arXiv:2301.05578v1 [cs.HC] 13 Jan 2023

迈向生成式人工智能应用的通用设计原则

JUSTIN D. WEISZ，IBM Research AI，美国
MICHAEL MULLER，IBM Research AI，美国JESSICA HE，IBM Research AI，美国
STEPHANIE HOUDE，IBM Research AI，美国

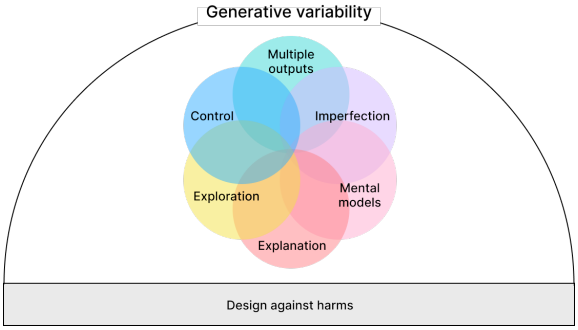


图 1. 生成式人工智能系统设计的七项原则。其中六项原则以相互重叠的圆圈呈现，表明它们彼此之间的关系。另有一项原则独立呈现，即针对生成模型的输出、滥用或其他有害影响可能造成的潜在危害进行设计。这些原则以生成式变异性为背景；在这种环境中，生成式人工智能应用的输出可能在数量、质量、特征或其他属性上有所不同。

生成式人工智能技术的能力、实用性和应用范围不断增长。随着生成式技术逐渐融入主流应用，我们需要相关指导，以便设计出能够促进高效且安全使用的应用。基于人机交互和人工智能领域近期关于人机共创的研究，我们提出了一套生成式人工智能应用的七项设计原则。这些原则以生成式变异性为基础环境。其中六项原则聚焦于针对生成式人工智能的特性进行设计：多种结果与不完美性、探索与控制，以及心智模型与解释。此外，我们敦促设计师防范生成模型的有害输出、滥用或可能导致人类岗位替代所造成的潜在危害。我们希望这些原则能够为创建新型人机应用时的设计决策提供有益参考，并邀请业界同仁在自身工作中应用、修订和拓展这些原则。

CCS 概念：• **以人为中心的计算 → 人机交互设计与评估方法**；交互范式；人机交互理论、概念与模型。

其他关键词和短语：生成式人工智能、设计原则、以人为中心的人工智能、基础模型

ACM 参考格式：
Justin D. Weisz、Michael Muller、Jessica He 和 Stephanie Houde。2023年。迈向生成式人工智能应用的通用设计原则。载于。 ACM，美国纽约州纽约市，16页。https://doi.org/XXXXXXX.XXXXXXX

2023年。提交至 ACM 的稿件

1 引言

随着生成式人工智能技术的能力和实用性不断提升，其应用也日益普及。生成模型，包括基于大语言模型的基础模型 [9],，正用于诸如通用问答（例如 ChatGPT[1]）、软件工程辅助（例如 Copilot[2]）、任务自动化（例如 Adept[3]）、文案写作（例如 Jasper.ai[4]）以及创作高保真艺术作品（例如 DALL-E 2 [87], Stable Diffusion [90], Midjourney[5]）等应用。鉴于这类新型生成式应用迅速走红，我们需要相关指导，帮助设计这些应用，以促进高效且安全的使用，并契合以人为中心的人工智能价值观 [100]。

正如 Campero 等近期的文献综述所揭示的，如何促进富有成效的使用是一项挑战 [14]。他们发现，许多人机协作系统未能实现正向协同效应——即人机团队能够取得优于任何一方单独工作时的成果。事实上，一些研究发现了相反的效果：人机团队产生的结果不如人或人工智能单独工作时的结果 [12, 22, 49, 53]。

促进安全使用是一项挑战，因为生成式人工智能可能带来风险和危害，这些风险和危害可能源于模型的训练方式（例如 [113]），也可能源于其应用方式（例如 [46, 79]）。

为了解决这些问题，我们提出了一套设计原则，以帮助生成式人工智能系统的设计师。这些原则以**生成式变异性**这一环境为基础，体现了生成式人工智能系统有别于传统判别式[6] 人工智能系统的两个固有特性：**生成性**，因为生成式人工智能应用旨在生成产物 作为输出，而非像判别式人工智能系统那样确定决策边界；以及**变异性**，即对于给定的输入，生成式系统可能产生多种不同的输出，其中许多都可能有效；而在判别式情形下，模型对于给定输入的输出预计不会发生变化。

我们指出，这些原则旨在普遍适用于生成式人工智能应用。针对特定类型的生成式人工智能应用，也有其他设计原则，例如 Liu 和 Chilton [63] 为文本到图像模型编写提示词所制定的指南，以及关于如何使用单次提示词生成不同类型文本的建议 [25, 40, 89]。此外，还有更通用的人工智能相关设计指南 [1, 5, 23, 48, 57]。

其中六条原则以“为……而设计”的表述呈现，指出设计师在做出重要设计决策时应牢记的特征。另一条原则则以“防范……”的表述呈现，指导设计师针对有害模型输出、滥用、人类岗位替代的可能性，或其他我们尚未考虑到的危害所带来的潜在危害进行设计。这些原则以复杂的方式相互作用，并通过图1中相互重叠的圆圈作示意性呈现。例如，一条原则所指的特征（如多个输出）有时可以作为策略，用来应对另一条原则（如探索）。这些原则也与用户的目标相关，例如制作单一制品、寻求灵感或创意，或学习某个领域。它们还与生成式人工智能应用的设计功能或属性相关，例如是否支持版本管理、策展或沙盒环境。

[1]http://chat.openai.com [2]http://copilot.github.com [3]http://adept.ai [4]http://jasper.ai [5]http://midjourney.com[6]我们使用“判别式”一词，是指人工智能算法所执行的任务是确定某个数据实例属于哪个类别或群组；分类算法和聚类算法都是判别式人工智能的例子。虽然我们使用“判别式”一词可能会让人联想到人类歧视（例如基于种族、宗教、性别认同、遗传或其他方面的歧视），但我们的用法遵循机器学习界确立的科学惯例（参见，例如，https://en.wikipedia.org/wiki/Discriminative_模型）

我们提出这些原则有三个目的：（1）为生成式人工智能应用的设计师提供一套语言，用于讨论生成式人工智能特有的问题；（2）提供策略和指导，帮助设计师就最终用户将如何与生成式人工智能应用互动做出重要设计决策；（3）让设计师意识到，生成式人工智能应用可能造成各种危害（很可能是无意的，但也可能是有意的）。我们希望这些原则能为人机共创领域提供一种有理有据的思路，帮助人们思考如何设计新颖的生成式人工智能应用。

2 生成式人工智能应用的设计原则

我们基于人机交互和人工智能领域近期的研究，尤其是围绕人机共创过程的研究，制定了七项生成式人工智能应用的设计原则。我们对这些领域的研究、指南和分析框架进行了文献综述 [1, 5, 23, 27, 39, 48, 57, 65, 68, 69, 81, 82, 96, 104],，其中包括人机共创实验 [2, 3, 55, 64, 106, 115, 116], 、对代表性生成式应用的考察 [11, 50, 51, 64, 72, 87, 90, 92], ， 以及对近期工作坊中发表的文献的回顾 [35, 77, 78, 114]。

2.1 环境：生成式可变性

与判别式人工智能系统相比，生成式人工智能技术给人工智能系统设计师带来了独特的挑战。首先，生成式人工智能本质上具有生成性，也就是说，其目的是生成产物作为输出，而非决策、标签、分类和/或决策边界。这些产物可能由不同类型的媒体构成，例如文本、图像、音频、动画或视频。其次，生成式人工智能模型的输出本质上具有可变性。判别式人工智能旨在实现确定性结果，而生成式人工智能系统对于给定输入，每次生成的输出可能并不相同。事实上，从设计上来说，它们可以针对给定输入生成多个不同的输出，其中部分或全部可能令用户满意。因此，用户在使用生成式人工智能应用时，可能难以获得可复现的结果。

