Standard Article

AI-replicas as ethical practice: introducing an alternative to traditional anonymisation techniques in image-based research

Qualitative Research
1–26

DOI: 10.1177/14687941241308705
journals.sagepub.com/home/qrj

Tobias Kamelski
Lingnan University, Hong Kong

Francisco Olivos
Lingnan University, Hong Kong

Abstract
This article introduces the use of AI-replicas as an alternative to traditional anonymisation methods in image-based qualitative research. It emphasises the ethical and practical dilemmas posed by current anonymisation methods, such as distortion or loss of emotional and contextual information in images, and proposes the use of AI-replicas to preserve the integrity and authenticity of visual data while ensuring participant anonymity. The article outlines the technological foundations of generative artificial intelligence (AI) and the practical application of Stable Diffusion to generate AI-replicas for anonymisation and fictionalisation purposes. Furthermore, it discusses the potential biases present in generative AI to suggest ways to mitigate these biases through careful prompt engineering and participatory approaches. The introduced approach aims to enhance ethical practices in visual research by providing a method that ensures participant anonymity without compromising the data's qualitative richness and interpretative validity.

Keywords
image anonymisation, generative AI, stable diffusion, research methodology, qualitative research, visual research, ethics, privacy

Ethical and research-practical challenges in using images for qualitative research

Over the last three decades, there has been a growing interest in the visual dimensions of social life, and research methods that utilise visual data of some kind have become legitimate and

Corresponding author:
Tobias Kamelski, Department of Sociology and Social Policy, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong.
Email: tobiaskamelski@ln.hk

powerful tools in the social sciences (Spencer, 2023). These visual research methods can include images that already exist, that are created by the researcher, or by those who are being researched (Rose, 2022). In this ever-evolving scenario, researchers in the field of visual research began discussing how ethical judgements arise within their field in response to a call for situated visual ethics (e.g. Prosser, 2000). At present, despite being contested and unresolved (Wiles et al., 2012), the debate has bourgeoned (Allen, 2015; Hannes and Parylo, 2014; Miller, 2015; Nutbrown, 2011; Spencer, 2020; Wiles et al., 2012), and it is expected to grow due to the 'blossoming of possibilities' using found images (Rose, 2022) from image-based online platforms, online archives, YouTube and online video games, amongst others. More recently, the COVID-19 pandemic and the transition to remote platforms to communicate across time and space have fostered even more the growth of visual research across fields (e.g. Burgos-Thorsen and Munk, 2023; MacEntee et al., 2022; Polat, 2022).

Against this background, one of the central debates in this pervasive ethical discussion in visual research methods is anonymisation as a default condition and a key concern of researchers (Wiles et al., 2012) and ethics committees (Allen, 2015). The anonymity of sources is a condition of the ethical standard of confidentiality: 'Visual researchers do not disclose […] identifiable information concerning their research participants' (Papademas and IVSA, 2009: 254). Researchers who use research-generated, participant-generated or found images are often mandated to ensure the anonymity of the research participants when navigating ethics committees or dissemination gatekeepers (Miller, 2015). The same requirement is also applied to researchers conducting video analysis, who also 'transcribe' videos into images for reporting purposes (e.g. Gan et al., 2020; Laurier, 2014). Recognising the limitations and ethical dilemmas posed by traditional anonymisation techniques in visual research (e.g. adding black bars, blurring features or applying abstraction filters) this article aims to introduce and evaluate the potential of generative artificial intelligence (AI) for creating AI-replicas. By leveraging contemporary technologies, we seek to enhance the process of anonymising visual data in qualitative research, demonstrating how novel generative AI technology can offer novel approaches to image anonymisation and fictionalisation. Thus, we introduce the opportunities and challenges of utilising AI for image anonymisation for any researcher working with image data.

This paper is structured as follows. The section 'Issues in the anonymisation of visual data' of this article discusses general issues of anonymisation of visual data. The section 'Generative AI and AI-replica as ethical solutions' further elaborates on the potential application of AI-replicas as an innovative alternative to traditional anonymisation methodologies whilst also elaborating on the technological foundations of this method. The section 'Practical application of stable diffusion for generating AI-replicas' exemplifies the generative process and the ability to control the generation of AI-replicas based on several example cases. The 'Discussion' section addresses AI-replicas and their effect on data richness and potential bias embedded in the models that power generative AI.

Issues in the anonymisation of visual data

Visual researchers report that ethical reviewers require the distortion of images beyond recognition, even after research participants consented to use their unmodified images (Spencer, 2020). Similarly, collecting consent or directly identifying picture producers are not always possible (Kozinets, 2020: 164; Tiidenberg and Baym, 2017). Image

data, such as found footage or online data originating from platforms without a real-name policy, are common in fields such as communication science or netnography (see Kozinets, 2020: 164). Such data often cannot be traced back to producers who can be asked for permission; however, they may still contain personally identifiable information about the producer or other individuals who need to be anonymised. These concerns can be explained by the longevity of images in public circulation and the potential reutilisation of these images as found images in the future, which makes standards of anonymisation even more salient (Wiles et al., 2012). These difficulties shape research practices because researchers may avoid engaging in image-based research anticipating these challenges (Miller, 2015; Nutbrown, 2011), and participants in visual participatory research may adapt their behaviour to prevent ethical breaches (Hannes and Parylo, 2014).

Nevertheless, apart from discouraging researchers and shaping research practices, anonymity requirements are in tension with other aspects of visual research (Allen, 2015). The analysis and scientific communication of images usually rely on textual transcriptions in form of 'protocol sentences' (Bohnsack, 2008: 3; Popper, 1959: 43, 59ff.). Protocol sentences are textual descriptions that capture elementary statements about complex empirical facts, such as pictures, and serve as the foundation for subsequent research. These texts represent a fundamental layer of abstraction, as every analysis is presented based (fully or partially) on textual reductions and not the visual data itself. This scenario means that a text-mediated analysis of images always involves a reduction and abstraction that can approximate only the perceptive gaze or 'seeing view' (*sehendes Sehen*) of the interpreter (Bohnsack, 2008: 23). The perceptive gaze refers to the observer's ability to engage with the meaning of a picture on a deeper level and is conceptualised in opposition to the 'recognising gaze', in which the beholder simply acknowledges what is objectively depicted by a picture (see Imdahl, 1996). Given this abstracting quality, preserving representative elements and the argumentative plausibility of images as self-referential systems (cf. Luhmann, 1995) are inherent problems of these protocol sentences. Protocol sentences are ideally presented alongside the pictures they describe to verify their accuracy. This approach inevitably poses problems when reporting anonymised image data, as anonymisation can impede the verification of the presented analysis and possible interpretations, as pivotal aspects might become unrecognisable (Nutbrown, 2011: 8). Therefore, anonymisation (e.g. adding black bars, blurring features or applying abstraction filters) must be applied to image features that are irrelevant or at least less relevant to the research questions. For example, applying black bars over a subject's eyes is less relevant if the latter's gaze is less important for the interpretation. Researchers and their audience, therefore, engage to a certain degree in a 'good-faith relationship' in which the anonymised elements are accepted to be irrelevant for interpretation or to be compensated for by the textual description.

Moreover, techniques that modify images present other challenges to image-based research. Firstly, pixelation, blurring, black-out strips, cropping and other fictionalisation techniques (Jordan, 2014; Nutbrown, 2011; Wiles et al., 2012) silence research participants, which contradicts the visual research's potential ethical aim of giving voice and empowering often unheard groups, such as children or marginalised populations. As explained by Allen (2015), blurring pictures in research on sexual and gender minorities can undermine the politics of naming and violate their identity rights. Thus, anonymity might contradict the principles of respect and dignity of research participants. The

arguments against anonymisation mainly rely on this tension between researchers' paternalism and respondents' agency (for a summary, see Wiles et al., 2012).

Secondly, the anonymisation techniques used by visual researchers could also bring epistemological challenges. Although researchers must protect research participants at all costs, traditional fictionalisation techniques could move us away from the original truth (Nutbrown, 2011) and introduce additional distortions to what is already generated by and through the researcher's interpretations and subsequent textual abstractions. These techniques add additional layers of obscurity between what is analysed and what is reported; these scenarios could negatively affect the integrity and validity of visual methods (Allen, 2015; Nutbrown, 2011). For instance, emotions that are expressed through facial expressions or body language cannot be reported to the audience if the images are blurred, which could hinder the validity of the interpretation because it is not possible to present the evidence. In an extreme case, the concern for confidentiality may entail the anonymisation of locations, which is more complex than what can be achieved with traditional methods. Researchers cannot insert black strips all over an image to anonymise a school in image-based research with children; simply, they cannot report such images. These concerns about the validity and reliability of the findings based on anonymised images not only affect what can be or cannot be reported in a conference presentation but also make journal editors and peer reviewers hesitant about publishing the studies.

