Dispute Settlement 争议解决, 2023, 9(6), 3418-3427
Published Online November 2023 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2023.96466

Hans汉斯

图像生成模型的法律性质界定

马佳祺，董星宏，周羽枫

上海政法学院刑事司法学院，上海

收稿日期：2023年10月21日；录用日期：2023年11月18日；发布日期：2023年11月24日

摘　要

以扩散模型和经扩散模型训练而成的海量的、个性化的预训练模型和微调模型为代表的图像生成式AI在图像生成、修改与替换等领域一枝独秀。核心技术开源、“一键式”便捷操作并且兼顾“质”与“量”是现阶段图像AI生成的重要特征。作为生成式AI在淫秽图像制作领域的产物，淫秽图像不仅数量上泛滥，淫秽程度和引发的风险也持续提升，继而使得技术滥用的危害不断扩大。而预训练模型和微调模型正是目前图像生成式AI被广泛应用于淫秽图像制作领域的关键，现行行政法规将具有图像生成能力的模型及相关技术定义为生成式人工智能技术，根据部分预训练模型和微调模型的特点以及在虚拟淫秽图像生成过程中所发挥的作用，应当承认直接认定为淫秽电子信息的合理性，从而摆脱技术中立原则的干扰，更好地遏制特定内容淫秽图像的泛滥、维持网络信息的健康和安全。

关键词

生成式人工智能，图像生成，潜在扩散模型，预训练模型，微调模型

Definition of Legal Nature of Image Generation Model

Jiaqi Ma, Xinghong Dong, Yufeng Zhou

School of Criminal Justice, Shanghai University of Political Science and Law, Shanghai

Received: Oct. 21st, 2023; accepted: Nov. 18th, 2023; published: Nov. 24th, 2023

Abstract

Image generation AI, which is represented by diffusion model and massive, personalized pre-training model and fine-tuning model trained by diffusion model, is unique in the fields of image generation, modification and replacement. Open source core technology, “one-click” convenient operation and taking into account “quality” and “quantity” are important features of image AI generation at this stage. As a product of generative AI in the field of obscene image production, obscene

文章引用: 马佳祺, 董星宏, 周羽枫. 图像生成模型的法律性质界定[J]. 争议解决, 2023, 9(6): 3418-3427.
DOI: 10.12677/ds.2023.96466

images are not only flooded in number, but also the degree of obscenity and the risk caused by it continue to increase, which makes the harm of technology abuse continue to expand. The pre-training model and fine-tuning model are the key to the widespread application of image-generating AI in the production of pornographic images. The current administrative regulations define models and related technologies with image generating ability as generative artificial intelligence technologies. According to the characteristics of some pre-training models and fine-tuning models and their roles in the generation of virtual pornographic images, We should recognize the rationality of directly identified as obscene electronic information, so as to get rid of the interference of the principle of technical neutrality, better curb the proliferation of pornographic images of specific content, and maintain the health and safety of network information.

Keywords

Generative Artificial Intelligence, Image Generation, Potential Diffusion Model, Pre-Training Model, Fine-Tuning Model

1. 前言

2022 年末，随着 ChatGPT 的公测，生成式 AI 迅速“出圈”，成为全世界范围内的热点。生成式 AI 在过去十年内迅速发展，同时覆盖文本、图像、音频和视频生成等领域。现阶段的生成式 AI 技术主要为企业所掌握，用户生成的内容也相对可控。但是，以稳定扩散模型为代表的部分生成式 AI 技术已经开源，由于可以在家用计算机上充分实现训练和生成功能，该技术在辅助更多人完成任务的同时，也为犯罪活动提供了明显的帮助。

2. 图像 AI 生成现状

(一) 生成式 AI 概述

生成式 AI(AIGC)技术根据给定的帮助训练和指导模型完成任务的人类指令，利用 GAI 算法生成相应的满足指令的内容[1]。

生成式 AI 可以追溯到 20 世纪中叶的隐马尔可夫模型(Hidden Markov Models, HMMs)和高斯混合模型(Gaussian Mixture Models, GMMs)。早期模型性能有限，难以应对复杂且多样化的任务。随着深度学习技术的发展，生成式对抗网络(Generative Adversarial Network, GAN)和扩散模型(Diffusion Model, DM)等深度合成模型弥补了传统模型的缺点，实现了生成过程的高精度控制和生成结果的高质量保证。

编码器－解码器结构(Transformer)是目前深度合成模型主流的基础结构，前者(Encoder)提取输入序列、生成隐藏表示(Hidden Representations)，后者(Decoder)提取隐藏表示、生成输出序列。对输入内容的有效识别和对输出内容的精准把控高度依赖模型的训练能力，以生成式对抗网络为例，存在生成器(Generator)和辨别器(Discriminator)的博弈，经过对真实数据分布的学习，前者持续生成新的数据，后者则根据真实数据分布加以辨别，通过不断博弈提高生成数据的能力。此外，深度合成模型还可以依托庞大的数据集进行大规模的、复杂的分布式训练，提高生成效率和质量的同时，应对复杂环境和需求的能力也显著提升。

