Original Manuscript

Using Artificial Intelligence to Generate Visual Vignettes in Factorial Survey Experiments

Social Science Computer Review
2025, Vol. 0(0) 1–25

DOI: 10.1177/08944393251392916
journals.sagepub.com/home/ssc

Nicole Schwitter[1,2]

Abstract

Factorial survey experiments are widely used in the social sciences to study decision-making and attitudes through controlled, experimentally manipulated scenarios – typically presented as text. While textual vignettes offer flexibility and ease of use, they often lack realism and may limit participant engagement. This article explores how generative artificial intelligence (AI) can be used to create customisable images for visual vignettes. It demonstrates techniques for producing and selectively editing images, highlighting their potential to address the demands of experimental social science research, while it also acknowledges key challenges, including ethical considerations, biases inherent in AI tools, and technical limitations. The article showcases potential applications of AI-generated images in social science research and draws on a pretest with human participants to present evidence on how AI-generated images are perceived and interpreted. By critically evaluating both opportunities and challenges, this article provides researchers with practical guidance on integrating AI-generated visuals into factorial survey experiments, enhancing methodological approaches in the social sciences.

Keywords

vignette experiment, factorial survey experiment, visual stimuli, generative artificial intelligence, Midjourney

Introduction

Humans are inherently visual creatures: Research suggests that images are processed faster, remembered better, and recalled more easily than verbal or textual information (Bower, 1970; Potter et al., 2014; Shepard, 1967; for a critical discussion on processing speed, see Tuscher, 2022). They provide quicker, more direct access to meaning (Schlochtermeier et al., 2013) and engage evolutionarily older brain regions than those processing verbal input (Harper, 2002). Despite the importance of visual information in everyday perception, social science data remains

[1]MZES, University of Mannheim, Mannheim, Germany
[2]Department of Sociology, University of Warwick, Coventry, UK

Corresponding Author:
Nicole Schwitter, University of Mannheim, Postfach, Mannheim 68131, Germany.
Email: nicole.schwitter@uni-mannheim.de

largely textual. Especially in surveys – one cornerstone of quantitative social science research (Sturgis & Luff, 2021) – text-based, written questions dominate the data collection process. This also applies to multifactorial survey experiments, where participants evaluate carefully designed, experimentally varied scenarios. In this article, I introduce AI-generated visuals to expand the range of how to present these scenarios, demonstrating how to best use AI-generated and AI-manipulated images in experimental social science research.

Factorial survey experiments (also known as vignette experiments) are the most widely used multifactorial survey method in the social sciences (Treischl & Wolbring, 2022). Within these, survey participants are presented with short descriptions of scenarios and respond accordingly. For example, in the study of Baron et al. (2001), participants rated conflict situations which were described in vignettes as the following one: '*Suppose you are passing by a local convenience store on an afternoon walk. The street is practically empty. Suddenly, a person steps out and pushes at [sic] you. This person is a wealthy looking 18 year old man*'. Baron et al. (2001) varied the offender's age, gender, and social class, the victim's identity, conflict intensity, and bystander presence to assess their impact on participants' aggressiveness and disputatiousness.

One key advantage of vignette experiments is their ability to present information in a way that facilitates imagination and mirrors real-life situations (Eifler & Petzold, 2022; Finch, 1987). Most factorial survey experiments use textual vignettes (Treischl & Wolbring, 2022) which are easy to implement, flexible, and allow for precise variation by simply modifying a few words. For example, by replacing just a few words, a 'wealthy looking 18 year old man' can become a 'lower class looking elderly woman' and thus change age, gender, and perceived social status.

However, textual vignettes may lack external validity and realism, as they exclude visual information, which plays a critical role in how people perceive the real life (Bansak et al., 2021; Harper, 2002). It is often unclear what mental images respondents form when reading descriptions, and cues like names used to signal gender or ethnicity may introduce unintended associations with age or class (Gaddis, 2017; Sidler et al., 2024).