An additional epistemological challenge is anticipating criticism from ethical review boards regarding the use of images. Visual researchers not only engage in significant self-censorship in the dissemination of results (Wiles et al., 2012) but also in avoiding research questions that could be more sensitive for ethical review boards. Similar effects have been reported for photovoice participants who adopted different coping strategies (Hannes and Parylo, 2014). Individuals could avoid certain conflict-ridden situations or affect their behaviour by shooting landscapes but not by taking photos of themselves or by taking photos of friends even if they are motivated to do otherwise. The tensions surrounding how data can be generated and reported have direct epistemological consequences when researchers attempt to acquire knowledge through the voices of their participants. Therefore, traditional anonymisation techniques could bias the process of generating knowledge in visual research as early as the definition of a research question or topic of interest.

Thirdly, anonymisation techniques are rarely effective and can even produce undesired results (Moore, 2012). As described by Allen (2015), in her experience with photo-anonymisation in school-based visual research, leaving out other students' faces has the effect of directing the attention to the bodies of students; thus, the sexualisation of young people which the ethics board intended to prevent was actually magnified by the anonymisation process. Banks (2001) also suggested that facial modifications such as black bars could invoke the association of anonymised subjects with criminal perpetrators due to common media connotations. Similarly, the pixelation of faces could be associated with victimhood.

Beyond these potential unintended consequences and epistemological and ethical challenges, offering anonymity to research participants is incontestable, and visual researchers must protect research participants even if this entails compromising their research (Nutbrown, 2011). In this context, this article leverages novel computational methods and introduces generative AI to create AI-replicas as an alternative to traditional

fictionalisation techniques for image anonymisation. We aim to assist in research implementing this technique to disseminate high-quality image data with strict observance of ethical principles. To the best of our knowledge, this novel AI technique has not been identified amongst the 'good ethical practices in relation to visual methods' (Wiles et al., 2012: 42). This technique enables us to anonymise images whilst retaining distinctive key elements for reporting meaningful data. Its use covers scenarios such as reporting on online data in which gathering informed consent is difficult or practically impossible, reporting visual material that shows vulnerable subjects, for example, political refugees or simply as an additional layer of anonymisation. Furthermore, this guide can facilitate the application of generative AI in participatory visual research where subjects can undertake the creation of AI-replicas themselves. This can be imagined as an augmented photovoice method (see also Wang and Burris, 1997), in which participants create and authorise AI-Replicas as vehicles to express underlying structures of knowledge. As we will explain, AI can be used to create detailed and accurate visual representations based on textual descriptions. This approach enables participants to generate visual support for otherwise textual accounts.

Generative AI and AI-replica as ethical solutions

Application potential of AI-replicas

The AI-replicas were inspired by Markham's (2012) ethical fabrication method in qualitative research. In the context of textual data, Markham suggested creating composite accounts by choosing 'representative elements from the dataset and composing a new original not traceable back to the originals' (7) and devising fictional narratives that preserve the original data's argumentative plausibility. The AI-replicas contain computer-generated versions of the original images, preserving crucial elements such as composition, colour and gestures and essentially replacing images of subjects with realistic, fabricated replicas. These can be placed on a spectrum between fictionalised abstractions and composite accounts on one end and more closely aligned anonymisations on the other end. Similar to an artist reimagining a scene (e.g. courtroom sketches), AI-replicas retain distinctive key elements, ensuring visual authenticity and anonymity. Utilising deep learning and AI-powered methods is not entirely new and has been tested in medical contexts (Yang et al., 2022) or when dealing with 'computer vision tasks requiring real-world data' (Douglas, 2022), for example, for processing Google Street View Data. Generative adversarial networks (GANs) offer sophisticated alternatives to classical anonymisation methods by generating realistic yet artificial faces and even bodies and superimposing them on a given picture (Hukkelås and Lindseth, 2022; Maximov et al., 2020). The concept of AI-replicas introduced in this work is based on diffusion models and differs from these anonymisation methods in that they enable users to anonymise more than simply human actors whilst enabling the creation of abstractions based on original data. Thus, whether AI-replicas are employed as an advanced form of anonymisation that aligns closely with the source material or as composite accounts remains unclear. The latter can be of particular use when reporting typologies developed based on image data or for the visualisation of otherwise textual accounts.

AI-replicas are conceived as a means to report image data in a qualitative research context and must therefore not be mistaken for a replacement of the latter throughout the initial analysis and interpretation. Similar to the aforementioned protocol sentences, the AI-replicas are,

to a certain degree, a product of the situated interpretation (Berger and Luckmann, 1967: 7) and the 'interpretative authority' of the research (Markham, 2012: 15). Similar to paraphrased sentences in an interview transcript, AI-replicas can be understood as paraphrased visual information that is subject to the situated perceptions and critical judgment of the researcher. As such, designing AI-replicas requires thorough consideration of the nature of the original material, the situational requirements for anonymisation, and the findings to be reported. We address these aspects in the section 'Discussion' after we illustrate the application potential of AI-replicas in the section 'Practical application of stable diffusion for generating AI-replicas'.

AI-replicas offer numerous benefits that could contribute to overcoming some of the concerns highlighted above. Firstly, AI-replicas preserve emotional and expressive content, which is crucial in visual data where facial features are key biometric identifiers (Smith et al., 2018; Smith and Miller, 2022). Standard anonymisation techniques often involve redactions such as adding black bars or blurring or pixelating faces. Such alterations significantly alter the original material, impacting the reader's understanding of the researcher's interpretation. AI-replicas can generate entirely new pictures whilst maintaining accurate facial expressions.

Secondly, AI-replicas provide additional customisation and control over the reported data. These tools enable researchers to modify aspects that are potentially irrelevant to specific research objectives, questions, methodologies and reporting requirements (e.g. race or location). Thirdly, in contrast to traditionally anonymised images, AI-replicas exhibit enhanced legibility. The excessive application of filters and other alterations can hinder the interpretability of visual materials. The AI-replicas can help maintain clear visual legibility by keeping the images free of obscuring artefacts. These aspects can be implemented without compromising the privacy of the data subjects whilst presenting a relatively low technological entry barrier.

Advancing visual research: an introduction to generative AI and its role in crafting AI-replicas

Users presently have a vast selection of publicly available generative AI tools for image generation, including Adobe Firefly, Lexica, Dreamlike, Midjourney, DALL•E (Betker et al., 2023; Mishkin et al., 2022) and Stable Diffusion (Stability-AI, 2023c). Whilst all these applications are chiefly designed for artistic purposes, a pivotal difference between these services is that most of them are commercialised cloud-based process pipelines that offer a streamlined experience for paying customers. Stable Diffusion is an open-source project that can be executed on local computers. This approach provides greater degrees of control at the cost of user convenience. Although suitable hardware is required to effectively run the software (i.e. a contemporary GPU with 10 GB VRAM), Stable Diffusion avoids the drawbacks of purely web-based services. Firstly, considering local computations, Stable Diffusion is constantly available and does not incur additional costs per image generation. Secondly, Stable Diffusion allows the use of fine-tuned models (checkpoints) trained on custom datasets.[1] Thirdly, Stable Diffusion ensures a higher degree of privacy because it is not a cloud-based service. A local installation of Stable Diffusion avoids all possible concerns of losing control over images because all the data generated and used for generation stay on a local computer. Finally, the open-source distribution and community support enable the use of additional neural networks for greater control over the generated images.

Stable Diffusion is a diffusion model (Rombach et al., 2022; Weng, 2021) that utilises 'noise reversal' or 'denoising' (see also Yang et al., 2024: 6). The modelling starts with a canvas of a random pattern of pixels (i.e. random noise), which is then refined across several sampling steps in which Stable Diffusion improves on the image via iterative refinement (sampling steps). Serving as a simple analogy, one can imagine noise as unrefined clay, which is carefully shaped through several (sampling) steps until it has a recognisable shape, for example, the sculpture of a face. Throughout this process, new values are imposed on the pixels until they resemble visual patterns associated with a text prompt. This process allows the model to create detailed and accurate visual representations of the text (Rombach et al., 2022), which has enormous potential for applications in text-based qualitative research.