深度合成模型也存在一些问题，例如 GAN 由于辨别器限制存在不稳定性，“泛化”能力[1]亦有限[2]，扩散模型则训练成本高昂、耗能巨大。随着扩散模型领域的技术突破，新的潜在扩散模型(Latent Diffusion Models, LDM)既维持了对 GAN 的优势，又通过忽略高频信息和低维空间训练等方法，显著减轻了训练的成本和运行的耗能，使得生成式 AI 技术的真实性训练和虚拟内容生成的门槛降低到消费级 GPU (Graphics Processing Unit，图形处理器)。

另外，生成式 AI 模型可以分为单模态和多模态[2]模型[3]，原始的 GAN 就是单模态模型，单模态模型的输入和输出内容形式上要求一致。CLIP (即 Contrastive Language-Image Pre-training，对比文本–图像预训练模型)等模型的出现为多模态模型的后续发展提供了助力，以图像 AI 生成为例，多模态模型可以完成输入文字、输出图像的操作，结合既有的 AIGC 技术对人类反馈的强化学习，图像 AI 生成兼顾实用性和真实性的同时，个性化生成能力也高度强大。

因此，在 ChatGPT“出圈”的同时，图像生成式 AI 也飞入寻常百姓家。

(二) 扩散模型的演进

1) 扩散模型的发展：从原始扩散模型到潜在扩散模型

随着扩散模型领域的技术突破，扩散模型成为目前最先进、最普遍的生成模型。

扩散模型通过注入噪声逐渐破坏数据，再通过学习逆转过程实现样本生成[4]。换言之，扩散模型存在两个阶段，正向扩散阶段逐步扰乱数据分布，反向扩散阶段通过反向学习逐步恢复数据分布。

原始的扩散模型存在明显的缺点，数据采样效率不及 GANs，最大化似然差，数据泛化能力有限。因此，扩散模型的后续发展既侧重数据“破坏–还原”过程的优化，也针对原始扩散模型的缺点重点改进。

去噪概率扩散模型(Denoising Diffusion Probabilistic Model, DDPM)是扩散模型发展的第一个里程碑。DDPM 的第一个阶段，即“前向链”，将数据扰动为噪声；第二个阶段，即“反向链”，将噪声转换为数据。DDPM 在数学层面首次经过严谨推导、论证，将扩散模型理论正式运用实践，也将“正向加噪–反向降噪–训练”的模式固化，为扩散模型的后续发展提供了条件。

潜在扩散模型则是第二个里程碑，LDM 通过将扩散模型应用于预训练自编码器的潜在空间[3] [5]，实现了扩散模型训练质量、灵活和成本的最佳化[6]。LDM 使用自编码器在低维空间采样数据，显著提高计算效率，再通过感知压缩技术忽略高频信息，保留低频信息，继而在低维潜在空间训练后者。尽管数据压缩程度提高，还原度方面反而更进一步，因此，LDM 在保证训练的稳定性和生成的高质量的同时，又使得扩散模型能够在消费级 GPU 上快速生成图像，随着 LDM 的代码开源，奠定了扩散模型未来大范围推广的基础。

2) 扩散模型的关键：预训练模型和微调模型

经过 DDPM 和 LDM 等阶段的发展，扩散模型成为现阶段生成式 AI 在图像领域的主流模型，LDM 的开源使得部分后续模型也选择开源，以稳定扩散模型(Stable Diffusion, SD)为代表的开源模型因此区别于非开源模型。非开源模型如 Midjourney，Imagen (Google)和 DALL E (OpenAI)等，不仅调用的数据样本丰富，生成内容也可控，一经问世就引起消费者广泛关注，在商业绘画领域存在广阔前景。开源模型则因为在消费级显卡上实现本地部署和免费运行的优势，在图像个性化生成领域也绝非付费模型所企及。

稳定扩散模型是基于 LDM 的文本–图像生成模型，主要应用于图像生成领域，支持文字和图像内容的输入。作为 LDM 的后续模型，稳定扩散模型在技术层面并没有创新，但是，开源的稳定扩散模型凭借本地部署和免费运行(10G 显存消费级显卡)的优势吸引了海量的参与者。

[1]“泛化”能力，即学得模型适用于新样本的能力，具有强泛化能力的模型能很好地适用于整个样本空间。
[2]模态是指事物发生或存在的方式，多模态是指两个或者两个以上模态各种形式的组合。多模态机器学习是指建立模型使机器从多模态中学习各个模态的信息，并且实现各个模态信息的交流和转换。
[3]编码器将输入压缩成较少数量的元素或位。输入被压缩到最大限度，即潜在空间。