A key challenge in using visual stimuli is the difficulty of varying specific, visible attributes while keeping all others constant. For instance, creating images of a wealthy looking 18-year-old man and a lower class looking elderly woman requires recruiting models and taking photos where they make the same facial expression, exhibit the same posture and wear suitable clothes which are similar except in signalling a specific social status. Alternatively or additionally, graphic designers may be needed to edit images. If images are supposed to include larger contexts or more background, these can further be affected by external, uncontrollable aspects (for example, a busy street scene may unintentionally include different passing cars). While stimulus sampling can partly address this problem, it requires many different vignettes which can be difficult to create with limited resources (Fong & Grimmer, 2023; Monin & Oppenheimer, 2014).

The creation of experimental visual stimuli can thus be costly and time-consuming, especially at scale. There are also inherent limitations to what can realistically be staged: Some scenarios, like a tense police encounter in a specific urban setting or a refugee family crossing a border, may be ethically sensitive or difficult to recreate convincingly. Additionally, staging rare or impossible events (e.g. alien encounters, historical scenes, or future technologies) and scenes involving a large number of actors are often out of reach both practically and imaginatively. These challenges limit the scope of what visual experiments can explore using traditional image creation methods. Recent advances in generative AI, especially transformer-based models, offer promising solutions (see also Davidson, 2024).

In this article, I pursue two goals. First, I provide a practical guide for researchers on how to create and refine AI-generated and AI-manipulated visual vignettes for use in factorial survey experiments. Second, I empirically assess how participants perceive such AI-generated images in terms of realism and authenticity, based on a pretest study. The empirical case focuses on the

critical step of validating images as credible experimental stimuli. To achieve these goals, the remainder of the paper is structured as follows: The next section reviews factorial survey experiments, their methodological foundations, and the use of visual vignettes. After that, I outline practical steps for creating visual stimuli – model selection, image generation, and editing – and highlight potential applications. Next, I present an empirical case, and then discuss limitations of AI-generated visuals as well as offer practical guidance. The paper concludes with key insights and directions for future research.

Factorial Survey Experiments and Visual Vignettes

Factorial survey experiments have gained tremendous popularity over the last two decades and have been extensively used across different research topics in the social sciences (see Treischl & Wolbring, 2022; Wallander, 2009). Factorial survey experiments integrate survey and experimental design. The strength of vignette experiments lies in their ability to simulate realistic yet controlled environments that allow researchers to isolate and examine specific variables while surveying large samples.

The basic idea is to present respondents with a series of hypothetical descriptions of persons, objects, or situations (so-called vignettes). These vignettes consist of and experimentally vary along predefined characteristics or attributes (dimensions/factors). One or more vignettes – most often selected randomly – are then shown to participants for evaluation (Auspurg & Hinz, 2015). Vignettes are commonly presented as text or tables, with photos or videos used less frequently (Treischl & Wolbring, 2022) and primarily within certain fields (Hu et al., 2022) – particularly consumer and user behaviour studies (e.g. Baptista et al., 2023; Loosschilder et al., 1995), urban planning and landscaping (e.g. Shr et al., 2019), health-related research (e.g. Jiwa et al., 2014) – and particularly within choice experiments (Shr et al., 2019).[1]

Survey experiments like vignette experiments offer a time and cost-effective way to generate data, especially with text vignettes; however, they come with challenges and work best under certain methodological conditions (Auspurg et al., 2009; Barabas & Jerit, 2010). For one, vignette descriptions can be relatively artificial: The controlled nature of factorial designs may reduce external validity, as participants may perceive the scenarios as unrealistic or detached from real-world experiences – a challenge which may be particularly pronounced for textual representations, since respondents must imagine entire situations from brief descriptions (Barabas & Jerit, 2010; van Zelderen et al., 2024; Wallander, 2009). Additionally, textual vignettes require participants to mentally construct the described scenarios, which cannot only vary widely in accuracy or detail based on personal experience, cultural context, or cognitive styles, but also be biased in non-random ways (see on names e.g. Gaddis, 2017; Jürges & Winter, 2013). Complex or long vignettes can further overburden participants with a high cognitive load (Shamon et al., 2022; Teti et al., 2016). Textual vignettes may also fail to fully engage participants or to evoke strong emotional or behavioural responses (Schlochtermeier et al., 2013), and research suggests that participants of vignettes studies prefer more immersive vignette formats (van Zelderen et al., 2024).