Stable Diffusion can be compared to machine learning and neural networks and less so for generative pretrained transformers such as ChatGPT.

Figure 1 illustrates this denoising process based on the following prompts with 1, 4, 6 and 35 sampling steps:

> Two men are squatting on the dirt floor, each holding a rooster. The men are focused intently on their birds, preparing for a traditional cockfight. They are dressed casually denoting a warm, tropical climate. Around them, a group of onlookers forms a semicircle, all standing and watching with keen interest, photography.

Beyond text-to-image generation, Stable Diffusion enables image-to-image generation for guided synthesis (Mao et al., 2023; Meng et al., 2022; Wang et al., 2021). This procedure allows for the control of the generation of replicas based on input images as image-to-image generation, optionally under consideration of text prompts. The input images will predominantly affect the composition and overall colour of the output images. This is achieved by generating new images not only with random noise but based on input images with layers of added noise. Both methods, text-to-image and image-to-image, as well as their potential application scenarios, will be further elaborated upon in the section 'Practical application of stable diffusion for generating AI-replicas'.

The creation of replicas with Stable Diffusion is therefore understood as image generation based on the composition of input images, notwithstanding that text-to-image generation also has application potential, as illustrated below. The output quality and possible resolutions depend chiefly on the models used (Bie et al., 2023; Podell et al., 2023: 16; Yang et al., 2024), which are trained on extensive image datasets (e.g. Beaumont, 2022; Common Crawl, 2023; CompVis, 2022). Models are trained for image generation

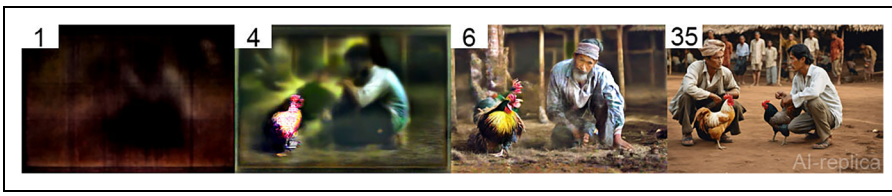

Figure 1. Sampling and denoising.

by adding noise to the training data and effectively learning the difference between the original images and several rounds of additional noise. During image generation, this process is reversed, and the images are generated by sequentially removing noise.

A plethora of free and paid models exist on the internet, with the option to train and fine-tune custom models. The majority of these models are fine-tuned versions of the popular Stable Diffusion v1-5 checkpoint (Rombach et al., 2022). Fine-tuning involves training existing models on specific types of images, such as human faces, landscapes, animals or art. For the anonymisation of human data, models based on the recently published, new Stable Diffusion-XL 1.0 base model are recommended. It must be noted that models used for generative AI are biased (Heikkilä, 2023), as the training data are vast but ultimately limited. These biases must be further reduced in the long term by pruning and refining the base models. An experiment by Nicoletti and Bass (2023, based on Stable Diffusion v1-5) revealed that Stable Diffusion displays racial bias when prompted to generate images of high- and low-income jobs. These model-conditioned biases can be avoided through more detailed prompts (e.g. the specification of skin colour). Nonetheless, such biases also lead to inherent limitations when generating images of vulnerable individuals and thus, in training datasets, less represented groups, for example, groups with physical disfigurations. In the following section, we illustrate the potential of Stable Diffusion to generate AI-replica that preserve the content of the originals. This account, not a comprehensive guide, aims to illustrate and inspire the application of AI-replicas in image anonymisation. These AI-replicas should not replace the original material for analysis but serve as substitutes in scenarios such as (1) circulating images in experiments, (2) supplementing field notes when visual documentation is unfeasible and (3) reporting image data whilst safeguarding privacy and maintaining legibility.

Practical application of stable diffusion for generating AI-replicas

The approach presented in this work is realised via the local installation of Stable Diffusion (Rombach et al., 2022; Stability-AI, 2023c) and the Stable Diffusion web UI (AUTOMATIC1111, 2023), both of which are publicly available via their GitHub repositories.[2] The web UI enables the use of Stable Diffusion via an intuitive browser interface that is accessible to users with minimal technical knowledge. In the section below, we introduce four application examples of Stable Diffusion for the creation of AI-replicas to illustrate its application potential.

Text-to-image generation

Figure 2 shows an example of a hypothetical case in which an image should be replaced with an AI-replica. Before and throughout the generation process, which aspect of the images is relevant must be balanced and must be reported. The focus may be on which aspects must be anonymised and what degree of abstraction and alteration is acceptable. This remains, of course, relative to the research questions and the employed research methodology and should be decided upon after the analysis of the original material.

Similar to chatbots such as ChatGPT, Bard or Poe, AI image generation tools operate in their most basic application on text prompts. The text prompt communicates to Stable

Figure 2. First example case for image anonymisation.

Diffusion <what> to generate. Natural language descriptions complemented with specific keywords usually work the best and ideally contain descriptions of the scene, the composition, the perspective, the focal length or the medium. A possible text prompt for the example picture could look as follows:

> A Caucasian man with a blue jacket, black shirt and black jeans standing next to a tree in a park holding a cell phone in his hand and looking at the camera with a serious look on his face, street and car in the background, canon 50 mm, light bokeh.

When generating pictures via text prompts alone, the most pivotal variable next to the prompt itself and the previously illustrated sampling steps (see Figure 1) is the classifier-free guidance (CFG) scale. The CFG scale determines how closely the generated images follow the text prompt and thus regulates how much of the text prompt finds its way into the final pictures and relies strongly on the model checkpoint used. A low CFG scale will effectively generate a random image (underfitting), whilst an extremely high value will inevitably create visual artefacts and potentially abstract outcomes (overfitting).

Using the previous prompt, Figure 3 illustrates how a higher CFG value will inevitably underscore all the elements of the prompt, with the CFG 7 variations resembling the original compositions the best. All the variations default to a tree with yellow leaves when not closely specified. These details can be adjusted using the text prompt, as illustrated in the adjusted version CFG 7* (i.e. 'next to a green tree'). These adjusted examples show that the colour of the leaves can easily be adjusted but also reveal that the similarity of CFG 7 with the composition of the original pictures was random, as CFG 7* deviates

further in terms of composition.[3] Depending on the case, one can argue that either of these variations already suffices as abstract representations of the original image but with a clear potential to improve the alignment with the original image.

These basic text-to-image generations demonstrate some of the inherent limitations of this technology. For example, whilst CFG 7 resembles the original picture superficially the most, the man was rendered with two mobile phones instead of one by chance. This phenomenon is related to the probability-based and thus partially random nature of Stable Diffusion and can be compensated for by continuous generations with different seed values.[4] Seed values are used to control the randomness of image generation. Using the same seed value with stable parameters will allow you to replicate the same image generation. The seed values are randomised by default. The generation of the same prompt with a different seed is illustrated by CFG 7*. However, this seed renders the man with a green jacket instead of a blue jacket and concentrates on the phone instead of the camera. This is an association error that can occur in complex prompts, in which different colours are assigned to different objects. These phenomena are related to the employed model and the complexity of the data on which it was trained. Similar to when multiple objects are used, these errors can be circumvented by generating the same images with different seed values. Whilst these examples illustrate the basic potential of generative AI, they also highlight its shortcomings concerning the predictability of the outcome that must be addressed before the generated images can be used as functional substitutes.

Text-to-image generations benefit from highly detailed prompts and, if possible, fine-tuned models that specialise in the desired outcome. This type of AI-replica is suitable when the pictures are supposed to be more abstract representations of findings, such as typologies or aggregated findings, and for the support of field notes or narratives that had no visual documentation in the first place. Similarly, visual support can be provided to textual descriptions of visual data recalled by participants (cf. Ward, 2017: 1653). Text-to-image generations are arguably less effective if one wishes to report on more specific aspects of an image that require stronger adherence to the original composition. For this sake, greater accuracy can be achieved via image-to-image generation.