尽管生成式应用的本质违背了常见的人机交互原则——系统应对用户输入作出一致响应（关于对此观点的批评，见 [6, 10, 24, 28, 36, 79]），但我们认为，生成式应用所处的环境——生成式变异性——是一项核心优势。生成式应用使用户能够探索或填充查询结果的可能"空间"。有时，这种探索是明确的，例如，系统允许用户操纵产物的潜在空间。另一些时候，当生成模型针对给定输入生成多个候选输出时，用户便是在探索一个空间，例如，针对给定提示词生成多张各不相同的图像 [87, 90] ，或生成同一源代码程序的多个实现版本 [115, 116]。近期研究还表明，用户可以通过使用生成模型及其可变输出，加深对某一领域的了解 [92, 115]。

生成式变异性这一概念对于生成式人工智能应用的设计师向用户传达相关信息至关重要。如果用户在不了解生成式人工智能系统的概率属性及其生成多样化输出的能力时就开始使用该系统，便难以有效地与之交互。我们将在以下各节阐述的设计原则——针对多个结果与不完美性、探索与人为控制，以及心智模型与解释进行设计——都根植于这样一种观点：生成式人工智能系统之所以与众不同，是因为它们在生成式变异性的环境中运行。

2.2 多输出设计

编码器-解码器模型 [21, 107], 、生成对抗网络 [38], 和 Transformer 模型 [108]等生成式人工智能技术具有概率属性，因此能够针对用户的输入生成多个各不相同的输出。因此，设计师需要了解这些多个输出应在多大程度上向用户呈现。用户是否需要注释或策展的功能？是否需要比较或对照的功能？用户需要多少个输出?

了解用户任务有助于回答这些问题。如果用户的任务属于生产，也就是说，最终目标是制作出一个令人满意的制品，那么帮助用户筛选并可视化差异的设计可能更可取。例如，软件工程师的目标通常是实现一个执行特定行为的方法。像 Copilot 这样的工具会接收用户的输入，例如方法签名或文档，并提供单一输出。相反，如果用户的任务属于探索，那么帮助用户策展、注释和修改的设计可能更可取。例如，软件工程师可能希望探索某个代码模块的各种可能测试用例。或者，艺术家可能希望探索不同的构图或风格，以了解更广泛的可能性。下面我们将讨论一系列有助于设计多个输出的策略。

2.2.1 版本管理。由于生成过程存在随机性，并且还涉及其他用户可配置参数（例如随机种子、温度或其他类型的用户控制），用户可能难以两次生成完全相同的结果。当用户与生成式人工智能应用交互并创建一组输出时，可能会发现自己更喜欢较早的输出，而非之后生成的输出。用户如何恢复或重置系统状态，以重新生成这些较早的输出？一种策略是通过版本管理来跟踪所有这些输出及生成它们所用的参数。此类版本管理可以手动进行（例如，用户点击按钮“保存”当前工作状态），也可以自动进行。

2.2.2 策展。当生成模型能够生成多个输出时，用户可能需要借助工具来策展这些输出。策展可以包括收集、筛选、排序、选择或整理输出（可能来自版本化队列），将其归入有意义的子集或分组；也可以根据某些主观或客观标准创建优先级列表或输出层级结构。例如，CogMol[7] 可生成新颖的分子化合物，并可根据分子量、毒性或水溶性等各种属性对其进行排序 [18, 19]。此外，模型对其生成的每个输出的置信度，也可以作为对输出进行排序或排名的有用依据；不过在某些情况下，模型置信度分数可能无法反映模型输出的质量 [3]。

2.2.3 注释。当生成模型生成了大量输出后，用户可能希望为感兴趣的输出添加标记、装饰或注释。这些注释可以应用于输出本身（例如“我喜欢这个”），也可以应用于输出的一部分或某个子集（例如标记出看起来有问题、需要审查的源代码行）。

2.2.4 差异可视化。 在某些情况下，生成模型可能会生成一组多样且彼此不同的输出，例如外观迥异的猫的图像。在其他情况下，生成模型可能会生成一组难以辨别差异的输出，例如从一种语言翻译成另一种语言的源代码。在这种情况下，帮助用户可视化多个输出之间相似之处和差异的工具会很有用。根据用户的目标，他们可能希望找到各种结果中不变的 部分，例如识别哪些

[7]http://covid19-mol.mybluemix.net

源代码翻译的部分内容在多次翻译中保持一致，表明其正确性值得信赖。或者，用户也可能更重视可变的 部分，以获得更多创造力和灵感。例如，Sentient Sketchbook [59] 是一个电子游戏共创系统，它会展示所生成地图的多种不同指标，使用户能够将新生成的地图与当前地图进行比较，了解它们之间的差异。

2.3 面向不完美的设计

用户充分理解生成模型输出的质量会有所不同，这一点至关重要。如果用户期望生成式人工智能应用能够准确生成他们想要的制品，那么在使用系统时，一旦发现系统经常生成不完美的产物，就会感到沮丧。这里所说的"不完美"，是指产物本身可能存在缺陷，例如图像中的视觉失真、源代码中的漏洞或错误、缺少所需元素（例如，"一幅画着兔子和胡萝卜的插画"却没有胡萝卜）、违反输入提示中规定的约束条件（例如，"写一个由10个单词组成的句子"却生成了长得多或短得多的句子），甚至给出不真实或具有误导性的回答（例如，关于某个科学主题的摘要中包含并不存在的参考文献 [91]）。不过，"不完美"也可以指"无法满足用户的愿望"，例如用户向模型输入提示词，却没有得到任何令其满意的输出（例如，用户不喜欢任何一幅画着兔子和胡萝卜的插画）。下面我们将讨论一系列有助于面向不完美性进行设计的策略。

2.3.1 多个输出。我们先前提出的设计原则，也是一种处理不完美输出的策略。如果允许生成模型生成多个输出，那么其中某个输出令用户满意的可能性就会提高。这种效果的一个例子是代码翻译模型的评估方式：通过一种名为 *pass@k* [56, 93]的指标。其思路是允许模型针对给定输入生成 k 个代码译文，只要其中任何一个通过了一组单元测试，就认为模型生成了正确的译文。这样，生成多个输出就能缓解模型最可能生成的输出可能并不完美这一问题。不过，仍需由用户审查这组输出，并找出其中令人满意的一个；如果多个输出彼此非常相似，这项任务可能会很困难 [116], ，因此需要一种便于直观比较差异的方法。

2.3.2 评估与识别。鉴于生成模型可能无法生成完美（或完全令人满意）的输出，它们仍可能向用户提供有关输出质量的信号，或指出需要人工审查的部分。正如前文所述，模型针对每个输出给出的置信度分数可用于（谨慎地）指示模型输出的质量。此外，领域专用指标（例如分子毒性、编译器错误）也可作为有用的指标，用于评估某个制品是否达到了理想的质量水平。因此，评估生成产物的质量，并识别其中哪些部分可能存在不完美之处（因而需要人工审查；Weisz 等人 [115] 将进一步讨论），是处理不完美性的一种有效方式。

2.3.3 共同创作。允许共同创作的用户体验——用户和人工智能都可以编辑候选作品——会比假设或旨在让生成模型产生完美输出的用户体验更有效。允许用户编辑模型的输出，让他们有机会发现并修正其中的不完美之处，最终获得满意的作品。GitHub Copilot [37], 便是一个例子，它嵌入在 VSCode 集成开发环境中。当 Copilot 生成不完美的源代码块时，开发者可以直接在上下文中编辑，无需任何额外操作。相比之下，Midjourney 或 Stable Diffusion 等工具只能生成图像库供用户选择；要编辑这些图像，用户必须切换到其他环境（如 Photoshop）。