Visual vignettes can overcome many of these limitations of text and they can also open up new theoretical and methodological possibilities. Studying visual vignettes allows researchers to capture how meaning is constructed in more naturalistic and culturally relevant ways, given that humans routinely encounter and interpret images in everyday life, from news photographs and campaign posters to memes and social media feeds. They also allow researchers to study spontaneous judgements, as images are processed very quickly (Potter et al., 2014). Moreover, images convey multiple layers of information simultaneously, such as expressions, gestures, environments, and social signals, allowing the analysis of interactions between subtle cues that are

more difficult to represent in text. Images can thus tap into implicit biases and trigger heuristics in ways that written descriptions cannot.

Against this background, researchers have increasingly turned to visual stimuli across diverse applications. For instance, Krysan et al. (2009) used videos of neighbourhoods varying in social class and racial composition to examine racial attitudes. Similarly, Havekes et al. (2013) combined text and photos in vignettes to explore mechanisms behind interethnic attitudes. Golden et al. (2001) used photo vignettes depicting ambiguous sexual harassment scenarios, manipulating the attractiveness of the individuals involved to study its influence on perceptions. Wouters and Walgrave (2017) presented edited news segments on protests, altering cues such as demonstrators' unity and commitment to assess their effect on audience perception. In the study by Thébaud et al. (2021), respondents evaluated photos of rooms labelled as male- or female-occupied to investigate gendered norms around housework. Krakowski et al. (2024) exposed participants to images – AI-generated or database-sourced – of fictitious individuals to test how native respondents sanction perceived norm violations by immigrants. López Ortega and Radojevic (2024) used a visual conjoint design to examine voter preferences, showing that AI-generated portraits had a stronger influence than textual labels, moderated by the visibility of social categories. Relatedly, Vecchiato and Munger (2024) developed a visual conjoint using fictional Twitter profiles of political candidates, demonstrating how visual cues can enhance realism in experimental designs. Political science research has also leveraged manipulated photos to study the effects of skin tone on perceptions of political representatives (Caruso et al., 2009; Mugglin et al., 2025). These examples, though selective, highlight the expanding role of visual elements in enriching experimental methods.

Visual vignettes, like their textual counterparts, feature different dimensions and levels. Ideally, they feature a higher degree of realism as the visual representation mirrors real-world appearances, environments, or interactions more closely (Caro et al., 2012; Manghani, 2012). Unlike written vignettes, which present information sequentially, images convey all information simultaneously. They can enhance participant engagement, evoke stronger cognitive responses, and create a sense of experiencing the scenario, requiring less extensive information processing (Burt et al., 2010; Eifler & Petzold, 2022). Thus, visual vignettes can function as a complementary method that aligns more closely with how people actually process information in real-world contexts.

However, creating visual vignettes is difficult due to the costs and logistics involved in producing high-quality visuals. This paper argues that recent advances in generative AI can help overcome many of the barriers, building upon a small but growing number of studies which have used AI-manipulated visuals (e.g. Eberl et al., 2022; Evsyukova et al., 2025; Haut et al., 2021; Kühn & Wolbring, 2024). It outlines a practical, scalable approach to make visual stimuli more accessible in social science research.