Image-to-image generation

Whilst text-to-image generations begin with complete random noise, image-to-image generations start with an input image as the basis. In contrast to text-to-image generation, new images are not calculated purely by approximating the text prompt through the denoising of complete random noise. Instead, noise is applied to various degrees to the

Figure 3. CFG value comparison.

input image or a particular section of it (Meng et al., 2022). One can simply imagine applying clay to a sculpture to mould a new face along the lines of the underlying original. Contrasting this analogy, it must be stressed that the input image cannot be reconstructed or retrieved by any technical means from the generated images, as Stable Diffusion utilises noise and partially random transformations that cannot be reversed. Likewise, the output image does not contain any hidden information about the input images, except for its intended resemblance. The intensity to which noise is applied to the input image is regulated by the denoising strength. The denoising strength determines to what degree the original image will be reimagined toward the desired outcome as specified in the text prompt (the CFG scale still applies). A reduced denoising strength adds less noise, keeping the output closer to the input, whilst high values lead to more abstract output.

Figure 4 shows the image-to-image generation results with the original prompt by using three different denoising strengths (CFG value 7). The range of 0.35 to 0.51 strongly approximates the original pictures whilst displaying varying levels of alteration. 0.35 appears as a light alteration, 0.51 gives the impression of an enactment, whilst 0.91 appears as a reinterpretation of the original. Although quite similar, one can observe a subtle difference in the 0.51 variation, for example, an additional car in the background, the absence of a light post or a shift in the mobile phone's position. Based on our experience, the AI-replicas, in the presented 0.30–0.60 range, can be used as a substitute for the original picture for reporting purposes to retain essential compositional and expressive information (e.g. positioning of actors and objects in the frame, facial expressions, body language, perspective, light, etc.), while replacing identifiable environments and actors. AI-replicas generated with a denoising strength larger than 0.6 tend to be more abstract and thus serve more as thematic summaries or composite accounts describing an overall archetype identified in the data. Depending on the research objective, this approach could be preferable to presenting AI-replicas that align more closely with specific images. More research is needed to systematically identify an optimal range. However, the most suitable value is strongly related to the picture in question and the desired outcome. If necessary, a prompt can be used to further alter the picture.

Figure 5 illustrates how 'Caucasian male' in the original prompt was replaced with 'Asian male' and 'Caucasian female'. Whilst this example could be considered a critical alteration in the pictures, it serves the sole purpose of illustrating the flexibility and fidelity with which the AI-replicas can be altered to suit the situated needs of the researcher. This example shows that the text prompt is a secondary factor when working with

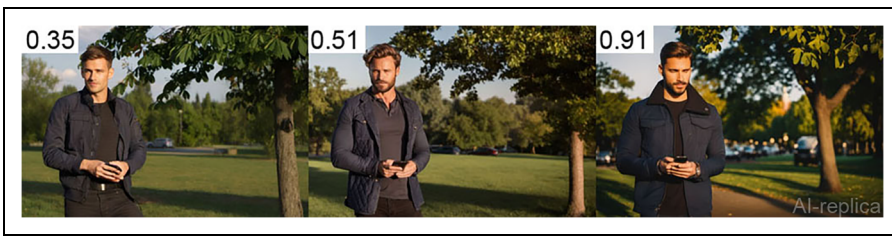

Figure 4. Denoising strength comparison.

Figure 5. Prompt alteration comparison.

image-to-image generations, which align closely with the original composition. However, a supporting text prompt allows us to alter key aspects of the image that are not relevant for the research question as well as the interpretation, and thus further support the anonymisation of the original data.

Image-to-image generation enables stronger alignment of the AI-replicas with the composition of the input image. This type of generation is suitable for AI-replicas that should communicate the nuances of a particular composition of the original pictures that could otherwise not be replicated or sufficiently approximated via simple text prompts alone. For example, two common reasons why a text prompt could not suffice could be the limitation of language in describing complex visual data or limitations in the training data of the used model. Limitation shall not mean that the model is not able to display it but that it struggles to generate the desired output based on the text prompt alone. The denoising procedure that generates the AI-replicas 'on top' of the original image can essentially replace everything within the picture with a semantic double, leading to full anonymisation. This procedure is our recommended default approach when generating AI-replicas. Furthermore, high denoising values allow for more abstracted AI-replicas whilst still having high control over the output image based on the original data material. Higher degrees of abstraction can be useful when the risk of exposing an informant is not rooted in biometric information alone but rather in iconic compositions that would give away the identity of the informant due to the alleged iconic singularity of a given picture, or subtle environmental cues. Whilst abstraction can be desirable in certain scenarios, as described above, it leaves the concern that images may lose the details and qualitative nuance that are deemed essential for the analysis. For even greater control over image generation, Stable Diffusion allows the use of additional conditions in the form of ControlNet models.

ControlNet

ControlNet (Zhang et al., 2023) enables conditions for image generation via a vast array of models and can be used by itself or in combination with image-to-image generation to exert more precise control over image generation. Given the considerable diversity among ControlNet models, a comprehensive introduction to all of them at this juncture is impractical. For image anonymisation, the two most important models are OpenPose for body poses and facial expressions and Depth for spaces. ControlNet preprocessors process input images and extract information that ControlNet models use to guide the

generation of new images. These parameters can be used for text-to-image and image-to-image generation. The benefit of using ControlNet is that it allows further abstraction when used in text-to-image generation and greater alignment with the input image in image-to-image generation.

Figure 6 exemplifies the application of the OpenPose model and preprocessor. The input image is on the left-hand side, whilst the abstraction on the right represents the body key points identified by the OpenPose model. The latter will be used to guide the image generation of Stable Diffusion. A prompt to describe the input image could be:

> Two men engaged in a friendly discussion whilst sitting at a round table in a modern library. The person on the left is a young Caucasian man with short brown hair wearing a black jacket and black pants. He is in a pensive pose with his hands clasped together, looking attentively at his interlocutor. The person on the right is a young Hispanic man with short black hair, a warm smile and a light stubble wearing a navy-blue and white sports jacket and black pants. They are surrounded by bookshelves filled with various books. On the table, there is a piece of paper with notes, a textbook and a blue aluminium can. The overall atmosphere is casual and intellectual, with a 35 mm eye-level angle.

Figure 7 illustrates the possible outcomes of image generation. Variation (1) is a replica purely based on the text prompt, whilst variation (2) uses the OpenPose model

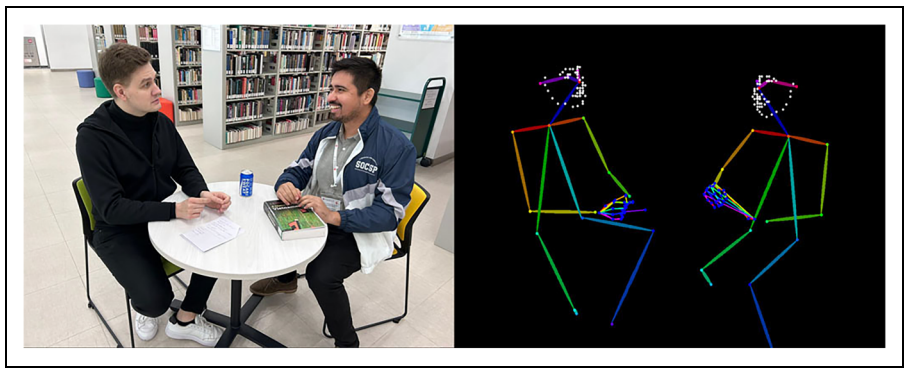

Figure 6. Second example case for image anonymisation openpose preprocessor skeleton frame.

Figure 7. Image generation with the OpenPose model comparison.

to align the actors in the replica with the original image. Variation (3) uses image-to-image generation with a denoising strength of 0.51 plus the OpenPose model. Variation (3) closely follows the original composition whilst emulating the facial expressions of the original actors. This procedure helps to further align facial expressions and the environment with the original. Some details, such as the beverage, are ignored when only the text prompt and the preprocessor are used, but they are included when image-to-image generation is utilised.

Image-to-image generations continue to provide results for the creation of replicas, especially when layered with models such as OpenPose. However, this approach does not undermine the application potential of ControlNet. Figure 8 illustrates how OpenPose can be used to transfer compositional elements to new settings, allowing us to retain the choreography between the actors whilst completely replacing the latter or the environment. Figure 8 was generated based on the same prompt as Figure 7 but with the gender changed to female, the location changed to a restaurant, and the bookshelves removed. As stressed above, the exact application scenario depends on what aspects of an image should be reported.