稳定扩散模型发挥作用还必须调用一个额外模型，即预训练模型，后者是以大规模数据集为基础经过扩散模型所训练的模型。通过对海量数据的学习，高效捕获信息知识，继而以参数形式储存，预训练模型以泛用性为特征[7]。另外，由于预训练模型可以被调整，还存在微调模型，顾名思义，是针对预训练模型的调整模型。

LoRA，即大型语言模型的低秩适应[8]，是典型的微调模型，原本是微软公司引入的处理大型语言模型微调问题的技术，通过冻结预训练模型权重并且在 Transformer 结构的每一层注入可训练秩分解矩阵，可以将下游任务的可训练参数减少 10,000 倍[9]。经过训练的 LoRA 模型可以在扩散模型调用预训练模型时介入，发挥微调的作用，灵活地生成特定图像，从而避免耗费大量资源专门训练一个新的预训练模型。由于微调模型是在低资源环境下所训练的小型模型，应用于图像生成时可能存在生成图像不收敛、欠拟合、过拟合等问题，亦即泛化能力和训练效果不佳，以 DreamBooth 为代表的正则化图像[4]的微调模型则避免了过拟合等问题，提高了微调模型训练和生成的稳定性。

个性化生成是图像 AI 生成的重要特点，微调模型则是个性化生成的必要途径。个性化生成不是随机生成，而是可控生成，否则只能通过不断试错获取目标图像，从而降低效率。ControlNet 的出现强化了图像 AI 生成的可控性，作为一种神经网络结构，ControlNet 通过对预训练模型的控制，使得图像生成阶段支持额外的输入条件[10]。ControlNet 同样旨在微调预训练模型，从而扩大图像生成的可控范围，同 LoRA 作用接近，只是具体可控的范围不同。

无论是预训练模型、还是微调模型，抑或是 ControlNet 等针对预训练模型的微调技术，都可以在互联网上被共享。随着稳定扩散模型的开源，海量的预训练模型随之产生，图像 AI 生成的门槛持续降低，生成内容的可控范围也在维持质量的同时不断扩大，风险也因此随之而来。

如同曾经“出圈”的 DeepFake，图像生成技术同样可以制造虚假和违法信息、数据，并推动其传播，任何人都可能沦为图像生成技术被滥用的受害者。稳定扩散模型的强大性能、海量资源和低门槛性则进一步强化了风险，考虑到图像生成式 AI 的实用性和普遍性，阻止技术普及显然是天方夜谭，如何有效应对相应风险将是严峻的挑战。

3. 图像 AI 生成在淫秽图像制作领域的风险检视

(一) 虚拟图像的制作效率显著提高

图像生成模型显著提高了淫秽图像的制作效率。传统淫秽图像的制作一般存在两种模式，一种是现实拍摄他人的淫秽色情行为，再以图片或录像的形式存储，也可以由制作人直接以绘画的方式创作；另一种是利用图像编辑软件、绘图软件、视频制作软件等计算机程序制作、合成新的淫秽图像或者修改真实淫秽图像。

传统的淫秽图像制作模式，无论现实制作，还是利用计算机程序制作，效率都相当有限。拍摄淫秽照片、录像前必须准备工具、场地，并且和被拍摄者联系，拍摄后还可能存在一定的后期处理工作。淫秽图像的现实创作则依赖制作人的绘画技艺，创作时间同作品的精细度成反比。即使是计算机技术介入后，利用电脑技术制作新的淫秽图像、或者修改既有的淫秽图像，效率亦提升不明显。

但是，随着扩散模型的发展，图像 AI 生成技术极大地促进了淫秽图像的制作效率。只需要设定好参数并且调用必要的预训练模型，就可以利用 Stable Diffusion 等图像生成模型制作海量的淫秽图像，又由于微调模型的介入，淫秽图像制作效率显著提高的同时，还保证了淫秽图像的高质量。利用 Stable Diffusion 等图像生成模型训练一个预训练模型需要花费大量的时间和资源，但是，预训练模型可以通过 Checkpoint、Safetensors 等技术予以保存，继而在互联网上被共享，从而整体上降低了成本。