Generating and Manipulating Images for Factorial Survey Experiments: An Example Process

The first step in implementing AI-generated image vignettes is determining whether image vignettes are an appropriate tool for the research question at hand. Several studies have systematically compared the presentation format differences in vignette experiments (Eifler & Petzold, 2022; Facciani et al., 2022; Hu et al., 2022; Slattery Rashotte, 2003) as well as in choice experiments (DeLong et al., 2021; Kabaya et al., 2024; Shr et al., 2019). Overall, the body of research yields mixed results, suggesting no clear advantage of one format over another. Applications that involve relevant visual information are argued to be better operationalised with visual vignettes, as written vignettes may fail to adequately process information mean to trigger the sense of vision (Eifler & Petzold, 2022). For image vignettes to be useful and offer an

advantage, visible characteristics should be of theoretical and substantial interest. Not all concepts or situations can be effectively conveyed in still images and the potential level of detail is limited. The appropriateness of image vignettes also depends on the cognitive processes and interpretive frameworks of participants. Researchers must consider whether an image-based stimulus is likely to evoke the intended associations and reactions, and could draw on psychological theories of recognition, information processing, and memory (e.g. Paivio, 2013), as well as explicitly pretest their expectations. Visual stimuli offer advantages when visual information is theoretically relevant (Eifler & Petzold, 2022) and scope conditions are met (Vecchiato & Munger, 2024).

If the research context lends itself to visual vignettes, the images can be created and implemented in the following steps:

1. AI model selection
2. Generation of baseline image
3. Generation of image alternatives
4. Validation of images

Note that visual vignettes can take many forms such as photos, cartoons, animations, avatars, or other forms of drawings and graphic designs. In the following, I will focus on photorealistic images, which in the past typically required staging real situations. Other types of visuals, such as avatars or cartoons, could be equally generated using the same AI-based procedures.

AI Model Selection

The first step in generating or manipulating images is selecting an appropriate AI model. While the field evolves rapidly and no single recommendation can be given, the following guidelines should help contextualise the choice based on practical considerations. A key distinction in AI image generation lies between open-source and closed-source models. Closed-source models (like the popular Midjourney and OpenAI's DALL-E and Sora, which cannot be self-hosted) have, in the recent past, provided polished results that open-source models lagged behind due to the significant resources invested by the companies in fine-tuning their models. Proprietary models further tend to be designed with user-friendly interfaces and a powerful backend infrastructure. However, transparency is restricted, the image generating model cannot be directly modified, and as users, we rely on an external platform which can raise concerns about data security, cost sustainability, and ethical oversight.

In contrast, open-source models (like StabilityAI's StableDiffusion) allow users to self-host the model on their own computational infrastructure, giving them control over data, outputs, and computational environment. This flexibility allows for more control over privacy and reproducibility, as well as the ability to adapt the model for specific use cases like specifying human poses or copying compositions (e.g. using ControlNet[2]). However, open-source models require more technical expertise to set up, maintain, and optimise.

In this article, I used Midjourney which has (as of December 2024) generally been the preferred choice among photographers and designers because of its quality, consistency, and customisability, providing superior out-of-the-box performance. It was also one of the first (and is still one of the few) models to provide selective image editing options.[3] However, it is a subscription-based closed-source model. A more detailed discussion of closed-source and open-source software will be given in section "*Dependency on corporations: Closed-source and open-source software*".

Midjourney can be accessed via a web interface and Discord, an instant messaging social platform. Subscriptions are required for image generation, with the 'pro' plan (60$/month as of April 2025) offering private and unlimited image generation. The following sections will outline

image generation with Midjourney (version 6.1/January 2025), noting that some features may change as the software evolves.

Generation of Baseline Image (Using Midjourney)

The efficient creation of vignettes with Midjourney relies on generating a strong baseline image and subsequent selective editing of this baseline image (alternatively, existing images can be uploaded by the user and then selectively edited which allows to manipulate real scenes and existing imagery more generally).