The functionality of ControlNet is limited not only to human actors and their poses. For instance, fictionalisation may also be required to anonymise locations whilst keeping information concerning the perspective of the depicted scene. Figure 9 illustrates how the depth model uses depth maps to control the perspective and depth of the generated images. In addition to the introduced models, a variety of other ControlNet models can be used to control the output of image generation with Stable Diffusion (see Zhang, 2023). All these models are situational in their application and can be combined.

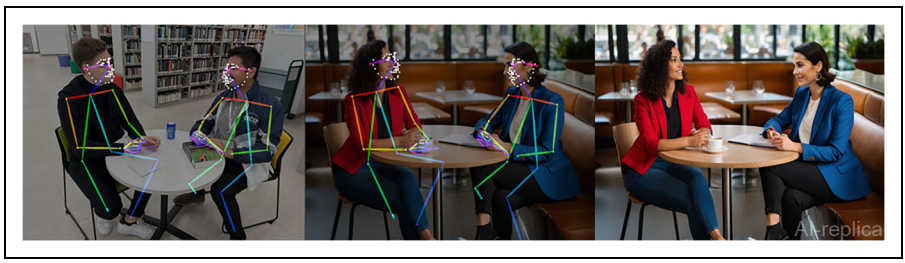

Figure 8. Illustration of pose transfer from original images to AI-replicas.

Figure 9. Illustration of depth ControlNet using a preprocessor depth mask.

Inpainting

Another key function for the creation of AI-replicas is inpainting. 'Inpainting' can be conceived as 'filling masked regions of an image with new content either because parts of the image are corrupted or to replace existing but undesired content within the image' (Rombach et al., 2022: 8). This method is essentially a localised image-to-image generation that allows the AI-replication of specific image sections. When applied effectively, this technology allows pivotal areas of an image to be rendered anew and merged seamlessly with the existing image. This technique can be applied to both AI-generated images and original pictures. The application potential of this function is threefold. Firstly, inpainting allows the occasional generation problems of the methods introduced above, such as two mobile phones or the colour of the jacket, to be fixed (Figure 3). Secondly, inpainting enables the restoration of previously anonymised elements (e.g. faces) and their replacement with AI-replicas. The latter is of critical importance when reporting visual data that have been stored in anonymised form only, whether for ethical or legal reasons (cf. Regulation (EU) 2016/679, 2016). Thirdly, inpainting allows the AI-replication of a particular information (e.g. faces) and thus serves as the most immediate alternative to traditional anonymisation methods (i.e. black bars, blur or pixelation) that retain the visual legibility of the original image. This application of AI-replicas can be especially useful if one wishes to prepare material to be rotated beyond the confidential relationship between the researcher and the data subject, for example, as stimulus material in group discussions.

Figure 10 illustrates the potential of the inpainting function to fill in previously anonymised pictures. On the basis of our previous example, we created a blurred version of Figure 2, which we subsequently processed with inpainting to create an AI-replica that emulates a face similar to the original image. This particular AI-replica was generated with a denoising strength of 0.51 and a CFG value of 7 (see Figure 21 in Rombach et al., 2022: 32 for further examples). This process is subject to the same limitations as full-image generation (i.e. parameters and the text prompt) and may require several iterative steps to achieve the desired output. Whilst showing the same anonymising potential as illustrated in Figure 5, this step also underlines the potential of AI-replicas to enhance anonymised material in a way that retains the legibility of non-anonymised material.

Discussion

The previous section illustrated the application potential of AI-replicas as an alternative to traditional anonymisation methods when reporting visual data. However, this possibility

Figure 10. Inpainting for restoration of blurred faces.

also invites a set of questions concerning the application potential, pitfalls and ethical concerns that we intend to address in this section.

AI-replicas and data richness

A major concern in image anonymisation and data fictionalisation is that removing information from an image must be carefully balanced against the ethical and research-practical necessities of a given usage scenario. The question concerning the data richness is pivotal for the previously addressed concern of validity and reliability of the presented interpretation (see the section 'Ethical and research-practical challenges in using images for qualitative research'). When applying traditional anonymisation methods, it is assumed that the application of black bars or pixelation of the face does not diminish the richness of the data and thus still allows the reader to comprehend the provided interpretation. However, this is relative to a given research objective. If the particular gaze and expressions on a given face are pivotal factors for a given interpretation, then obscuring them with traditional means like pixelation would render the interpretation impossible.

The examples presented in the section 'Practical application of stable diffusion for generating AI-replicas' show that guided image synthesis can help create AI-replicas that can retain expressive nuances whilst completely obscuring the identity of the data subjects whilst retaining the compositional complexity of the originals. However, providing a definitive answer to the question of whether AI-replicas distort or manipulate the data to a critical degree is impossible because this approach is relevant to a given research objective. The necessity of creating an AI-replica of a picture, whether for specific segments or the entire frame, depends on the sensitivity of the interview material and its relevance to the research question. Which visual information should be replicated, retained or altered remains relative to the findings that are supposed to be reported. As such, AI-replicas must enable the audience to follow the interpretation of the researcher without limiting the interpretative possibilities to the latter's position alone. Similar to text-based strategies that are reported instead of full interview transcripts, such as thematic summaries or composite accounts, AI-replicas should ideally be planned and provided with clear information concerning their creation (i.e. the parameters used), as well as a rationale that weighs the need for anonymisation. These rationales must remain as situational as the extent to which AI-replicas are applied. Ideally, AI-replicas are accompanied by descriptions of the original pictures that aim to describe the basic visual elements and their arrangement within a picture solely on the denotative level, providing readers with additional information about the original material (see 'pre-iconographic description' in Imdahl, 1996). These processes can be further controlled by being conducted by groups of researchers rather than individuals. As the generation of AI-replicas is an iterative process, researchers are well advised to keep and document any intermediate generations with remarks on why a generation had to be improved upon further. Such documentation can further support the reflexivity of the researcher when generating AI-replicas while also providing additional insight into the research process. While these intermediate AI-replicas might not be suitable for publication due to their incomplete nature (e.g. they do not sufficiently anonymise or abstract too much), they may still serve as valuable incubators for reflexivity on the researcher's end.

Concerning the quality of AI-replicas, it remains to consider at what point an AI-replica is too close to the original. As with all anonymisation methods, an

AI-replica would be too close to the original if it allows the identification of individuals by revealing biometric information, idiosyncratic behaviour such as particular clothing styles or postures, or contextual clues like the background or specific objects. This highlights a particular benefit of AI-replicas, as they enable the anonymisation of not only individuals but also entire spaces, such as private and vulnerable environments. An AI-replica can therefore be deemed too close to the original if it permits any of the scenarios listed above. We argue that 'as little as possible, as much as necessary' should be the guiding paradigm when using AI-replicas. However, the exact extent must remain relative to the image data in question and what the researcher wishes to report. For example, one could argue that the additional car and missing lamppost in Figure 5 (0.51) critically altered the essence of the original picture. Similarly, one could argue that these changes are too subtle to have any meaningful impact on any potential research questions and instead contribute to the fictionalisation effect of AI-replicas. If an extremely holistic application of the proposed method seems undesirable, then retreating to more local applications is always an option, as illustrated in Figure 10. On this level, AI-replicas generated with Stable Diffusion would fulfil the same function as contemporary automated full-body anonymisation methods (Hukkelås and Lindseth, 2022) but allow the user fundamentally greater control over the outcome. This approach utilises AI-replicas and is particularly desirable if we wish to retain selective information, such as facial expressions.

The results of qualitative data analysis, particularly interpretative approaches that go beyond simple codification of written text, as is commonly used in qualitative content analysis (cf. Graneheim and Lundman, 2004), are the outcome of a deep engagement of researchers with the data material. The richness of the data, or rather what is considered relevant for the research question at hand, plays a critical role in the identification and explication of meaning as represented by the data material. The situatedness of the researcher also plays a critical role in this process (Berger, 2015; Berger and Luckmann, 1967: 7). Therefore, AI-replicas will echo the selectivity and validation of the research-relevant information to a certain degree, similar to written protocol sentences. However, in contrast to protocol sentences, AI-replicas can align closely with the original material by using image-to-image generation with lower denoising strength. In this sense, AI-replicas are a form of display 'to provide evidence for claims in a format readers can easily access' (Watt, 2015).