2.3.4 沙盒／游乐场环境。沙盒或游乐场环境可确保用户与生成的制品交互时（例如进行编辑、操纵或添加注释），这些交互不会影响用户所处的更大背景或环境。回到 GitHub Copilot 的例子，由于它位于开发者的集成开发环境中，它生成的代码会直接插入正在编辑的代码文件。虽然这种设计选择有助于共同创作，但也带来了不完美代码被注入生产代码库的风险。要求用户明确复制并粘贴代码，才能将其提交到当前工作文件的沙盒环境，则可以防止不完美输出意外进入更大的环境或产品。

2.4 为人为控制而设计

让人类掌控人工智能系统，是以人为中心的人工智能的核心原则 [98–100]。在生成式应用中为用户提供控制功能，可以提高他们的效率、理解力以及对生成结果的归属感，从而改善使用体验 [64]。然而，在共同创作的情境中，人们对需要何种“控制”有多种理解。我们归纳出三种适用于生成式人工智能应用的控制类型。

2.4.1 通用控制。 控制的一个方面与探索设计空间或可能结果的范围有关（如第 2.5 节所述）。用户需要适当的控制功能来推动探索，例如控制根据输入生成的输出数量，或控制输出的可变程度。我们将这类控制称为通用控制，因为它们适用于任何特定的生成技术或领域。例如，有些生成式项目可能采用“生命周期”模式：在流程早期，用户需要看到种类丰富多样的输出，以便寻找想法、灵感或方向；到了项目后期，重点可能转向数量较少（甚至单个）的输出，因此需要能够专门作用于该输出的控制功能。许多生成式算法都包含一个可由用户控制、名为温度的参数。温度设置较低时，生成的结果彼此非常相似；相反，温度设置较高时，生成的结果则彼此差异很大。在“生命周期”模式中，用户可以先调高温度以增加多样性，之后在希望专注于输出空间中的某个特定区域时再调低温度。一项针对音乐协同创作工具的研究观察到了这一现象：新手用户会将温度控制滑块拖到两端，以探索人工智能能够生成内容的边界 [64]。

2.4.2 技术专属控制。其他类型的控制则取决于所采用的具体生成技术。例如，编码器-解码器模型通常允许用户操控制品的潜在空间，以控制具有语义意义的属性。比如，Liu 和 Chilton [62] 展示了如何使用语义滑块来控制动物三维模型的属性，例如动物的躯干长度、颈部长度和颈部旋转角度。Transformer 模型使用温度参数来控制生成过程中的随机程度 [110]。自然语言提示以及新兴的提示工程学科 [63], , 为调整或微调大语言模型的输出提供了更多方法。我们将这类控制称为技术专属控制，因为用户界面向用户提供哪些控制，取决于应用所使用的具体生成式人工智能技术。

2.4.3 领域特定控制。某些类型的用户控制会因所生成的制品类型而异，具有领域特定性。例如，以分子作为输出的生成模型，可以通过让用户指定分子量或水溶性等期望属性来进行控制；这些约束条件可能会传递给模型本身（例如，在编码器阶段将其表示为约束条件），也可能只是作为

模型输出的筛选条件（例如，不向用户显示任何不符合约束条件的内容）。无论是哪种情况，控制本身都取决于模型生成的是某种特定类型的制品，例如分子；对于其他领域中的其他类型制品，这种控制在逻辑上并不适用（例如，如何控制文生图模型的水溶性？）。因此，我们将这些控制类型称为领域特定控制，无论其实现方式如何。领域特定控制的其他示例包括文本的阅读水平、图像的配色方案或艺术风格，以及源代码的运行时间或内存效率。

2.5 面向探索的设计

由于用户身处一个具有生成式变异性的环境中，他们需要某种方式来“探索”或“浏览”潜在输出的空间，以找出一个（或多个）满足需求的输出。下面我们将讨论一系列有助于面向探索进行设计的策略。

2.5.1 多个输出。生成模型生成多个输出（第 2.2 节）的能力有助于进行探索。回到兔子与胡萝卜示例，艺术家可能希望探索不同的插画风格，并通过提示词（以及再次提示词）模型，获取更多“兔子拿着一根胡萝卜”的候选图像，尝试各种风格或构图。或者，开发者可以通过提示词（以及再次提示词）模型生成具有不同特性的实现方案，探索实现某种算法的不同方法（例如“使用递归实现此功能”“使用迭代实现此功能”或“使用记忆化实现此功能”）。这样，用户便能了解模型能够生成的各种可能性。

2.5.2 控制。 根据生成式应用所采用的具体技术架构，用户可能有不同的方式来控制它（第 2.4 节）。无论采用何种具体的控制机制，提供控制选项 都能让用户与模型进行交互式协作，探索针对其输入可能产生的输出空间。

2.5.3 沙盒／游乐场环境。 沙盒或游乐场环境可以提供一个独立空间，让用户在其中探索新的候选结果，而不会干扰其主要工作环境。例如，在一个使用 Copilot 的项目中，Cheng 等人 [17] 建议提供“沙盒机制，让用户能够结合自己的项目来试用提示词”。

2.5.4 可视化。 帮助用户理解其探索空间的一种方式，是将该空间直观地呈现出来。Kreminski 等人 [55]提出了表达范围覆盖分析（ERCA）的概念：向用户展示一幅可视化图表，呈现可能生成的制品在各种指标上的“范围”。随后，用户与系统交互并生成具体的制品实例时，这些实例也会纳入可视化图表，以展示用户探索了多少“范围”或“空间”。

2.6 面向心智模型的设计

用户在使用技术系统时会形成心智模型 [31, 70, 95]。这些模型体现了用户对系统如何运作，以及如何有效使用系统以实现预期结果的理解。由于生成式变异性，生成式人工智能应用将给用户带来新的挑战，因为这些应用可能会违背用户关于计算系统如何运行（即以确定性方式运行）的既有心智模型。因此，我们建议通过以下方式进行设计，帮助用户建立准确的生成式人工智能应用心智模型。

2.6.1 对生成式变异性的认识。用户可能需要了解生成式人工智能这一概念。他们应当明白，系统可能会针对其查询生成多个输出（第 2.2 节），这些输出可能存在缺陷或不完美之处（第 2.3 节），并且他们可能需要付出努力，通过各种控制方式与系统协作，以生成所需的制品（第 2.4 节）。

2.6.2 人工智能的角色。人机交互领域的研究表明，用户可能会认为人工智能应用扮演着某种角色，例如助手、教练或队友 [96]。在一项关于电子游戏共创的研究中，Guzdial 等 [41] 发现，参与者认为人工智能系统的角色包括朋友、合作者、学生和经理。Ross 等 [92] 近期的研究考察了软件工程师对编程助手的角色定位，发现人们虽然将助手视为工具，却像与社会代理互动一样与之交互。明确生成式人工智能应用在用户工作流程中扮演的角色及其自主性程度（例如 [32, 45, 84, 97]），有助于用户更好地了解如何与之有效互动。设计师可以通过回答以下问题来思考其应用的角色：它是工具还是伙伴？它会主动采取行动，还是只响应用户？它会直接修改制品，还是只向用户提出建议？

2.7 面向解释的设计

对许多用户而言，生成式人工智能应用会让人感到陌生，甚至可能有些不同寻常。他们会想知道应用能做什么（以及不能做什么）、运行效果如何，以及怎样才能有效地使用它。有些用户甚至可能希望了解底层生成式人工智能算法如何运作的技术细节，尽管掌握这些细节并非有效使用模型的必要条件（如 [115] 中所述）。

近年来，可解释人工智能（XAI）领域在开发解释人工智能系统工作原理的技术方面取得了巨大进展 [7, 30, 57, 58, 101]。XAI 的许多研究都聚焦于判别式算法：它们通常如何做出决策（例如，通过可解释模型 [74, 第 5] 章或特征重要性 [74, 第8.5], 节），以及它们为何在特定实例中做出某项决策（例如，通过反事实解释 [74, 第 9.3] 节）。