Images are generated (and manipulated) by using *prompts*, text phrases provided by the user to guide the AI in generating an image. Prompts can include specific details such as objects, settings, styles, colours, or artistic influences, which the AI interprets to produce an image output. The Midjourney bot breaks down the prompt into smaller pieces (*tokens*) which are compared to its training data to then generate an image. The Midjourney bot generally works best with simple, short phrases that describe what should be shown in the image (Midjourney, 2025b). Words in the beginning of the prompt are suggested to have a greater effect on the resulting image. Users can refer to the Midjourney 'explore' page[4] for inspiration based on recent examples by other users.

When writing a prompt for a social science research application, it is generally important to be clear about the context of an image and details such as the subject of the image (person, location), the medium (e.g. a photo), the environment, and the composition (headshot, close-up, birds-eye view, etc.). 'Prompt engineering', the crafting of efficient commands, is an iterative process of trial and error, guided by best practices and complicated by the fact that generative AI models are sensitive to changes in phrasing and word choice and that these models are not necessarily deterministic, so that the same prompts can lead to different outputs and can also vary between versions (Jahani et al., 2025; Xie et al., 2023). Midjourney prompts can contain a seed parameter which allows to somewhat reproduce the initial image generation starting from a prompt. Experimenting with phrasing can help and is necessary to achieve the best results (see also Don-Yehiya et al., 2023).

After submitting a prompt, Midjourney returns four images, which can then be further edited. Figure 1 shows the output images for a study which might focus on family norms. Four images of a German-looking family were created after one request.

When crafting prompts, it is important to recognise that Midjourney has learnt what images look like from existing imagery and its labels. Machine learning algorithms tend to reproduce existing biases (Arseniev-Koehler & Foster, 2022; Bolukbasi et al., 2016) and this also holds for generative image AI. When generating scenes with people, less stereotypical prompts may result in images that do not adhere to the prompt in the intended way. For example, when trying to depict a person breaking into a house, a black man was more likely to appear as an intruder, while a white man was shown simply on the porch. This suggests that the AI associates certain racial profiles with specific actions (see Figure 2, panels A and B). Similarly, when aiming to create a scene where a young man is shown to break into a car, more prompt-adhering results were obtained when describing the man as 'Romanian' than, for example, 'Italian', making use of existing car thief stereotypes of Eastern Europe which the generative AI seems to have incorporated (see Figure 2, panels C and D).

More advanced prompts can also include image URLs and parameters to refine the generated image. Image URLs serve as an inspiration, influencing composition, style, and colour of the resulting image. Key parameters for a social scientific use case include the aspect ratio (the connection between the width and height of an image), image weight (adjusting the prominence or dominance of an image prompt), and the 'no' parameter, which refers to negative prompting. Negative prompting allows researchers to explicitly exclude certain elements from the generated