Creating AI-replicas must not be 'something automated delegated to the machine' but rather a delicate process that is carefully administered by the researcher to overcome certain ethical and methodological issues that may arise with the use of this technology. As mentioned, AI-replicas do not replace qualitative analysis. Instead, they may be incorporated in the reporting stages of the research process. However, the representation must maintain a sufficient level of similarity to ensure that the audience can accurately follow and critically assess the interpretative process undertaken by the researcher. Therefore, similar to qualitative analysis itself (Elo and Kyngäs, 2008: 113), it is recommended that fellow researchers validate AI-replicas. Moreover, AI-replicas could obscure the voice of those who are researched. Replicated images created by research participants could interfere in methodologies that aim to give voice to marginalised groups, as the technology could produce results that reproduce those inequalities. Thus, research participants may also be part of that process of validation if necessary and informed consent may explicitly state the use of AI-replicas in the reporting process.

Bias handling and transparency of AI usage

A common concern when addressing any technology that falls under the broader label of generative AI is how biased the results are due to their training data. This concern for a biased generation must be considered separate from the situated interpretation by the researcher and reader. We have to stress that AI-replicas are not proposed to serve as the subjects of the researchers' interpretation, as these are still the original images. Instead, AI-replicas should assist the reader in comprehending the researchers' interpretation as if they are presented with the original images. Thus, it must be understood how much the used models, as well as the situatedness of the researchers in using this technology, may affect the outcome. As previously stressed, various sources have pointed out how susceptible technologies such as Stable Diffusion are to biased results (Heikkilä, 2023; Jenka, 2023; Nicoletti and Bass, 2023). Stable Diffusion uses contrastive language–image pretraining (CLIP) (OpenAI, 2020; Radford et al., 2021) to establish the connection between real language input (text prompts) and the image material it was trained on, effectively guiding the generation process. For example, CLIP connects the word 'house' with generations it learnt to represent 'houses'. The particular models of Stable Diffusion (e.g. SD 1.5 or SD-XL 1.0) guide the creative capacity of image generation. Metaphorically speaking, CLIP operates as a teacher who provides Stable Diffusion with the association of words and semantic concepts, whilst Stable Diffusion's creative capacity translates these text-concept combinations into a visual representation. Potentially biased generations are therefore found in the datasets on which a CLIP or Stable Diffusion model was trained. If a Stable Diffusion model is not trained sufficiently on a given object (e.g. physical features associated with Down syndrome), then the model will not know how to generate it (underfitting) or how to approximate it, in the best case, via association with other objects in the dataset. Similarly, as illustrated in Figure 3, Stable Diffusion tends to represent certain concepts in a given way (e.g. the yellow tree). This situation can be regarded as the result of the dataset and an association in and through the text prompt. For example, 'a tree during the autumn season' will most likely also lead to a yellow tree, despite not being prompted as such. Similarly, the prompt used for Figure 1 did not specify the skin colour of the two men or the bystanders. However, the model relies on data associating cockfight, clothing and climate with a skin colour that suits the cultural background of cockfights (see Geertz, 1973). The semantic association of concepts can therefore also be beneficial for the generation of AI-replicas. The concern described above is most evident in text-to-image generations that can suffer from insufficient prompts, as illustrated by the tree. Using image-to-image generation is therefore a potent strategy to counteract any bias in the data.

As such, the model used, and by proxy the model designer, will have an impact on the results when utilising text-to-image generations. It is inevitable that the model designer influences the generative capacity of the model through the selection of the training data as well as the design of the model itself. This impact should not be understood as deterministic, as the Stable Diffusion base models are designed with the intention to generate the largest possible spectrum of subjects, with the exception of violent and sexual content. The final generation is solely determined by the prompt and parameters given by the user within the generative boundary defined by the designer and model (see also Gandikota et al., 2023). These limitations, whether due to the intentional limitations of

the model designer or insufficient training data, are not final and can be further modified through model fine-tuning of the base models by more experienced users (Sayak and Park, 2023). As the previous example of the yellow tree exemplified, diffusion models gravitate towards certain standard representations that require sensitivity and validation on the part of the researchers. Given that these standard representations are difficult to systematically identify and due to the rapid development and subsequent replacement of base models, they are not economically feasible to track. Therefore, it is pivotal to reference the employed model and document the used parameters. Such documentation must be understood as an ethical and professional standard on the same level as properly referencing academic sources, as it will enable increased scrutiny, a more comprehensive understanding of the employed models, and a more systematic recognition of their strengths and weaknesses. This circumstance highlights the importance of understanding proper prompt engineering as a pivotal skill that can actively help counteract biases in the models used. Generating AI-replicas will require the researcher to cultivate a high conceptual clarity of what should be generated. For example, when generating 'a man next to a tree', we must actively decide what age or skin colour Stable Diffusion should render, but this must remain relative to the situated need of the AI-replica. Moreover, this level of detail must always be anticipated by the researchers during generation. Generating AI-replicas is therefore an active and reflective process and not necessarily a one-click automatisation. As such, the convenience of simple pixelations might be lost, but the approach could yield potentially more authentic representations. The aforementioned limitation of the models used will prevent Stable Diffusion from generating everything conceivable, as some classes of objects will inevitably be insufficiently represented. These limitations are usually addressed using fine-tuned models that specialise in particular outputs. The extent of this limitation, or rather the actual creative potential, requires further investigation. However, due to the rapid improvement of the models, this is difficult to realise. Furthermore, a more practical application of AI-replicas for research purposes is required to identify field- and discipline-specific needs to train custom checkpoints.

Besides the technical consideration in avoiding unnecessarily distorted generations, we also have to ensure that the use of AI-replicas and their extent are transparently communicated. An inherent benefit of traditional anonymisation methods is that the degree of their impact is often self-evident. When pixelating a face, researchers can easily judge whether this technique disables the audience from comprehending and potentially validating a presented interpretation. This aspect becomes more nuanced with the proposed AI-replicas, as the AI-replicas can ideally not be distinguished from the original ones in terms of perceived authenticity. The practical application of AI-replicas must be accompanied by a declaration of how and to what extent this technology has been used. The type of image generation employed (i.e. text-to-image, image-to-image or inpainting) and the type of ControlNet applied should therefore be described in a transparent manner. Reporting the guiding prompt used and the specific parameters for the generation (i.e. denoising strength and CFG scale) is also recommended. This approach will help the reader understand how to assess the degree and form to which an AI-replica is an alteration of the original. This practice can be complemented by other developing standards of AI usage in academic work, for example, AI usage cards (Wahle et al., 2023). A similar practice of standardised reporting of generation data is already in

place amongst online communities that use Stable Diffusion for artistic purposes and must be only slightly streamlined for academic purposes and variations such as image-to-image and inpainting. As Stable Diffusion automatically embeds the parameters used for generation as metadata into the generated files, we also propose making generated AI-replicas publicly available, ideally via digital repositories that can accompany any given publication. While it is not possible to reconstruct the original image based on these parameters, they provide a clear idea of the generation process while also serving as an identifier of the image as AI-generated. This embedded meta-information is particularly useful if AI-replicas begin to circulate in the research domain and become disconnected from their original publication. Beyond digital watermarks and embedded metadata, we also propose the adoption of a non-intrusive visual signifier that can identify an AI-replica and its purpose as such (see also Madiega and European Parliamentary Research Service, 2023).

As previously remarked in the section 'Image-to-image generation', AI-replicas generated using the image-to-image method do not allow the reconstruction of the original images due to the nature of the underlying diffusion model. However, researchers still remain accountable for any alterations concerning the veracity of the AI-replicas with respect to the reported information. While literal tracing of the images is technically not feasible and practically ill-advised, we advocate that image-to-image conversions should be documented and approved by the responsible ethical review board, similar to paraphrased, non-verbatim transcriptions or otherwise altered interview data. This includes documentation and rationale of deliberate changes, as well as retaining the original images for potential review by authorised parties.

Conclusion

Offering anonymity to research participants is an incontestable principle in image-based research (Nutbrown, 2011). Anonymity is a recurrent requirement of ethical review boards and outlets (Allen, 2015). Nevertheless, anonymisation poses several potential challenges to the research process: silencing the voice of research participants, shaping the research interests of researchers, reducing agency, increasing paternalism and limiting the possibility of reporting evidence that affects the validity of the findings. Moreover, current image modification and fictionalisation techniques (Jordan, 2014; Nutbrown, 2011; Wiles et al., 2012) do not fully account for or overcome these limitations.