[4]以是否图像正则化为标准，区分直接微调模型和以 DreamBooth 为代表的微调模型。

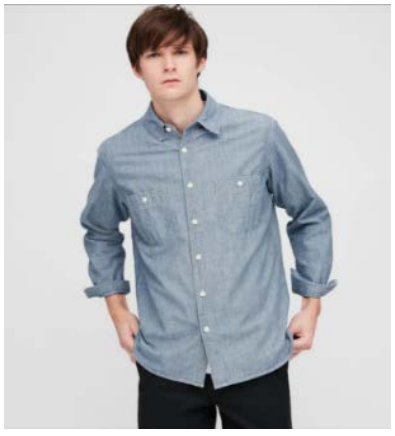
因此，通过下载并调用预训练模型和微调模型，可以迅速以 Stable Diffusion 等图像生成模型为基础构建一个图像制作平台，通过输入提示词源源不断地输出全新的图像，制作效率同图像信息数量成反比，图像所包含的信息(亦即输入条件)越复杂、图像的分辨率越高，制作速度越慢，此外，计算机显卡性能对图像生成效率的影响也相当明显。常规情况下，一张 1024*1024 像素的图片往往只需要 10~20 秒即制作完毕，而正则化图像的微调模型如 DreamBooth 等进一步缩短了图像输出的间隔，由于欠拟合、过拟合、不收敛等问题已经被初步解决，生成模型所输出的图像也并非是对既有数据的简单复制，而是符合创作的条件，亦即属于全新的图像。

由此可见，图像 AI 生成技术使得淫秽图像的制作效率产生了质变，曾经由制作人花费数小时创作的淫秽图像，如今缩短到以秒或分钟计，即使考虑到输出图像在质量把控方面仍不稳定的现实，整体效率的显著提升依然是毫无疑问的。

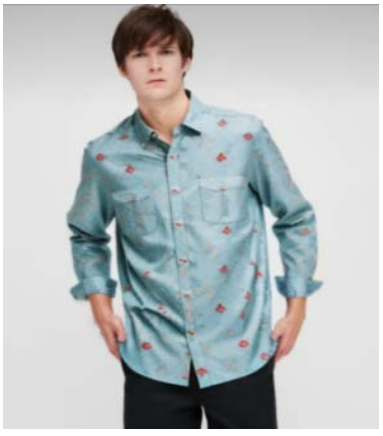
(二) 真实图像的深度伪造效果增强

图像生成模型不仅提升了虚拟图像的制作效率，也强化了真实图像的修改、替换效果，从而使得图像生成模型被滥用时的社会危害性扩大。深度伪造技术(Deep Fake)旨在借助人工智能技术，尤其是深度学习对音频、视觉或文本内容等进行制造和修改，从而达到技术开发者或者技术使用者的某种特定目的[11]。几乎所有传播技术和网络技术在其出现之后都是首先在色情产业获利[12]，深度伪造技术亦然，针对公众人物的淫秽色情信息伪造是深度伪造技术被滥用的重要表现。正如国外有学者指出：“深度伪造压倒性地描绘并且创造了出现在淫秽图片中的人未经同意的色情内容。”[13]，2019 年 3 月，深度伪造软件 DeepNude 上架，仅上传一张女性照片，DeepNude 就可以在极短的时间内生成一张女性全身裸露的图片[14]。此类软件可以通过特定的算法帮助用户轻易制造他人的淫秽色情图片、视频，特别是女性，从而侵害其名誉权和肖像权，甚至成为色情报复和敲诈勒索的工具[15]。

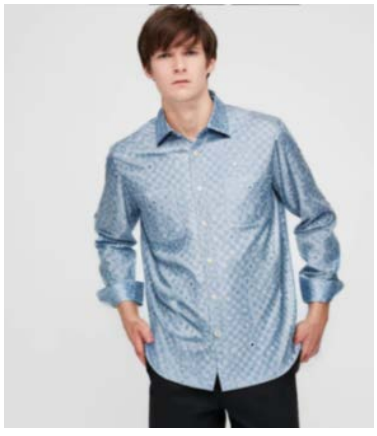
而当 Stable Diffusion 等图像生成模型加载 Partial Redrawing (局部重绘)模型后，可以完全覆盖甚至超越传统深度伪造技术的功能。例如，为一张图片中的人物替换衣服花纹时，首先为衣服部分添加蒙版，继而在正向指示区[5]输入指向特定花纹的提示词(Tag)，即可在极短的时间内完成替换，效果如下图：



(原图)



(tag 为 flower，point 的效果)



(tag 为 lattice 的效果)

图像生成模型对传统深度伪造技术的突破在于真实图像的伪造更加灵活和便捷，个性化调整的特点更加突出。制作人可以肆意调整蒙版覆盖的区域，甚至在不同区域多次重绘。以上图为例，如果在衣服

[5]在 Stable Diffusion Webui 界面中，存在正向指示区和反向指示区，均可以输入特定的提示词。前者增加某些内容的权重，后者降低某些内容的权重，继而通过手动操作强化图像生成的可控性。

部分添加蒙版并输入提示词“Nude”，加载特定预训练模型后，生成的图像将在衣服部分裸露，并且裸露的内容可以通过修改提示词随意调整。结合另一个微调工具 ControlNet，还可以调整人物动作，继而生成不同姿势的淫秽图片，再通过既有的动态图片技术，进一步制作动图、甚至短视频。