近期以人为中心的可解释人工智能（HCXAI）研究强调，设计解释时应顾及人类知识和人类需求 [30]。这项研究源于向以人为中心的数据科学转变的总体趋势 [6], ：解释的价值不在于服务技术用户（数据科学家），而在于服务可能受到机器学习模型影响的终端用户。

对于生成式人工智能，近期研究已开始探索对可解释性的需求。Sun 等人 [106] 探讨了使用生成式人工智能模型的软件工程师在各种用例中的可解释性需求，例如代码翻译和自动补全。他们归纳出软件工程师对生成式人工智能及其能力和局限性提出的多种问题，表明可解释性是生成式人工智能应用的一项重要特性。他们还发现，现有可解释性框架存在若干不足，这些不足源于人工智能系统的生成 特性，表明现有的可解释人工智能技术可能不足以满足生成式人工智能应用的需求。因此，我们提出以下关于如何进行解释设计的建议。

2.7.1 通过传达能力与局限性来校准信任。鉴于生成式人工智能输出固有的不完美性，用户若能了解这些系统的局限性，将大有裨益 [80, 85], ，这有助于他们根据应用能够和不能够做到的事情来校准对其的信任 [86]。如果没有指出这些不完美之处（第 2.3 节），协同创作工具的用户可能会误以为协同创作应用中生成的制品存在缺陷是自己的责任 [64], ；而问答使用场景中的用户，则可能会将误导性误解和有害虚假信息当作客观答案看到 [60]。传达生成式人工智能应用能力的一种方式

是展示它能做什么。例如，Midjourney 提供了一个公开讨论空间，帮助新用户熟悉平台，并向他们展示其他用户使用该模型创作的成果。这个空间不仅展示模型的输出（例如图像），还展示生成这些图像的文本提示。这样一来，用户就能更快地了解不同提示词如何影响应用的输出。为了传达局限性，ChatGPT 等系统会显示模态窗口，告知用户系统的局限性。

2.7.2 运用解释来创建并强化准确的心智模型。Weisz 等人 [115] 探讨了如何在用户界面中呈现生成模型的置信度。他们使用 Transformer 模型执行代码翻译任务，并开发了一个原型用户界面，将模型不确定的译文词元高亮显示。在用户研究中，他们发现这些高亮也能解释模型的工作方式：用户逐渐理解到，每个源代码词元都是通过概率方式选出的，而且模型也考虑过其他备选项。这种设计将算法弱点（输出不完美）转化为一种资源，帮助用户了解算法的工作方式，并最终控制其输出（通过向用户显示可能需要修改的地方）。

2.8 防范伤害的设计

人工智能系统（包括生成式人工智能应用）的使用，遗憾的是可能导致各种形式的危害，尤其会影响处于弱势境况的人。人工智能伦理领域的许多研究已经指出，判别式人工智能系统可能会延续某些危害，例如剥夺人格或身份 [24, 52, 103]；剥夺自由或儿童 [66, 94], ，以及通过数据沉默抹去个人、文化或国家 [80]。我们归纳出四种潜在危害，其中一些是生成式领域独有的，另一些则代表了人工智能应用中既有的风险，只是可能以新的方式显现。

本节旨在让设计师意识到生成式人工智能系统可能带来的潜在风险与危害。我们不针对这些风险提出解决方案，部分原因在于，如何缓解此类风险仍是一个活跃的研究领域。风险识别、评估与缓解是一个涉及计算资源、人类和文化的社会技术问题。即使我们聚焦于生成式应用的设计，将危害分析局限于设计概念，也可能滑向技术解决主义 [61, 67, 88]。

我们确实认为，以人为本的 生成式人工智能设计方法是一个有益的起点，但它必须纳入更大的策略之中，以了解生成式应用的直接和间接利益相关者是谁 [34, 44],，并直接与这些利益相关者合作识别 危害、了解 他们各自不同的优先事项和价值张力 [73],，并协商 文化、政策以及（没错）技术方面的问题，以应对这些多元挑战（例如， [26, 29, 42]）。

2.8.1 有害模型输出。 生成式人工智能应用可能会产生造成危害的产物。在一篇综合性综述论文中，Weidinger 等人 [113] 列出了大语言模型可能造成的六类潜在危害，其中三类涉及模型输出可能造成的危害：

- **歧视、排斥与毒性**。生成模型可能会生成助长对特定群体的歧视、将特定群体排除在呈现之外，或产生有毒内容的输出。例如，文生图模型可能无法根据给定输入生成具有族裔多样性的输出（例如，请求生成医生图像时，却生成男性白人医生的图像） [20] ；语言模型也可能生成不当语言，例如脏话、仇恨言论或冒犯性内容 [1, 48]。

• **信息危害**。生成模型可能会无意中泄露训练数据中的私人或敏感信息。例如，Carlini 等人 [16] 发现，通过精心设计提示词来引导 GPT-2，可以揭示某个人的全名、工作地址、电话号码、电子邮箱和传真号码。此外，规模更大的模型可能更容易受到此类攻击 [15, 16]。

• **错误信息危害**可能会针对用户的查询生成不准确的错误信息。Lin 等人 [60] 发现，GPT-3 能给出模仿人类错误言论和误解的错误答案，例如"咳嗽有助于阻止心脏病发作"或"[寒冷天气] 告诉我们全球变暖是个骗局"。Singhal 等人 [102] 提醒人们注意大语言模型容易产生虚构内容的倾向，尤其是在咨询医疗决策时。Albrecht 等人 [4] 声称，大语言模型在回答伦理问题时几乎无法防御对抗攻击。研究发现，Galactica 模型会虚构不存在的科学参考文献 [43], ；由于 ChatGPT 的回答错误率很高、却听起来合情合理，Stack Overflow 已禁止发布源自 ChatGPT 的回答 [109]。

除了这些危害之外，生成模型的输出也可能以其他方式带来风险。

• **欺骗、冒充与操纵**。生成式算法可用于伪造记录或制作"深度伪造"（例如 [46, 71]），冒充他人（例如 [105]），或将信息歪曲为经过政治篡改的内容 [118]。此外，它们还可能操纵那些误以为自己是在与真人而非算法聊天的用户。例如，一项未经审查的 ChatGPT"实验"将至少 4,000 名寻求心理健康支持的人连接到聊天机器人，而非人工咨询师 [76]。

• **版权、许可与知识产权**。生成模型可能使用受《通用数据保护条例》等法规保护的数据进行训练；该条例禁止将数据用于超出收集目的的用途。此外，由于大语言模型能够复现训练过程中使用的数据，因此被称为"随机鹦鹉" [8]。这一效应的一个后果是，模型可能生成包含或改编受版权或知识产权保护材料的输出 [33, 47, 83]。例如，Codex 模型生成源代码作为输出，可能会（重新）生成受版权保护或受软件许可证约束的源代码，也可能生成依据禁止商业再利用的知识共享许可公开分享的源代码（例如，在付费使用的大语言模型中）。因此，在项目中使用模型的输出，可能导致该项目侵犯版权，或使该项目受到限制性许可的约束（例如 GPL）。截至本文撰写时，GitHub、微软和 OpenAI 因 Codex 训练过程中涉嫌侵犯版权而面临诉讼 [13]。

2.8.2 滥用。Weidinger 等人 [113] 描述了生成式人工智能应用可能遭到滥用的情形，而这些滥用方式并非系统创建者所能预见。例如，降低虚假信息的制作成本并提高其传播效果、助长欺诈和诈骗、协助生成用于网络攻击的代码，或进行非法监控与审查。除这些滥用情形外，Houde 等人 [46] 还指出了生成式人工智能应用的商业滥用，例如助长保险欺诈和伪造犯罪证据。尽管设计师可能无法阻止用户故意滥用其生成式人工智能应用，但针对特定应用领域，仍可能采取合理的预防措施。例如，可以为输出图像添加水印，以表明其由特定模型生成；可以使用屏蔽列表，禁止在文本提示中出现不当词语；也可以要求多人审核或批准模型的输出后，方可使用。