This article introduces AI-replicas as a novel approach for ethically reporting visual data, adeptly balancing participant anonymity and data integrity. We have demonstrated that the AI-replica has the potential to effectively preserve emotional content and image authenticity, overcoming the limitations inherent in traditional anonymisation. This article proposes the use of AI-generated replicas to report visual data as complementary material to otherwise purely textual protocol sentences (Bohnsack, 2008: 3; Popper, 1959: 59ff.) or as a novel reverse photovoice procedure where the voice of participants can be utilised as an input to generate visual representations.

Certain concerns about the applicability of AI in qualitative research are legitimate. Firstly, qualitative researchers could be concerned about the possible distortions of images that could affect studies' conclusions. Nevertheless, our proposal only entails the use of AI-replicas for reporting purposes. Secondly, qualitative researchers could

be concerned about well-known biases of AI (Heikkilä, 2023; Nicoletti and Bass, 2023). Nevertheless, traditional anonymisation methods can also be subject to bias. Moreover, text prompts provide flexibility to adjust the output image to counteract biased generations. In this sense, the reflexibility of researchers and critical evaluation of outputs are key elements for identifying biases and correcting them. Furthermore, if possible, research participants can be involved in the generation of AI-replicas to give them creative authority over the end product. Thirdly, qualitative research could involve data documentation and transparency. This situation is another limitation of traditional fictionalisation methods. It is therefore pivotal to adhere to good practices of generative AI usage and document its use as diligently as possible (i.e. prompts, parameters and ControlNet usage).

The main contribution of this article is the introduction of AI-replicas as a technique to overcome limitations of traditional anonymisation methods in visual research. However, AI applications can go beyond the generation of replicas for this purpose and open opportunities for innovation in visual research more broadly. AI-replicas can be used for reporting aggregated data that represent and accentuate specific aspects of the analysed image data but without referring to any image in particular (e.g. for the development of typologies; visualisation of prototypes) or supporting data that are not visual in nature (e.g. field notes). For instance, social psychologists and anthropologists have long used the idea of prototypes in reference to the collection of quintessential elements of a concept (see MacLaury, 1991; cf. 'ideal type' in Schütz, 1972; Weber, 1949). A concept has features with varying levels of centrality that define a prototype, and people use them to judge what an object is and how it can be categorised. Here, more abstracted AI-replicas can find an application to convey the overall features of a prototype but without the need to adhere to the iconic singularity of any analysed image.

When used complementarily to field notes or mind protocols, AI-replicas can fulfil a supplementary function. Similar to interpretations or interviews regarding this matter, field notes are reductionist in nature and bound the situated perception of the researcher (Emerson et al., 1995; Mulhall, 2003). Here, the AI-replicas function complementarily to field notes, similar to visual sketches (Phillippi and Lauderdale, 2018), and can provide a pivotal visual dimension where visual documentation would otherwise (for which reason whatsoever) not be possible. This process requires the same critical reflection concerning the results of the image generations as would be for the creation of field notes (Watt, 2015) and qualitative research in general (Berger, 2015). This call for reflexivity applies not only to the generation of completely synthetic AI-replicas but also to the types discussed above. All these applications are illustrations of how AI can be used to enhance the depth and scope of qualitative methodologies. By incorporating AI tools, further studies in qualitative research can explore new dimensions of data (re) presentation and analysis.

The use of AI-replicas necessitates cautious consideration of AI biases and broader applicability in various research scenarios. Future work should aim to further evaluate the application potential and quality of this method concerning the limitations of the employed models to improve the utilisation of AI-replicas in various application scenarios. Furthermore, future studies can investigate the application potential of AI-replicas for the fictionalisation of video material, which can be made publicly available complementarily to the respective main publication. As we recognise the current early stages of this methodology, a fundamental consideration is the ongoing

development of the underlying technology, datasets and models, which implies continuous refinements and advancements. This article marks an initial step in leveraging generative AI's potential for anonymisation and fictionalisation in visual research. We expect to contribute to well-position qualitative research in the evolving AI landscape.

Acknowledgements

All presented AI-replicas (Figures 1, 3, 4, 5, 7–10) have been generated using the SD-XL 1.0-base model (CreativeML Open RAIL++-M License, Stability-AI, 2023b). We thank the anonymous reviewers, the editor, and especially Ronyu Li, Alexis Sossa, and Denise Klinge for their valuable feedback on earlier drafts of this paper.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Tobias Kamelski https://orcid.org/0000-0002-4537-5709

Notes

1. DALL•E and Midjourney allow model configuration but restrict the choice to proprietary, in-house trained models (Midjourney, 2024; Niles and OpenAI, 2024).
2. The Stable Diffusion community provides a vast set of comprehensive and user-friendly guides for the installation and use of Stable Diffusion and its modules:
   - Stable Diffusion installation guide: https://stable-diffusion-art.com/install-windows (Andrew, 2023b).
   - ControlNet installation and application guide: https://stable-diffusion-art.com/controlnet (Andrew, 2023a).
3. We intentionally ignore the possibility of using a consistent seed variable throughout multiple generations.
4. Suitable image-to-video models are already in active development (Stability-AI, 2023a, 2023d).

References

Allen L (2015) Losing face? Photo-anonymisation and visual research integrity. *Visual Studies* 30(3): 295–308.

Andrew (2023a) ControlNet v1.1: A complete guide. Available at: https://stable-diffusion-art.com/controlnet/ (accessed 11 January 2024).

Andrew (2023b) How to install stable diffusion on windows (AUTOMATIC1111). Available at: https://stable-diffusion-art.com/install-windows/ (accessed 11 January 2024).

AUTOMATIC1111 (2023) Stable diffusion web UI. Available at: https://github.com/AUTOMATIC1111/stable-diffusion-webui (accessed 16 December 2023).

Banks M (2001) *Visual Methods in Social Research*. London: Sage Publications.

Beaumont R (2022) Laion-5B: A new era of open large-scale multi-modal datasets. Available at: https://laion.ai/blog/laion-5b/ (accessed 14 January 2024).

Berger PL and Luckmann T (1967) *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Anchor Books.

Berger R (2015) Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research* 15(2): 219–234.

Betker J, Brooks T, Wang J, et al. (2023) DALLE3: Improving image generation with better captions. arXiv. https://cdn.openai.com/papers/dall-e-3.pdf

Bie F, Yang Y, Zhou Z, et al. (2023) RenAIssance: A survey into AI text-to-image generation in the era of large model. pp. 1–25. https://arxiv.org/pdf/2309.00810.pdf

Bohnsack R (2008) The interpretation of pictures and the documentary method. *Forum: Qualitative Social Research* 9(3): 1–17.

Burgos-Thorsen S and Munk A (2023) Opening alternative data imaginaries in urban studies: Unfolding COVID place attachments through Instagram photos and computational visual methods. *Cities* 141: 104470.

Common Crawl (2023) Common crawl. Available at: https://commoncrawl.org/ (accessed 14 January 2024).

CompVis (2022) Stable diffusion v1 model card. Available at: https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md#training (accessed 14 January 2024).

Douglas K (2022) Anonymizing people in images using generative adversarial networks. MSc Thesis, University of Amsterdam.

Elo S and Kyngäs H (2008) The qualitative content analysis process. *Journal of Advanced Nursing* 62(1): 107–115.

Emerson R, Fretz R and Shaw L (1995) *Writing Ethnographic Fieldnotes*. Chicago: The University of Chicago Press.

Gan Y, Greiffenhagen C and Licoppe C (2020) Orchestrated openings in video calls: Getting young left-behind children to greet their migrant parents. *Journal of Pragmatics* 170: 364–380.

Gandikota R, Materzyńska J, Fiotto-Kaufman J, et al. (2023) Erasing concepts from diffusion models. In: Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 1–6 October 2023, pp. 2426–2436. https://doi.org/10.1109/ICCV51070.2023.00230

Geertz C (1973) Deep play: Notes on the balinese cockfight. *Daedalus* 101(1): 1–37.

Graneheim UH and Lundman B (2004) Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today* 24(2): 105–112.

Hannes K and Parylo O (2014) Let's play it safe: Ethical considerations from participants in a photovoice research project. *International Journal of Qualitative Methods* 13(1): 255–274.

Heikkilä M (2023) These new tools let you see for yourself how biased AI image models are. Available at: https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/ (accessed 13 January 2024).

Hukkelås H and Lindseth F (2022) DeepPrivacy2: Towards realistic full-body anonymization. Available at: https://arxiv.org/pdf/2211.09454v1.pdf (accessed 14 January 2024).