由此可见，图像生成模型驱动下的深度伪造是个性化、多样性和精细度兼顾的深度伪造，尤其在该模型更擅长的图片生成领域，是对传统深度伪造技术的进一步强化。

(三) 特殊淫秽图像潜藏泛滥危机

扩散模型时代的图像 AI 生成，输出内容的多样性和可控性不仅日益提高，同真实图像的界限也不断缩短。从 GAN 时期到稳定扩散模型起步阶段，尤其是人物的图像生成方面，手部和脚部图像始终欠拟合，涉及特定人物图像时，过拟合的现象也存在，预期图像的输入条件过多、或者过于复杂时，输出图像甚至直接不收敛，以至于社会一般人可以轻松辨别出一张图像是否由人工智能制作。随着稳定扩散模型的开源及推广，技术不断迭代与进步，以 Stable Diffusion 为例，首先加载特定的微调模型，再输入“Reality”、“High Quality”等正向提示词，可以显著优化图像的拟真度，部分预训练模型更以输出高精度图像为特点；而通过输入反向提示词以避免欠拟合、过拟合等问题时，效果也更加明显。因此，AI 生成图像同真实图像曾经不可逾越的鸿沟已经近在咫尺。

儿童色情向来是不可逾越的禁区。传统儿童色情图像的制作，或者真实拍摄、或者绘图创作，图像生成模型时代，通过收集儿童图像和淫秽色情图像数据，可以专门训练生成儿童淫秽色情图像的预训练模型。如果预训练模型可以输出儿童图像或者淫秽色情图像，只要在微调模型部分调用一个输出淫秽色情图像或者儿童图像的模型，也可以输出儿童淫秽色情图像，因此，Stable Diffusion 在 2.0 版本后选择“一刀切”，直接删除了儿童和 NSFW 的训练数据[16]，但是，AI 生成儿童淫秽色情图像的可能性并没有因此中止，用户依然可以自行训练。

此外，公众人物淫秽图像的 AI 生成也不局限于真实图像的伪造，通过一个特定人物的预训练模型，再调用其他微调模型，也能够输出以假乱真的全新人物图像，输出图像可以是淫秽色情图像，也可以是正常图像，继而局部重绘。

概而言之，图像生成模型使得特定的、社会危害性更加严重的淫秽色情图像存在泛滥的可能性。儿童淫秽色情图像，即使是虚拟、合成图像，在境外部分国家也是刑法禁止的内容，至于公众人物的淫秽色情图像，对当事人名誉权的侵害也更加明显。仅因为存在此类风险，继而禁止技术传播是不切实际的，更何况人工智能作为前沿科技，立法尚起步，学理层面的研究亦有限，如何科学、合理规制，以及寻求法律和技术的平衡，皆是难题。结合图像生成模型的基本原理和淫秽色情图像的生成过程，预训练模型是可行的突破口，作为规制预训练模型的前提，明确预训练模型的性质是必要的。

4. 预训练－微调模型的刑法规范界定和规制

(一) 预训练－微调模型的性质界定

1) 预训练模型和微调模型具备技术和电子信息的双重性

运用图像 AI 生成技术实现淫秽图像的“一键生成”包含几个关键要素，首先是以稳定扩散模型为代表的图像生成模型，其次是由前者所训练的预训练模型和微调模型，最后，还包括以 ControlNet 为代表的具备微调功能的其他技术。2023 年 8 月 15 日起施行的《生成式人工智能服务管理暂行办法》将三者统称为技术[6]，实践中也存在统称为模型的情况，但是，三者存在明显差异。

图像生成模型，本质上属于算法范畴，以稳定扩散模型为例，稳定扩散模型作为慕尼黑大学和 Runway

[6]《生成式人工智能服务管理暂行办法》第二十二条本办法下列用语的含义是：(一) 生成式人工智能技术，是指具有文本、图片、音频、视频等内容生成能力的模型及相关技术。

的研究人员基于开源数据集 LAION-aesthetics 所训练的文本–图像生成模型，总耗时 15 万 NVIDIA A100 GPU 小时，核心文件却是后缀为 py 的 Python 脚本文件，可见，稳定扩散模型是对海量图像深度学习后形成的算法。即使模型所学习的海量图像中可能存在色情内容，一方面，根据“技术中立”[7] [17]和利益衡量原则，图像生成模型明显促进社会生产力，何况对人体结构的学习也是保证算法质量的必要前提，另一方面，基础的图像生成模型作为算法，并不具备输出淫秽色情图像的能力。