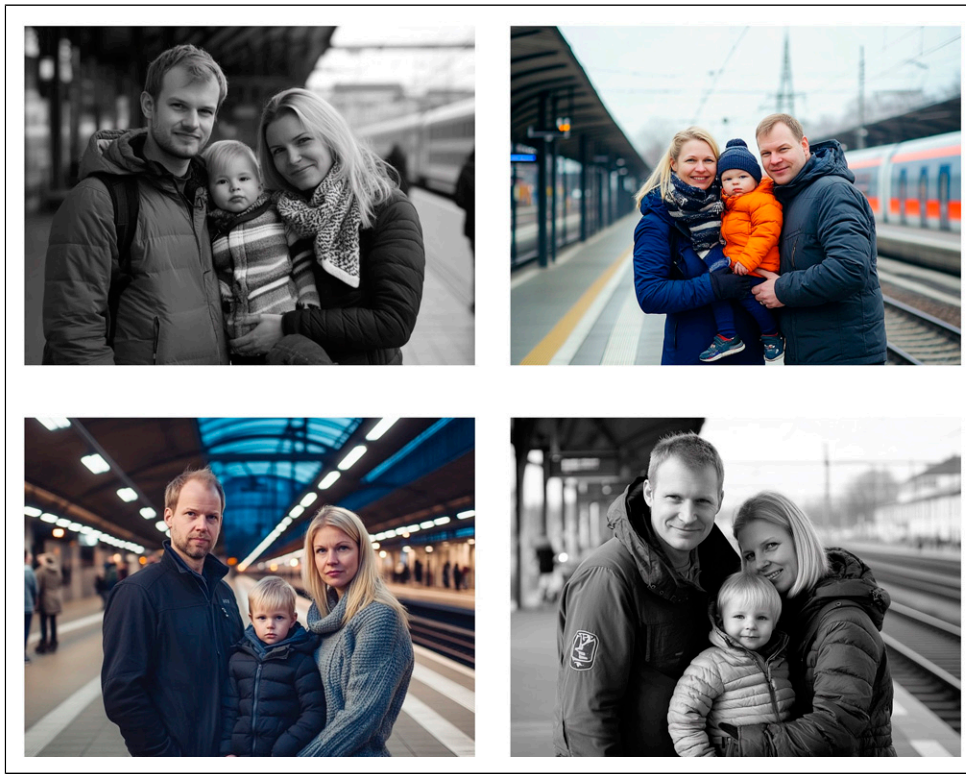

Figure 1. Four Images Returned by Midjourney Showing a Family. Prompt: 'german family standing at the train station, waiting. father, mother and one child. they have blond hair'.

image (e.g. by adding '--no text' to prevent letters or writing from appearing). The aspect ratio can play a surprisingly important role; it impacts the shape and the composition of a generated image (e.g. landscape, square, and portrait) and the same text prompts with different aspect ratios can lead to very different results. Again, this might be the case due to the fact that Midjourney has learned from existing images and different image sources are associated with specific ratios (e.g. Instagram photos are a source of square images). This also must be considered when researchers want to create different versions for desktop and mobile surveys.

Generated images can be further modified, either subtly or strongly – altering composition, colours, or fine details. Midjourney also supports regional editing, allowing users to select and regenerate specific areas based on a new prompt. The outcome depends on the original content, the selected region, and prompt; only the chosen area is altered. This functionality is particularly useful for refining image accuracy, removing distractions, or addressing small inconsistencies or AI artifacts such as poorly rendered hands. Panel A in Figure 3 shows an initial image of a person giving a speech, which was iteratively refined (Panel B) to produce a more polished baseline image which might then be edited further. This feature grants researchers significant control, enabling unlimited adjustments – unlike in collaborations with graphic designers, where extensive iteration is rarely feasible. Please note that these subsequent editing steps generally do not have a 'seed' and are less reproducible than the initial generation step. While the generated images are thus not completely reproducible, this also applies to photos created with other means, such as hired actors.

Generating an appealing and suitable baseline image is a challenging process which often requires many iterations of trial and error.

Figure 2. Images Showing People Conducting Illegal Activities, Created by Midjourney (Selection of One Representative Image Out of Four Returned by Midjourney). Prompts: [A + *B*] 'you see a house in the distance. there is a middle-aged [white/*black*] man who carefully climbs over the railing onto a porch from an upper-class house. The porch is wide and pretty. The camera view is wide-angle and shows the man from the side, gripping the upper edge of the railing with both hands. The setting is in a suburban, green area' [C + *D*] 'Wide angle Photo of a young [Italian/*Romanian*] man who is standing close to car and is preparing a break in. He peeks into the car. In the background is an empty street with parked cars'.

Generation of Image Alternatives

Once a satisfying baseline image has been created, the next step involves varying specific dimensions by editing selected areas to introduce different vignette attributes. Alternatively, users can also upload their own image materials, be it real photos taken by the researcher or photo material from other sources. Real images can thus also serve as a baseline which allows researchers to start with a realistic ground truth. In this next phase of generating image alternatives, researchers can manipulate certain features of the image to reflect the varying dimensions/attributes of the experimental stimuli.