Imdahl M (1996) *Giotto, Arenafresken: Ikonographie – Ikonologie – Ikonik*. Munich: Wilhelm Fink Verlag.

Jenka (2023) AI and the American Smile: How AI misrepresents culture through a facial expression. Available at: https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf (accessed 7 July 2023).

Kozinets RV (2020) *Netnography: The Essential Guide to Qualitative Social Media Research* (3rd ed.). London: Sage Publications.

Laurier E (2014) The graphic transcript: Poaching comic book grammar for inscribing the visual, spatial and temporal aspects of action. *Geography Compass* 8(4): 235–248.

Luhmann N (1995) *Social System*. Stanford, CA: Stanford University Press.

MacEntee K, Kendrick C and Flicker S (2022) Quilted cellphilm method: A participatory visual health research method for working with marginalised and stigmatised communities. *Global Public Health* 17(7): 1420–1432.

MacLaury RE (1991) Prototypes revisited. *Annual Review of Anthropology* 20(1): 55–74.

Madiega T and European Parliamentary Research Service (2023). *Generative AI and Watermarking*. Brussels: European Parliament. Available at: https://www.europeansources.info/record/generative-ai-and-watermarking/

Mao J, Wang X and Aizawa K (2023) Guided image synthesis via initial image editing in diffusion model. In: MM 2023 – Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023, pp. 5321–5329.

Markham A (2012) Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information Communication and Society* 15(3): 334–353.

Maximov M, Elezi I and Leal-Taixe L (2020) CIAGAN: Conditional identity anonymization generative adversarial networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020, pp. 5446–5455. Available at: https://ieeexplore.ieee.org/document/9157102/

Meng C, He Y, Song Y, et al. (2022) SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR 2022 – 10th International Conference on Learning Representations, Virtual, 25–29 Apr 2022. https://arxiv.org/abs/2108.01073

Midjourney I (2024) Version. Available at: docs.midjourney.com (accessed 13 January 2024).

Miller KE (2015) Dear critics: Addressing concerns and justifying the benefits of photography as a research method. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 16(3): Art. 27.

Mishkin P, Ahmad L, Brundage M, et al. (2022) DALL·E 2 preview – Risks and limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md

Moore N (2012) The politics and ethics of naming: Questioning anonymisation in (archival) research. *International Journal of Social Research Methodology* 15(4): 331–340.

Mulhall A (2003) In the field: Notes on observation in qualitative research. *Journal of Advanced Nursing* 41(3): 306–313.

Nicoletti L and Bass D (2023) Humans are biased. Generative AI is even worse. Available at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/ (accessed 30 January 2024).

Nutbrown C (2011) Naked by the pool? Blurring the image? Ethical issues in the portrayal of young children in arts-based educational research. *Qualitative Inquiry* 17(1): 3–14.

OpenAI (2020) CLIP. Available at: https://github.com/OpenAI/CLIP

Niles R and OpenAI (2024) DALL-E 3 API. Available at: https://help.openai.com/en/articles/8555480-dall-e-3-api (accessed 13 January 2024).

Papademas D and IVSA (2009) IVSA code of research ethics and guidelines. *Visual Studies* 24(3): 250–257.

Phillippi J and Lauderdale J (2018) A guide to field notes for qualitative research: Context and conversation. *Qualitative Health Research* 28(3): 381–388.

Podell D, English Z, Lacey K, et al. (2023) SDXL: Improving latent diffusion models for high-resolution image synthesis. https://arxiv.org/abs/2307.01952

Polat Z (2022) Facilitating a 'virtual space' for social change during the COVID-19 pandemic: Working with high-risk population using an arts-informed method. *Visual Studies* 37(1–2): 22–32.

Popper K (1959) *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Prosser J (2000) The moral maze of image ethics. In: Simons H and Usher R (eds) *Situated Ethics in Educational Research*. London: Routledge, 116–132.

Radford A, Kim JW, Hallacy C, et al. (2021) Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research* 139: 8748–8763.

Regulation (EU) 2016/679 (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&qid=1706168522662

Rombach R, Blattmann A, Lorenz D, et al. (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022, pp. 10674–10685.

Rose G (2022) *Visual Methodologies: An Introduction to Researching with Visual Materials* (5th ed.). London: Sage Publications.

Sayak P and Park C (2023) *Fine-Tuning Stable Diffusion*. Available at: https://keras.io/examples/generative/finetune_stable_diffusion/ (accessed 28 July 2024).

Schütz A (1972) *The Phenomenology of the Social World*. London: Heinemann Educational Books.

Smith M, Mann M and Urbas G (2018) Facial recognition. In: Urbas G, Mann M and Smith M (eds) *Biometrics, Crime and Security*. Oxon: Routledge, pp. 54–70.

Smith M and Miller S (2022) The ethical application of biometric facial recognition technology. *AI and Society* 37(1): 167–175.

Spencer D (2020) The face in visual representations of children. *Qualitative Research* 22(2): 236–250.

Spencer S (2023) *Visual Research Methods in the Social Sciences: Awakening Visions* (2nd ed.). Oxon: Routledge.

Stability-AI (2023a) SDXL-turbo model card. Available at: https://huggingface.co/stabilityai/sdxl-turbo (accessed 13 January 2024).

Stability-AI (2023b) stable-diffusion-xl-base-1.0. Available at: https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0 (accessed: 11 January 2024).

Stability-AI (2023c) Stable diffusion version 2. Available at: https://github.com/Stability-AI/stablediffusion (accessed 16 December 2023).

Stability-AI (2023d) Stable video diffusion image-to-video model card. Available at: https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt (accessed 13 January 2024).

Tiidenberg K and Baym NK (2017) Learn it, buy it, work it: Intensive pregnancy on Instagram. *Social Media and Society* 3(1): 1–13.

Wahle JP, Ruas T, Mohammad SM, et al. (2023) AI Usage Cards: Responsibly Reporting AI-Generated Content. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2023, Santa Fe, NM, USA, 26–30 June 2023. Association for Computing Machinery.

Wang C and Burris MA (1997) Photovoice: Concept, methodology, and use for participatory needs assessment. *Health Education & Behavior* 24(3): 369–387.

Wang P, Li Y, Singh KK, et al. (2021) IMAGINE: Image synthesis by image-guided model inversion. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021, pp. 3680–3689.

Ward J (2017) What are you doing on Tinder? Impression management on a matchmaking mobile app. *Information Communication and Society* 20(11): 1644–1659.

Watt D (2015) On becoming a qualitative researcher: The value of reflexivity. *The Qualitative Report* 12(1): 82–101.

Weber M (1949) *The Methodology of the Social Sciences*. New York: Free Press.

Weng L (2021) What are diffusion models? Available at: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ (accessed 30 January 2024).

Wiles R, Coffey A, Robison J, et al. (2012) Ethical regulation and visual methods: Making visual research impossible or developing good practice?. *Sociological Research Online* 17(1): 3–12.

Yang L, Zhang Z, Song Y, et al. (2024) Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* 56(4): 1–39.

Yang Y, Lyu J, Wang R, et al. (2022) A digital mask to safeguard patient privacy. *Nature Medicine* 28(9): 1883–1892.

Zhang L (2023) Released checkpoints. Available at: https://huggingface.co/lllyasviel/sd-controlnet-openpose#released-checkpoints (accessed 27 January 2024).

Zhang L, Rao A and Agrawala M (2023) Adding conditional control to text-to-image diffusion models. https://arxiv.org/abs/2302.05543

Author biographies

Tobias Kamelski is a cultural sociologist specialising in reconstructive and praxeological approaches as well as computational methods. His research explores self-presentation and impression management within online dating, focusing on variations across gender, sexual orientation and culture. Employing reconstructive, ethnomethodological and computational methods, he investigates how knowledge structures influence everyday actions in digital spheres and the power dynamics embedded within these actions. Grounded in critical theory, his research aims to enhance media literacy and promote responsible use of technology.

Francisco Olivos earned his PhD in Sociology from The Chinese University of Hong Kong. He also holds a master's degree in Sociology and Social Research from Utrecht University and a master's degree in Sociology from Pontificia Universidad Catolica de Chile. His research interests lie in the intersections between social stratification, cultural sociology, subjective well-being and quantitative methods. His work has been published in the British Journal of Sociology, Research in Social Stratification and Mobility, and European Societies, among others.