因此，以稳定扩散模型为代表的图像生成模型应当属于中立的技术。同理，ControlNet 等技术也具备中立性。例如 Partial Redrawing，本质上也是 Python 脚本文件，亦即算法。若脱离具备淫秽色情图像生成能力的预训练模型，Partial Redrawing“淫秽色情伪造”的功能将失效。

预训练模型作为图像生成模型基于大规模数据集的训练产物，本质上也依然是数据集。预训练阶段，通过对海量数据的学习，捕捉数据信息，继而以参数(亦即权重)形式储存，由于深度学习的本质就是学习、优化权重的值，使其达到一个最优解的状态[18]，预训练模型储存数据是有序的，即参数、权重的值最优化的状态。同理，微调模型作为预训练后续阶段的模型，本质上同预训练模型一致，只是目的从泛用性专向针对性转变。

预训练模型和微调模型的常见格式为 pt、ckpt 和 safetensors，pt 和 ckpt 分别是保存 PyTorch 和 Tensorflow 模型结构和参数等内容的文件格式，由于 pt 和 ckpt 文件存在被植入恶意代码的风险，safetensors[8] 取而代之，safetensors 同样是数据存储格式，甚至更加纯粹，仅由张量数据构成。

概而言之，图像生成模型所训练的预训练模型本质上仍然是数据集，无论是预训练模型，还是微调模型，pt、ckpt 和 safetensors 文件代表了两种模型内部数据的具体储存形式。

因此，区别于 Python 脚本文件格式的图像生成模型(基础模型)和其他算法，预训练模型(包括微调模型)并非纯粹的技术，它同作为训练对象的数据集联系更紧密，可以视作经过处理和以特殊形式储存的数据集，预训练模型从而具备了被解释为电子信息的基础。

2) 部分微调模型属于淫秽电子信息

预训练作为图像 AI 生成的关键环节，以稳定扩散模型等图像生成模型为基础，对海量图像数据高效采集信息，继而以特定形式储存。因此，采集信息的对象，即原始的图像数据将一定程度上决定输出的内容。无论是商业化的图像生成程序所加载的预训练模型，抑或是在网络上公开提供下载的预训练模型，基本都包含了对色情图像的学习，因此，均具备输入特定提示词后生成淫秽色情图像的能力。区别在于，前者部分可以放弃对色情图像的学习，部分则可以利用技术手段屏蔽特定提示词，如 Nude，NSFW 等，从而使得提供的图像生成服务符合法律法规的要求。至于后者，尽管随着潜在扩散模型的开源，预训练的门槛理论上已经降低至家用计算机，由于涉及海量数据处理，时间和成本仍然高昂，专门以色情图像为对象训练预训练模型的可能性甚微。

因此，目前一个以 GB 为单位的预训练模型事实上很可能包含了淫秽色情图像的信息，但是比重基本可以忽视，也因此，同稳定扩散模型一样，即使预训练模型属于电子信息，也不能轻易认定为淫秽电子信息。

微调模型却不然，区别于泛用性的预训练模型，微调模型强调针对性，存储的图像信息也更少。因此，专门训练一个应用于淫秽图像制作的微调模型是可行的，此微调模型也必然以一定数量的淫秽色情图像信息为内容，甚至全部都是，微调模型从而具备认定为淫秽电子信息的前提。

根据刑法第 367 条，“书刊、影片、录像带、录音带、图片及其他淫秽物品”属于淫秽物品的载体，

[7] “技术中立原则”，即某项产品或者技术是被用于合法用途还是非法用途，并非产品或者技术的提供者所能预料和控制，因而不能因为产品或技术成为侵权工具而要求提供者为他人的侵权行为负责。
[8] Safetensors 是一种用于安全储存张量(Tensors)的新型简单格式。

后续司法解释又将“视频文件、音频文件、电子刊物、图片、文章、短信息等互联网、移动通讯终端电子信息和声讯台语音信息”[9]解释为“其他淫秽物品”，从而扩大了淫秽物品载体的范围。可见，淫秽物品最初以有形物为载体，信息化的淫秽物品随后产生，两者都是淫秽内容的直接载体，而以淫秽物品种子文件为代表的部分淫秽电子信息则区别于前两者，种子文件所指向的淫秽内容以下载为前提，种子文件本身只是储存于云端的数据化淫秽内容和使用者之间的媒介。尽管如此，种子文件“由于详细记载了淫秽视频的特征信息，利用 P2P 软件即可通过这些信息直接下载淫秽视频”[19]，从而在司法实践中，被广泛认定为淫秽内容的载体。

纵观刑法和司法解释对淫秽物品载体的规定，微调模型似乎不能够对应任何一项。但是，根据淫秽电子信息的成立条件，亦即载体是物、可独立存在的客观实体、包含或者产生淫秽信息的源发物[20]，服务于淫秽图像生成的微调模型全部符合，只不过“源发物”一项，由于图像生成模型和预训练模型原则上的中立性，亦即预训练模型本身输出淫秽图像的可能性，必然要求特定微调模型在淫秽图像制作中的作用，从而排除一些微调模型。