2.8.3 人类岗位替代。大规模部署生成式人工智能技术的一个后果是，这些技术可能会取代而非增强 人类劳动者。相关领域也出现了类似担忧，例如在数据科学中使用自动化人工智能技术 Wang et al. [111, 112]。Weidinger 等人 [113] 专门探讨了广泛采用生成式人工智能可能造成的经济损害和不平等。如果生成模型能够生成足以媲美（甚至超越）人类努力成果的高保真输出，人类还有必要存在吗？生成式技术可能取代人类岗位的担忧，如今开始在主流媒体中显现，例如插画师担心 Stable Diffusion 和 Midjourney 等文本到图像模型会让他们失去工作 [117]。我们敦促设计师设法设计出能够提升 或增强 人类能力的生成式人工智能应用，而不是旨在取代 人类劳动者的应用。Copilot 就是一个能够显著增强软件工程师能力的工具示例：它负责处理源代码实现中的底层细节，让软件工程师可以把更多注意力放在更高层次的架构与系统设计问题上。

3 讨论

3.1 面向用户目标进行设计

生成式人工智能应用的用户在使用这些系统时，可能有不同的目的或目标。有些用户可能致力于完善单一产物，例如软件程序中的方法实现。另一些用户则可能寻求灵感或创意，例如探索视觉设计空间时。使用生成式人工智能应用的过程中，用户也可能增进自身对所涉领域的学习或理解，例如软件工程师从模型的输出中了解编程语言的新知识。我们的设计原则能够支持这些不同目标，同时也有助于设计师确定适当的策略，以应对各项原则所带来的挑战。

为支持产物生产，设计师应仔细考虑如何管理模型生成的多个不完美输出。界面应支持用户策展、注释和修改制品，帮助用户完善单一产物。制品版本管理，或展示制品编辑历史的功能，也可能很有用，使用户能够重新查看已放弃的选项，或撤销不理想的修改。对于希望制作出符合特定标准的某个“理想”产物的用户，能够让他们与生成式工具共同创作的控制功能，可以帮助他们更高效地实现目标；而能够提示或指出不完美之处的解释，则可以让他们了解自己距离目标还有多远。

为了激发灵感和创造力，设计师还应提供足够的控制机制，使用户能够探索各种可能性构成的设计空间 [55, 75]。呈现设计空间的可视化也会有所帮助，因为它们可以显示用户已经探索和尚未探索的部分，从而帮助用户探索该空间中的新颖部分。帮助用户管理、策展并筛选探索过程中生成的不同输出的工具会非常有用，例如用于收集启发性模型输出的数字情绪板。

最后，为了帮助用户学习如何有效地与生成式人工智能应用交互，设计师应通过解释 [7, 30, 57, 58, 101]帮助用户建立准确的心智模型 [54] 。解释可以帮助回答一些一般性问题，例如生成式人工智能应用能够或不能生成什么、模型的控制项如何影响其输出，以及模型是如何训练的、其训练数据来自何处。解释还可以回答有关特定模型输出的问题，例如模型对该输出有多大把握、该输出的哪些部分

可能需要人工审查或修改、如何调整或修改输入或提示词以改变输出的属性，或者该输出还有哪些其他选项或替代方案。

3.2 价值敏感设计在减轻潜在危害方面的重要性

面对生成式人工智能技术的快速成熟和广泛采用可能造成的潜在危害，设计师需要保持敏感。尽管用于减轻这些危害的社会技术手段尚未形成，我们仍建议设计师在思考如何设计生成式人工智能应用时采用价值敏感设计方法 [34, 44]。通过明确识别生成式人工智能应用涉及的不同利益相关者和受影响方，并明确列举他们的价值观，设计师便能更审慎地判断有害模型输出、模型滥用以及人类岗位替代问题可能对这些利益相关者造成何种影响。

4 局限性与未来工作

生成式人工智能应用仍处于起步阶段，新的协同创作用户体验正在迅速涌现。因此，我们认为这些设计原则也尚处于起步阶段，可能遗漏了其他重要的设计原则、策略和/或用户目标。此外，尽管这些原则能为设计师做出具体设计决策提供有益指导，但仍需在现实环境中加以验证，以确保其清晰性和实用性。

5 结论

我们提出了一套适用于生成式人工智能应用的七项设计原则。这些原则以生成式变异性为背景，其关键特征在于：生成式人工智能应用会生成制品作为输出，而这些输出的性质可能各不相同（例如质量或特征有所差异）。这些原则着重于针对多个输出及其不完美性进行设计，针对对可能输出空间或范围的探索进行设计并在探索过程中保持人为控制，以及通过解释帮助建立对生成式人工智能应用的准确心智模型。我们还敦促设计师采取设计防范 措施，避免潜在危害：这些危害可能由有害的模型输出（例如产生不当语言或图像、强化现有刻板印象，或未能包容性地呈现不同群体）、滥用模型（例如制造虚假信息或伪造证据），或取代人类劳动者（例如设计取代 而非增强 人类劳动者的方案）造成。我们希望这些原则能帮助设计师在创建新颖的生成式人工智能应用时做出审慎的选择。

参考文献

[1] ACM。2023年。措辞至关重要：计算机行业中带有争议色彩术语的替代说法。检索于2023年1月4日，网址：https://www.acm.org/diversity-inclusion/words-matter[2] Mayank Agarwal、Jorge J Barroso、Tathagata Chakraborti、Eli M Dow、Kshitij Fadnis、Borja Godoy、Madhavan Pallan 和 Kartik Talamadupula。2020年。Project clai：将命令行构建为人工智能智能体的新环境。*arXiv* 预印本 *arXiv:2002.00762* (2020年)。[3] Mayank Agarwal、Kartik Talamadupula、Stephanie Houde、Fernando Martinez、Michael Muller、John Richards、Steven Ross 和 JustinD Weisz。2020年。代码翻译的质量估计与可解释性。载于 *NeurIPS 2020* 计算机辅助编程研讨会论文集（*NeurIPS 2020*）。[4] Joshua Albrecht、Ellie Kitanidis 和 Abraham J Fetterman。2022年。尽管表现"超越人类"，当前的大语言模型并不适合用于有关伦理与安全的决策。*arXiv* 预印本 *arXiv:2212.06295* (2022年)。[5] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N Bennett、Kori Inkpen 等。2019年。人机交互指南。载于*2019*年 *CHI* 人机交互系统中的计算机人因会议论文集。1–13。[6] Cecilia Aragon、Shion Guha、Marina Kogan、Michael Muller 和 Gina Neff。2022年。以人为中心的数据科学：导论。麻省理工学院出版社。

