cal Newport. 2023. What Kind of Mind Does ChatGPT Have? *The New Yorker* (April 2023). <https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have> Section: annals of artificial intelligence.
- [102] Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan. 2021. Algorithmic bias playbook. *Center for Applied AI at Chicago Booth* (2021).
- [103] OpenAI. 2023. GPT-4 Technical Report. *ArXiv abs/2303.08774* (2023).
- [104] 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.
- [105] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. <https://doi.org/10.48550/arXiv.2305.10973> arXiv:2305.10973 [cs].
- [106] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. <https://doi.org/10.48550/arXiv.2304.03442> arXiv:2304.03442 [cs].
- [107] Ethan Perez, Sam Ringer, Kamilė Lukošiuūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. <https://doi.org/10.48550/arXiv.2212.09251> arXiv:2212.09251 [cs].

- [108] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. <https://doi.org/10.48550/arXiv.2103.00020> arXiv:2103.00020 [cs].
- [109] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv abs/2204.06125 (2022)*, 10.
- [110] Horst W Rittel and Melvin M Webber. 1974. Wicked problems. *Man-made Futures* 26, 1 (1974), 272–280.
- [111] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. (2017), 2662–2670. <https://www.ijcai.org/Proceedings/2017/371>
- [112] George IN Rozvany, O Sigmund, and Ming Zhou. 1992. Topology optimization in structural design.
- [113] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? <https://doi.org/10.48550/arXiv.2303.17548> arXiv:2303.17548 [cs].
- [114] Greg Saul, Manfred Lau, Jun Mitani, and Takeo Igarashi. 2010. SketchChair: an all-in-one chair design system for end users. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction (TEI '11)*. Association for Computing Machinery, New York, NY, USA, 73–80. <https://doi.org/10.1145/1935701.1935717>
- [115] B. Schilit, N. Adams, and R. Want. 1994. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*. 85–90. <https://doi.org/10.1109/WMCSA.1994.16>
- [116] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models. <https://doi.org/10.48550/arXiv.2302.04222> arXiv:2302.04222 [cs].
- [117] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. 440–451. <https://doi.org/10.1145/3531146.3533110>
- [118] Yujun Shen and Bolei Zhou. 2021. Closed-Form Factorization of Latent Semantics in GANs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/CVPR46437.2021.00158>
- [119] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3239–3254. <https://doi.org/10.18653/v1/2020.findings-emnlp.291>
- [120] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. “Nice Try, Kiddo”: Investigating Ad Hominems in Dialogue Responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 750–767. <https://doi.org/10.18653/v1/2021.naacl-main.60>
- [121] Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E. Mackay. 2023. Integrating AI in Human-Human Collaborative Ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–5. <https://doi.org/10.1145/3544549.3573802>
- [122] Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Sorous Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. <https://doi.org/10.48550/arXiv.1707.06742> arXiv:1707.06742 [cs, stat].
- [123] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (Dec. 2013). <http://arxiv.org/abs/1312.6034> _eprint: 1312.6034.
- [124] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- [125] Christopher Steiner. 2012. *Automate This: How Algorithms Came to Rule Our World*. Portfolio/Penguin, New York. OCLC: ocn757470260.
- [126] Jiao Sun, Yu Hou, Jiin Kim, and Nanyun Peng. 2023. Helpfulness and Fairness of Task-Oriented Dialogue Systems. <https://doi.org/10.48550/arXiv.2205.12554> arXiv:2205.12554 [cs] version: 2.
- [127] 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 (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 212–228. <https://doi.org/10.1145/3490099.3511119>
- [128] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. <https://doi.org/10.48550/arXiv.2210.04885> arXiv:2210.04885 [cs].
- [129] Steven Umbrello and Ibo van de Poel. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics* 1, 3 (Aug. 2021), 283–296. <https://doi.org/10.1007/s43681-021-00038-3>

- [130] David A van Dyk and Xiao-Li Meng. 2001. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (March 2001), 1–50. <https://doi.org/10.1198/10618600152418584> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1198/10618600152418584>.
- [131] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv preprint arXiv:2306.07899* (2023).
- [132] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. CLIPasso: semantically-aware object sketching. *ACM Transactions on Graphics* 41, 4 (July 2022), 86:1–86:11. <https://doi.org/10.1145/3528223.3530068>
- [133] Jianyuan Wang, Ceyuan Yang, Yinghao Xu, Yujun Shen, Hongdong Li, and Bolei Zhou. 2022. Improving GAN Equilibrium by Raising Spatial Awareness.
- [134] Yau-Shian Wang and Yingshan Chang. 2022. Toxicity Detection with Generative Prompt-based Inference. <https://doi.org/10.48550/arXiv.2205.12390> arXiv:2205.12390 [cs].
- [135] Daniel Wigdor and Dennis Wixon. 2011. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier.
- [136] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44. Publisher: ACM New York, NY, USA.
- [137] Catherine Wong, William P. McCarthy, Gabriel Grand, Yoni Friedman, Joshua B. Tenenbaum, Jacob Andreas, Robert D. Hawkins, and Judith E. Fan. 2022. Identifying concept libraries from language about object structure. <https://doi.org/10.48550/arXiv.2205.05666> arXiv:2205.05666 [cs].
- [138] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2016/hash/44f683a84163b3523afe57c2e008bc8c-Abstract.html>
- [139] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–10.
- [140] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712* (2023).
- [141] Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023. Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning. <https://doi.org/10.48550/arXiv.2306.01150> arXiv:2306.01150 [cs].
- [142] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting With Contextual Attention. 5505–5514. https://openaccess.thecvf.com/content_cvpr_2018/html/Yu_Generative_Image_Inpainting_CVPR_2018_paper.html
- [143] J.D. 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3544548.3581388>
- [144] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. <https://doi.org/10.48550/arXiv.2204.00598> arXiv:2204.00598 [cs].
- [145] Honghua Zhang, Meihua Dang, Nanyun Peng, and Guy Van den Broeck. 2023. Tractable Control for Autoregressive Language Generation. <https://doi.org/10.48550/arXiv.2304.07438> arXiv:2304.07438 [cs].
- [146] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]
- [147] Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training Language Models with Memory Augmentation. <https://doi.org/10.48550/arXiv.2205.12674> arXiv:2205.12674 [cs].
- [148] Zhongyi Zhou and Koji Yatani. 2022. Gesture-aware Interactive Machine Teaching with In-situ Object Annotations. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3526113.3545648>
- [149] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. 2007. A Text-to-Picture Synthesis System for Augmenting Communication. In *AAAI*, Vol. 7. Association for Computing Machinery, New York, NY, USA, 1590–1595. <https://doi.org/10.5555/1619797.1619900>
- [150] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-based Colorization of Scene Sketches. *ACM Transactions on Graphics* 38, 6 (Dec. 2019), 233:1–233:16. <https://doi.org/10.1145/3355089.3356561>