因此，当微调模型被训练用于专门服务于淫秽图像制作时，它可能具备淫秽电子信息的条件。

另外，微调模型不是淫秽信息的直接载体，从而与淫秽信息的种子文件具备一定的共性，相较于种子文件，特定的微调模型同样记载了淫秽图像的特征信息，而以图像生成模型和预训练模型为前提也可以再结合微调模型生成特定的淫秽色情图像。因此，将种子文件解释为淫秽电子信息的理由同样适用于特定微调模型的解释。

最后，还存在一些微调模型，例如，以特定公众人物图像、儿童图像、孕妇图像为训练对象的微调模型，尽管没有以淫秽图像为训练对象，从而不包含淫秽图像信息，但是，该类微调模型一旦同预训练模型结合，通过输入特定提示词也能生成特定的危害性更严重的淫秽图像，尽管不能解释为淫秽电子信息，从数据收集违法的角度，也应当予以控制。

3) 作为淫秽电子信息的部分微调模型属于“无受害人”的虚拟淫秽物品

即使是淫秽电子信息，仍然普遍指向现实的淫秽、色情内容，或者现实发生的性犯罪，再或者广义的性剥削行为，从而存在具体的、特定的受害人。由此可见，从法益理论的角度，淫秽物品可能诱发更多的性侵或暴力犯罪，那么损害原则就可能支持对淫秽物品的禁止[21]；而从道德主义的角度，绝大多数淫秽物品的存在本身即是性剥削的证明。

然而，虚拟淫秽物品却没有被害人，或者说具体的被害人。法益理论和道德主义所支持的论据并不足以成为禁止、管制虚拟淫秽物品的理由。考虑到淫秽、色情物品还存在积极的一面，例如为人们提供了从正式的社会生活中解脱出来的机会，对于缓解性压抑和生活压力发挥作用[22]，虚拟淫秽物品的社会危害性似乎显著下降。

但是，随着图像 AI 生成技术的成熟，虚拟淫秽物品同现实淫秽物品的界限不断缩短，通过以真人图像为训练对象的预训练和微调模型所生成的图像几乎能够以假乱真。由于图像生成的高效性，特定内容的虚拟淫秽物品，如以未成年人、特定公众人物、孕妇等为内容的淫秽物品的数量也迅速增长。

因此，一方面，量变引起质变，即使虚拟淫秽物品的社会危害性有限，在图像 AI 生成时代，应当强化对虚拟淫秽物品的规制。另一方面，即使开始规制微调模型(以及可能的预训练模型)，也应当重点把握特定内容。

(二) 预训练－微调模型的规制与未来进路

现行司法解释将制作、复制、出版、贩卖、传播淫秽电子图片之行为的入罪门槛规定为“一百件

[9]《最高人民法院、最高人民检察院关于办理利用互联网、移动通讯终端、声讯台制作、复制、出版、贩卖、传播淫秽电子信息刑事案件具体应用法律若干问题的解释》(法释[2004] 11 号)第九条。

以上”[10]，即使是淫秽种子文件，由于种子文件对淫秽内容的指向性明确，定量时也可以依据所指向的下载内容。微调模型则不然，所指向的并非具体淫秽内容，而是不特定的、潜在的淫秽内容，一方面，作为训练对象的既有淫秽图像难以还原，另一方面，特定的微调模型一旦结合图像生成模型和预训练模型，可以输出无穷的、全新的淫秽图像。

因此，微调模型作为淫秽内容的载体，如果以潜在内容为基准，那么所承载的淫秽物品数量不是空白，就是无穷大，前者明显掩盖了特定微调模型的危害，从而轻纵相关涉嫌违法、犯罪的行为，而后者则明显违反罪责刑相称原则。而从训练对象角度，亦即微调模型所储存的淫秽图像数据信息入手，同样困难重重，因为事后还原基本是天方夜谭。

倘若训练人同淫秽图像的制作人属于同一人、或者存在共犯情形，定量时可以参照已经被制作的淫秽物品数量。但是，如果仅针对传播特定微调模型的行为，倘若训练行为同制作行为属于完全独立的两个行为，由于定量的复杂性，似乎只能将传播训练模型的行为认定为无罪，从而使得微调模型即使被认定为淫秽电子信息，也将失去实际意义。