[7] Vijay Arya、Rachel KE Bellamy、Pin-Yu Chen、Amit Dhurandhar、Michael Hind、Samuel C Hoffman、Stephanie Houde、Q Vera Liao、Ronny Luss、Aleksandra Mojsilovic 等。2020年。AI Explainability 360：用于理解数据和机器学习模型的可扩展工具包。《机器学习研究杂志》21卷，第130期（2020年），1–6。[8] Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。2021年。论随机鹦鹉的危险：语言模型会不会太大？载于2021年 ACM公平性、问责制与透明度会议论文集。610–623。[9] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。2021年。论基础模型的机遇与风险。arXiv 预印本 arXiv:2108.07258 (2021年)。[10] Danah Boyd 和 Kate Crawford。2012年。大数据的关键问题：对一种文化、技术与学术现象的质询。《信息、传播与社会》15卷，第5期（2012年），662–679。[11] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，1877–1901。https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf[12] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z Gajos。2021年。是信任还是思考：认知强制功能可减少人工智能辅助决策中对人工智能的过度依赖。《ACM人机交互汇刊》5卷，CSCW1（2021年），1–21。[13] Matthew Butterick。2022年。GitHub Copilot诉讼。https://githubcopilotlitigation.com[14] Andres Campero、Michelle Vaccaro、Jaeyoon Song、Haoran Wen、Abdullah Almaatouq 和 Thomas W Malone。2022年。人机系统性能评估测试。arXiv 预印本 arXiv:2206.12390 (2022年)。[15] Nicholas Carlini、Daphne Ippolito、Matthew Jagielski、Katherine Lee、Florian Tramer 和 Chiyuan Zhang。2022年。量化神经语言模型中的记忆化。arXiv 预印本 arXiv:2202.07646 (2022年)。[16] Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Ulfar Erlingsson 等。2021年。从大语言模型中提取训练数据。载于第30届 USENIX安全研讨会（USENIX Security 21）。2633–2650。[17] Ruijia Cheng、Ruotong Wang、Thomas Zimmermann 和 Denae Ford。2022年。"这对我来说也行得通"：在线社区如何塑造软件开发者对人工智能驱动的代码生成工具的信任。arXiv 预印本 arXiv:2212.03491 (2022年)。[18] Vijil Chenthamarakshan、Payel Das、Samuel C Hoffman、Hendrik Strobelt、Inkit Padhi、Kar Wai Lim、Benjamin Hoover、Matteo Manica、Jannis Born、Teodoro Laino 等。2020年。Cogmol：利用深度生成模型进行针对特定靶点且具有选择性的 COVID-19 药物设计。arXiv 预印本 arXiv:2004.01215 (2020年)。[19] Vijil Chenthamarakshan、Payel Das、Inkit Padhi、Hendrik Strobelt、Kar Wai Lim、Ben Hoover、Samuel C. Hoffman 和 Aleksandra Mojsilovic。2020年。利用深度生成模型进行针对特定靶点且具有选择性的 COVID-19 药物设计。arXiv:2004.01215 [cs.LG] [20] Jaemin Cho、Abhay Zala 和 Mohit Bansal。2022年。Dall-eval：探测文本到图像生成式Transformer的推理能力与社会偏见。arXiv 预印本 arXiv:2202.04053 (2022年)。[21] Kyunghyun Cho、Bart Van Merriënboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。2014年。使用用于统计机器翻译的 RNN编码器-解码器学习短语表示。arXiv 预印本 arXiv:1406.1078 (2014年)。[22] Elizabeth Clark、Anne Spencer Ross、Chenhao Tan、Yangfeng Ji 和 Noah A Smith。2018年。机器参与的创意写作：关于口号与故事的案例研究。载于第23届智能用户界面国际会议。329–340。[23] Apple Computer。2022年。人机界面指南。https://developer.apple.com/design/human-interface-guidelines/guidelines/overview[24] Sasha Costanza-Chock。2020年。设计正义：构建我们所需世界的社区主导实践。麻省理工学院出版社。[25] Paul Denny、Viraj Kumar 和 Nasser Giacaman。2022年。与 Copilot 对话：探索如何运用提示工程，通过自然语言解决 CS1问题。https://arxiv.org/abs/2210.15157[26] Norman K Denzin、Yvonna S Lincoln、Linda Tuhiwai Smith 等。2008年。批判性与原住民方法论手册。Sage。[27] Sebastian Deterding、Jonathan Hook、Rebecca Fiebrink、Marco Gillies、Jeremy Gow、Memo Akten、Gillian Smith、Antonios Liapis 和 Kate Compton。2017年。混合主动式创意界面。载于2017年 CHI大会扩展摘要：人机交互系统中的计算机人因会议论文集。628–635。[28] Catherine D'ignazio 和 Lauren F Klein。2020年。数据女性主义。麻省理工学院出版社。[29] Carl DiSalvo。2022年。作为民主探究的设计：将实验性公民实践付诸实践。麻省理工学院出版社。[30] Upol Ehsan、Philipp Wintersberger、Q Vera Liao、Elizabeth Anne Watkins、Carina Manger、Hal Daumé III、Andreas Riener 和 Mark O Riedl。2022年。以人为中心的可解释人工智能（HCXAI）：不止于揭开人工智能的黑箱。载于 CHI 人机交互系统中的计算机人因会议扩展摘要。1–7。[31] Stephen M Fiore、Eduardo Salas 和 Janis A Cannon-Bowers。2001年。群体动力学与共享心智模型的发展。《组织中人们如何评价他人》234（2001年）。[32] Paul M Fitts、MS Viteles、NL Barr、DR Brimhall、Glen Finch、Eric Gardner、WF Grether、WE Kellum 和 SS Stevens。1951年。高效空中导航与交通管制系统的人因工程及附录1至3。技术报告。俄亥俄州立大学研究基金会，哥伦布。[33] Giorgio Franceschelli 和 Mirco Musolesi。2022年。生成式深度学习中的版权。《数据与政策》4（2022年）。[34] Batya Friedman 和 David G Hendry。2019年。价值敏感设计：以道德想象力塑造技术。麻省理工学院出版社。