因此，倘若将微调模型等具有图像生成能力的模型认定为淫秽电子信息，如何解决定量的问题依然值得思考。

5. 结语

人工智能技术的广泛应用给人们带来无限的发展机遇和便利，也深刻影响着人们的行为和思考方式、价值和道德观念，从而引发了大量潜在的风险。习近平总书记指出，“要全面提升技术治网能力和水平，规范数据资源利用，防范大数据等新技术带来的风险”[23]。图像 AI 生成技术给人类无限的艺术遐想，也使得我们时刻面对道德、伦理和法律风险。对于图像 AI 生成技术，特别是众多具有图像生成的模型及相关技术，应当尽快从法律角度予以精确识别，从而帮助此类模型和技术更好地造福社会、服务于社会主义现代化建设，同时避免图像 AI 技术被恶意滥用于特殊淫秽图像的制造，从而引发更大的风险。

参考文献

[1] Cao, Y.H., Li, S.Y., Liu, Y.X., Yan, Z.L., Dai, Y.T., Yu, P.S. and Sun, L.C. (2023) A Comprehensive Survey of AI-Generated Content (AIGC). https://arxiv.org/abs/2303.04226

[2] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 3.

[3] 刘建伟, 丁熙浩, 罗雄麟. 多模态深度学习综述[J]. 计算机应用研究, 2020(6): 1601-1614.

[4] Yang, L., Zhang, Z.L., Song, Y., Hong, S.D., Xu, R.S., Zhao, Y., Zhang, W.T., Cui, B. and Yang, M.-H. (2023) Diffusion Models: A Comprehensive Survey of Methods and Applications. https://arxiv.org/abs/2209.00796

[5] 西班扬·达斯, 乌米特·卡卡马克. 自动机器学习入门与实践: 使用 Python [M]. 谢琼娟, 译. 武汉: 华中科技大学出版社, 2019: 201.

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2023) High-Resolution Image Synthesis with Latent Diffusion Models. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10674-10685. https://arxiv.org/abs/2112.10752 https://doi.org/10.1109/CVPR52688.2022.01042

[7] Han, X., Zhang, Z.Y., Ding, N., Gu, Y.X., Liu, X., Huo, Y.Q., *et al*. (2023) Pre-Trained Models: Past, Present and Future. https://arxiv.org/abs/2106.07139

[8] Cuenca, P. and Paul, S. (2023) Using LoRA for Efficient Stable Diffusion Fine-Tuning. https://huggingface.co/blog/lora

[9] Hu, E., Shen, Y.L., Wallis, P., Allen-Zhu, Z.Y., Li, Y.Z., Wang, S.A., Wang, L. and Chen, W.Z. (2023) LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/2106.09685

[10]《最高人民法院最高人民检察院关于办理利用互联网、移动通讯终端、声讯台制作、复制、出版、贩卖、传播淫秽电子信息刑事案件具体应用法律若干问题的解释(二)》第一条。

[10] Zhang, L.M. and Agrawala, M. (2023) Adding Conditional Control to Text-to-Image Diffusion Models. https://arxiv.org/abs/2302.05543

[11] 李明鲁. “深度伪造”的刑法治理路径[J]. 科技与法律(中英文), 2021(6): 40-47+73.

[12] 桑本谦. 网络色情、技术中立与国家竞争力——快播案背后的政治经济学[J]. 法学, 2017(1): 79-94.

[13] Olson, A. (2022) The Double-Side of Deepfakes: Obstacles and Assets in the Fight against Child Pornography. *Georgia Law Review*, **56**, 865-892.

[14] Langa, J. (2021) Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes. *Boston University Law Review*, **101**, 761-802.

[15] Harris, D. (2019) Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review*, **17**, 99-128.

[16] Stability AI (2023) Stable Diffusion 2.0 Release. https://stability.ai/blog/stable-diffusion-v2-release

[17] 陈洪兵. 论技术中立行为的犯罪边界[J]. 南通大学学报(社会科学版), 2019(1): 58-65.

[18] 邢彤彤, 孙仁诚, 邵峰晶, 隋毅. 深度学习中的权重初始化方法研究[J]. 计算机工程, 2022, 48(7): 104-113.

[19] 张远金. 贩卖淫秽视频种子文件的定性和数量认定[J]. 人民司法(案例), 2017(20): 43-45.

[20] 周新. 略论淫秽电子信息犯罪的界定[J]. 社会科学家, 2012, 27(9): 103-106.

[21] 罗翔. 论淫秽物品犯罪的惩罚根据与认定标准——走出法益理论一元论的独断[J]. 浙江工商大学学报, 2021(6): 82-90.

[22] 董玉庭, 黄大威. 论传播淫秽、色情物品犯罪的刑事立法政策——以无被害人犯罪为视角[J]. 北方法学, 2014(1): 60-67.

[23] 中华人民共和国中央人民政府. 习近平主持中共中央政治局第十二次集体学习并发表重要讲话[EB/OL]. https://www.gov.cn/xinwen/2019-01/25/content_5361197.htm, 2019-01-25.