[35] Werner Geyer、Lydia B Chilton、Justin D Weisz 和 Mary Lou Maher。2021年。HAI-GEN 2021：第二届人机协同生成模型创作研讨会。载于第26届智能用户界面国际会议伴会。15–17。[36] Lisa Gitelman。2013年。《原始数据是个矛盾说法》。麻省理工学院出版社。[37] Github。2021年。Copilot。检索于2021年8月3日，网址：https://copilot.github.com[38] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2020年。生成对抗网络。《ACM 通讯》63，11（2020年），139–144。[39] Imke Grabe、Miguel González-Duque、Sebastian Risi 和 Jichen Zhu。2022年。迈向协同创作式 GAN 应用中的人机交互模式框架。2022 年 ACM IUI 研讨会联合论文集，2022年3月，芬兰赫尔辛基（2022年）。[40] Cobus Greyling。2022年。提示工程、文本生成与大语言模型。https://cobusgreyling.medium.com/prompt-engineering-text-generation-large-language-models-3d90c527c6d5[41] Matthew Guzdial、Nicholas Liao、Jonathan Chen、Shao-Yu Chen、Shukan Shah、Vishwa Shah、Joshua Reno、Gillian Smith 和 Mark O Riedl。2019年。朋友、协作者、学生、经理：人工智能驱动的游戏关卡编辑器设计如何影响创作者。载于《2019 年 CHI 人机交互系统中的计算机人因会议论文集》。1–13。[42] Gillian R Hayes。2014年。通过实践求知：作为人机交互研究方法的行动研究。载于《人机交互中的认知方式》。施普林格，49–68。[43] Will Douglas Heaven。2022年。为什么 Meta 最新的大语言模型只在线存活了三天。https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/[44] David G Hendry、Batya Friedman 和 Stephanie Ballard。2021年。价值敏感设计作为形成性框架。《伦理与信息技术》23，1（2021年），39–44。[45] Eric Horvitz。1999年。混合主动式用户界面原则。载于《SIGCHI 计算系统中的计算机人因会议论文集》（美国宾夕法尼亚州匹兹堡）（CHI '99）。计算机协会，美国纽约州纽约市，159–166。https://doi.org/10.1145/302979. 303030[46] Stephanie Houde、Vera Liao、Jacquelyn Martino、Muller Muller、David Piorkowski、John Richards、Justin D Weisz 和 Yunfeng Zhang。2020年。生成式人工智能的商业（误）用例。载于与第25届智能用户界面国际会议（IUI 2020）同期举办的人机协同生成模型创作与用户感知型对话代理研讨会联合论文集。[47] Kalin Hristov。2016年。人工智能与版权困境。《Idea》57（2016年），431。[48] IBM。2023年。设计中的种族公平。检索于2023年1月4日，网址：https://www.ibm.com/design/racial-equity-in-design/[49] Maia Jacobs、Melanie F Pradier、Thomas H McCoy、Roy H Perlis、Finale Doshi-Velez 和 Krzysztof Z Gajos。2021年。机器学习推荐如何影响临床医生的治疗选择：以抗抑郁药选择为例。《转化精神病学》11，1（2021年），1–9。[50] Tristan E Johnson、Youngmin Lee、Miyoung Lee、Debra L O'Connor、Mohammed K Khalil 和 Xiaoxia Huang。2007年。衡量团队相关知识的共享程度：共享心智模型测量工具的设计与验证。《国际人力资源开发》10，4（2007年），437–454。[51] Benjamin Kaiser、Akos Csiszar 和 Alexander Verl。2018年。直接生成数控刀具路径的生成模型。载于《2018年第25届机电一体化与机器视觉实践国际会议》（M2VIP）。IEEE，1–6。[52] Shalini Kantayya。2020年。《编码偏见》。检索于2023年1月4日，网址：https://www.pbs.org/independentlens/documentaries/coded-bias/[53] Bennett Kleinberg 和 Bruno Verschuere。2021年。人类如何损害自动化欺骗检测的性能。《心理学学报》213（2021年），103250。[54] Steven Kollmansberger。2010年。帮助学生构建计算心智模型。载于第十五届计算机科学教育中的创新与技术年度会议论文集。128–131。[55] Max Kreminski、Isaac Karth、Michael Mateas 和 Noah Wardrip-Fruin。2022年。通过表达范围覆盖分析评估混合主动式创意界面。载于 IUI 研讨会。34–45。[56] Sumith Kulal、Panupong Pasupat、Kartik Chandra、Mina Lee、Oded Padon、Alex Aiken 和 Percy S Liang。2019年。Spoc：基于搜索的伪代码转代码。《神经信息处理系统进展》32（2019年）。[57] Q Vera Liao、Daniel Gruen 和 Sarah Miller。2020年。质疑人工智能：为可解释人工智能用户体验设计实践提供参考。载于《2020 年 CHI 人机交互系统中的计算机人因会议论文集》。1–15。[58] Q Vera Liao、Moninder Singh、Yunfeng Zhang 和 Rachel Bellamy。2021年。可解释人工智能简介。载于《2021 年 CHI 人机交互系统中的计算机人因会议扩展摘要》。1–3。[59] Antonios Liapis、Georgios N Yannakakis、Julian Togelius 等。2013年。Sentient Sketchbook：计算机辅助游戏关卡创作。载于 FDG。213–220。[60] Stephanie Lin、Jacob Hilton 和 Owain Evans。2021年。TruthfulQA：衡量模型模仿人类谎言的程度。arXiv 预印本 arXiv:2109.07958（2021年）。[61] Silvia Lindtner、Shaowen Bardzell 和 Jeffrey Bardzell。2016年。重塑创造的乌托邦愿景：技术解决主义之后的人机交互。载于《2016 年 CHI 人机交互系统中的计算机人因会议论文集》。1390–1402。[62] Vivian Liu 和 Lydia B Chilton。2021年。通过语义控制进行三维动物形状的神经符号生成。载于 IUI 研讨会。[63] Vivian Liu 和 Lydia B Chilton。2022年。文生图生成模型提示工程设计指南。载于 CHI 人机交互系统中的计算机人因会议。1–23。[64] Ryan Louie 、 Andy Coenen 、 Cheng Zhi Huang 、 Michael Terry 和 Carrie J Cai。2020年。通过人工智能引导工具与深度生成模型进行新手人工智能音乐协同创作。载于《2020 年 CHI 人机交互系统中的计算机人因会议论文集》。1–13。

[65] Todd Lubart。2005年。计算机如何成为创造过程中的伙伴：对特刊的分类与评述。《国际人机交互研究杂志》63，4-5（2005年），365–369。[66] Alexandra Lyn。2020年。高风险业务：美国量刑与保释程序中的人工智能和风险评估。SSRN 3831441（2020年）可查。[67] Michael A Madaio、Luke Stark、Jennifer Wortman Vaughan 和 Hanna Wallach。2020年。共同设计核对清单，以了解人工智能公平性面临的组织挑战与机遇。载于《2020 年 CHI 人机交互系统中的计算机人因会议论文集》。1–14。[68] Mary Lou Maher。2012年。计算创造力与集体创造力：谁在发挥创造力？载于 ICCC。Citeseer，67–71。[69] Mary Lou Maher、Brian Magerko、Dan Venura、Douglas Fisher、Rogelio Cardona-rivera、Nancy Fulda、Johannes Gooth、Minwoo Lee、David Wilson、James Kaufman 等。2022年。以人为中心、可解释的协同创造型人工智能：整合生成式人工智能与认知人工智能的研究计划。载于 ACM CHI 人机交互系统中的计算机人因会议。[70] John E Mathieu、Tonia S Heffner、Gerald F Goodwin、Eduardo Salas 和 Janis A Cannon-Bowers。2000年。共享心智模型对团队过程与绩效的影响。《应用心理学杂志》85，2（2000年），273。[71] Edvinas Meskys、Julija Kalpokiene、Paulius Jurcys 和 Aidas Liaudanskas。2020年。监管深度伪造：法律与伦理考量。《知识产权法与实践杂志》15，1（2020年），24–31。[72] Cade Metz。2022年。认识 GPT-3。它学会了编程（以及写博客和辩论）。（发表于2020年）。https://www.nytimes.com/2020/11/24/science/ artificial-intelligence-ai-gpt3.html[73] Jessica K Miller、Batya Friedman、Gavin Jancke 和 Brian Gill。2007年。设计中的价值张力：某企业群件系统的价值敏感设计、开发与应用。载于《2007 年 ACM 支持小组工作国际会议论文集》。281–290。[74] Christoph Molnar。2020年。《可解释机器学习》。Lulu.com。[75] Meredith Ringel Morris、Carrie J. Cai、Jess Holbrook、Chinmay Kulkarni 和 Michael Terry。2022年。生成模型的设计空间。载于 NeurIPS 2022以人为中心的人工智能研讨会论文集（NeurIPS 2022）。[76] Robert R. Morris。2023年。我们使用 GPT-3 为约 4,000 人提供了心理健康支持——以下是结果。检索于2023年1月7日，网址：https://twitter.com/RobertRMorris/status/1611450197707464706[77] Michael Muller、Plamen Agelov、Hal Daume、Q Vera Liao、Nuria Oliver、David Piorkowski 等。2022年。HCAI@NeurIPS 2022，以人为中心的人工智能。载于神经信息处理系统年度会议。[78] Michael Muller、Lydia B Chilton、Anna Kantosalo、Charles Patrick Martin 和 Greg Walsh。2022年。GenAICHI：生成式人工智能与人机交互。载于 CHI 人机交互系统中的计算机人因会议扩展摘要。1–7。[79] Michael Muller、Steven I. Ross、Stephanie Houde、Mayank Agarwal、Fernando Martinez、John T. Richards、Kartik Talamadupula 和 Justin D. Weisz。2022年。与（人工智能）编程伙伴一起喝柴茶：一篇关于软件工程生成式人工智能的设计虚构，107-122。载于 IUI 2022 研讨会联合论文集：APEx-UI、HAI-GEN、HEALTHI、HUMANIZE、TExSS、SOCIALIZE，与 ACM智能用户界面国际会议（IUI 2022）同期举办，线上活动，芬兰赫尔辛基，2022年3月21日至22日（CEUR研讨会论文集，第3124卷），Alison Smith-Renner 和 Ofra Amir（编）。CEUR-WS.org，107–122。[80] Michael Muller 和 Angelika Stroymayer。2022年。数据科学中的遗忘实践。载于《2022 年 CHI 人机交互系统中的计算机人因会议论文集》。即将出版。[81] Michael Muller 和 Justin Weisz。2022年。通过动态性和社会性扩展人机协作框架。载于《2022年面向工作的人机交互研讨会》。1–12。[82] Michael Muller、Justin D. Weisz 和 Werner Geyer。2020年。混合主动式生成式人工智能界面：生成式人工智能应用的分析框架。ICCC 2020研讨会：《协同创造系统的未来》。https://computationalcreativity.net/workshops/cocreative-iccc20/papers/ Future_of_co-creative_systems_185.pdf[83] Michael D Murray。2022年。生成式人工智能创作的作品与版权法。SSRN（2022年）可查。[84] Raja Parasuraman、Thomas B Sheridan 和 Christopher D Wickens。2000年。人与自动化系统交互类型与层级模型。《IEEE 系统、人和控制论汇刊 A 部分：系统与人》30，3（2000年），286–297。[85] Claudio Pinhanez。2021年。揭示不确定性、引发不信任、避免解释：迈向人工智能伦理指南，发表于 HCAI@NeurIPS 2021研讨会。https://www.google.com/url?q=https%3A%2F%2Farxiv.org%2Fabs%2F2112.01281&sa=D[86] Claudio Pinhanez。2022年。故障、语言使用与怪异错误：BRL 对话代理研究的过去、现在与未来，发表于 IBM剑桥研究实验室 Guess演讲系列。[87] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。利用 CLIP潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv:2204.06125（2022年）。[88] Anais Resseguier 和 Rowena Rodrigues。2021年。将伦理视为对情境的关注：人工智能伦理建议。《Open Research Europe》1，27（2021年），27。[89] Laria Reynolds 和 Kyle McDonell。2021年。面向大语言模型的提示编程：超越少样本范式。载于《2021 年 CHI 人机交互系统中的计算机人因会议扩展摘要》。1–7。[90] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。利用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。10684–10695。[91] Janus Rose。2022年。Facebook 撤下其全新"科学人工智能"项目，因为它漏洞百出、糟糕透顶。《Vice》（2022年11月）。检索于2023年1月6日，网址：https://www.vice.com/en/article/3adyw9/facebook-pulls-its-new-ai-for-science-because-its-broken-and-terrible

[92] Steven I Ross、Fernando Martinez、Stephanie Houde、Michael Muller 和 Justin D Weisz。2023。《程序员助手：用于软件开发的大语言模型对话交互》。载于第 28 届智能用户界面国际会议。[93] Baptiste Roziere、Marie-Anne Lachaux、Lowik Chanussot 和 Guillaume Lample。2020。《编程语言的无监督翻译》。载于 NeurIPS。[94] Devansh Saxena、Karla Badillo-Urquiola、Pamela J Wisniewski 和 Shion Guha。2021。通过儿童福利案例研究构建的公共部门高风险算法决策框架。《ACM人机交互汇刊》5，CSCW2（2021），1–41。[95] Matthias Scheutz、Scott A DeLoach 和 Julie A Adams。2017。人机团队共享心智模型的构建与应用框架。《认知工程与决策期刊》11，3（2017），203–224。[96] Isabella Seeber、Eva Bittner、Robert O Briggs、Triparna De Vreede、Gert-Jan De Vreede、Aaron Elkins、Ronald Maier、Alexander B Merz、Sarah Oeste-Reiß、Nils Randrup 等。2020。机器作为队友：关于人工智能与团队协作的研究议程。《信息与管理》57，2（2020），103174。[97] Thomas B Sheridan 和 William L Verplank。1978。人类与计算机对水下遥操作设备的控制。技术报告。麻省理工学院剑桥分校人机系统实验室。[98] Ben Shneiderman。2020。以人为中心的人工智能：可靠、安全且可信。《国际人机交互期刊》36，6（2020），495–504。[99] Ben Shneiderman。2021。以人为中心的人工智能。《科学与技术议题》37，2（2021），56–61。[100] Ben Shneiderman。2022。《以人为中心的人工智能》。牛津大学出版社。[101] Auste Simkute、Aditi Surana、Ewa Luger、Michael Evans 和 Rhianne Jones。2022。面向学习的 XAI：缩小“新”专家与“老”专家之间的数字鸿沟。载于 2022 年北欧人机交互会议附属论文集。1–6。[102] Karan Singhal、Shekoofeh Azizi、Tao Tu、S Sara Mahdavi、Jason Wei、Hyung Won Chung、Nathan Scales、Ajay Tanwani、Heather Cole-Lewis、Stephen Pfohl 等。2022。大语言模型编码了临床知识。arXiv 预印本 arXiv:2212.13138（2022）。[103] Katta Spiel。2021。“他们为什么都痴迷于性别？”——穿越技术基础设施的（二元）性别导航。载于 2021 年交互系统设计会议。478–494。[104] Angie Spoto 和 Natalia Oleynik。2017。混合主动式创意界面库。检索于 2021年6月19日，网址：http://mici.codingconduct.cc/[105] Catherine Stupp。2019。诈骗者利用人工智能模仿 CEO 声音，酿成一起罕见的网络犯罪案件。《华尔街日报》（2019年8月）。检索于 2023年1月6日，网址：https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402[106] Jiao Sun、Q Vera Liao、Michael Muller、Mayank Agarwal、Stephanie Houde、Kartik Talamadupula 和 Justin D Weisz。2022。通过基于情景的设计探究生成式人工智能用于代码的可解释性。载于第 27 届智能用户界面国际会议。212–228。[107] Ilya Sutskever、Oriol Vinyals 和 Quoc V Le。2014。利用神经网络进行序列到序列学习。《神经信息处理系统进展》第 27 卷（2014）。[108] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力机制就是你所需要的一切。《神经信息处理系统进展》第 30 卷（2017）。[109] James Vincent。2022。编程问答网站 Stack Overflow 暂时禁止人工智能生成的答案。检索于 2023年1月6日，网址：https: // www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers[110] Patrick von Platen。2020。如何生成文本：使用不同的解码方法，通过 Transformer 生成语言。Hugging Face 博客（2020年3月）。检索于 2023年1月6日，网址：https://huggingface.co/blog/how-to-generate[111] Dakuo Wang、Justin D Weisz、Michael Muller、Parikshit Ram、Werner Geyer、Casey Dugan、Yla Tausczik、Horst Samulowitz 和 Alexander Gray。2019。数据科学中的人机协作：探索数据科学家对自动化人工智能的看法。《ACM人机交互汇刊》3，CSCW（2019），1–24。[112] Qiaosi Wang、Koustuv Saha、Eric Gregori、David Joyner 和 Ashok Goel。2021。迈向人机交互中的相互心智理论：语言如何反映学生对虚拟助教的认知。载于 2021 年 CHI 人机交互系统中的计算机人因会议论文集。1–14。[113] Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh 等。2021。语言模型造成伤害的伦理与社会风险。arXiv 预印本 arXiv:2112.04359（2021）。[114] Justin D Weisz、Mary Lou Maher、Hendrik Strobelt、Lydia B Chilton、David Bau 和 Werner Geyer。2022。HAI-GEN 2022：第三届人机协同生成模型创作研讨会。载于第 27 届智能用户界面国际会议。4–6。[115] Justin D Weisz、Michael Muller、Stephanie Houde、John Richards、Steven I Ross、Fernando Martinez、Mayank Agarwal 和 Kartik Talamadupula。2021。不必追求完美？代码翻译中的人机伙伴关系。载于第 26 届智能用户界面国际会议。402–412。[116] Justin D Weisz、Michael Muller、Steven I Ross、Fernando Martinez、Stephanie Houde、Mayank Agarwal、Kartik Talamadupula 和 John T Richards。2022。合作更好吗？对人工智能辅助代码翻译的评估。载于第 27 届智能用户界面国际会议。369–391。[117] Alex Wilkins。2022。人工智能文生图生成器会让插画师失业吗？《新科学家》（2022年5月）。[118] Shuo Yang、Kai Shu、Suhang Wang、Renjie Gu、Fan Wu 和 Huan Liu。2019。社交媒体上的无监督虚假新闻检测：一种生成式方法。载于 AAAI人工智能会议论文集，第 33 卷。5644–5651。