Regional variation tools allow researchers to highlight and edit specific image areas. For instance, in the hypothetical family norms study (Figure 1), altering the faces of the depicted parents enables straight-forward adjustments to represent a mixed-race, Arab, or same-sex couple – simultaneously varying gender, race/ethnicity, and sexual orientation (see Figure 4). Subsequently, smaller refinements can be made to enhance facial consistency, such as adjusting

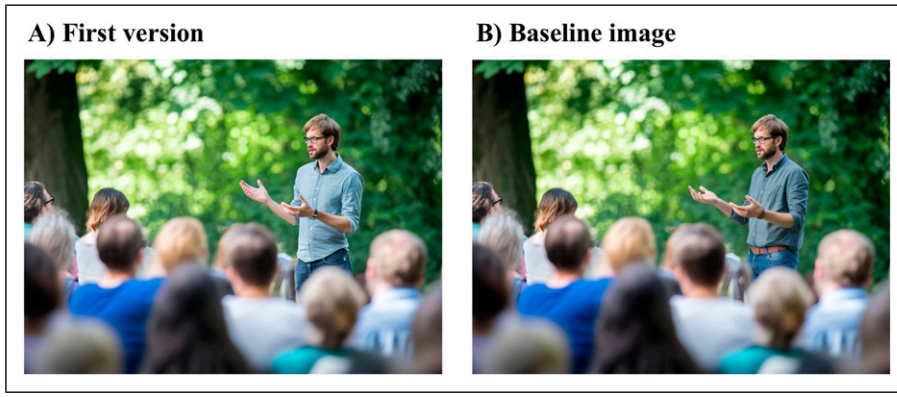

Figure 3. Images of a Speaker, Created by Midjourney. (A) Prompt: 'person giving a speech to an audience. it is set in the public in a park. focus is on the speaker who is a german man in casual clothes'. (B) Selectively Edited Image. Shirt and Pants Were Highlighted and Edited With the Prompt: 'man in casual clothes. wearing jeans'. Lower Stomach of Man Was Highlighted and Edited With the Prompt: 'Shirt tucked in, belt visible'. Hands Were Highlighted and Edited With the Prompt: 'normal hand with normal amount of fingers --no weird hands'.

facial expressions or features like facial hair. Finally, researchers can use editing software like Photoshop or even basic tools like Paint alongside AI-generated content to manually combine or modify image elements, enabling more precise control over multiple factors. Manual editing was done in one instance in this paper (Figure 5(G)): After selectively editing the face and body of the speaker using AI prompts, I manually combined the output with a consistent face generated in another image (using Paint).

The prompt to vary a specific region can also include image URLs. For example, by selecting the clothes of a person and passing an image of inspiration clothes, the style of a person can be changed in a specific way. Taking the baseline image from Figure 3 (panel B), we can generate further images to study the perception of public speakers (see Figure 5). While varying clothes and gender simultaneously makes the image generation a bit more complex (as simply changing faces does not produce believable female bodies) and requires the combined use of a generative AI model and a (simple) image editing tool (like Paint), it is still easily possible without advanced image editing skills. The more features overlap within an image, the more difficult the editing becomes; however, this might change with technological advancements.

While the previous examples focused on altering the appearance of individuals, researchers can also manipulate environments and contexts. Figure 6 presents two examples of contextual variation. Panels A and B show different levels of disorder in a person's neighbourhood, which could be used in studies examining the spread of disorder (Keizer et al., 2008). In these images, the more disorderly environment served as the baseline, as it is generally easier to remove objects in Midjourney than to add them (i.e. removing trash and graffiti from a 'poor neighbourhood' tends to work better than inserting trash and graffiti into a 'rich neighbourhood'; adding items like trash often results in unrealistic, oversized pieces of very specific debris). Panel C and D depict variations of village environments.

With well-crafted prompts and targeted edits, Midjourney enables the creation of realistic, context-specific images of people and settings. These can be further refined in external programs and are generally free to use for research under the Creative Commons Noncommercial 4.0 Attribution International License. As of November 2024, Midjourney also introduced an

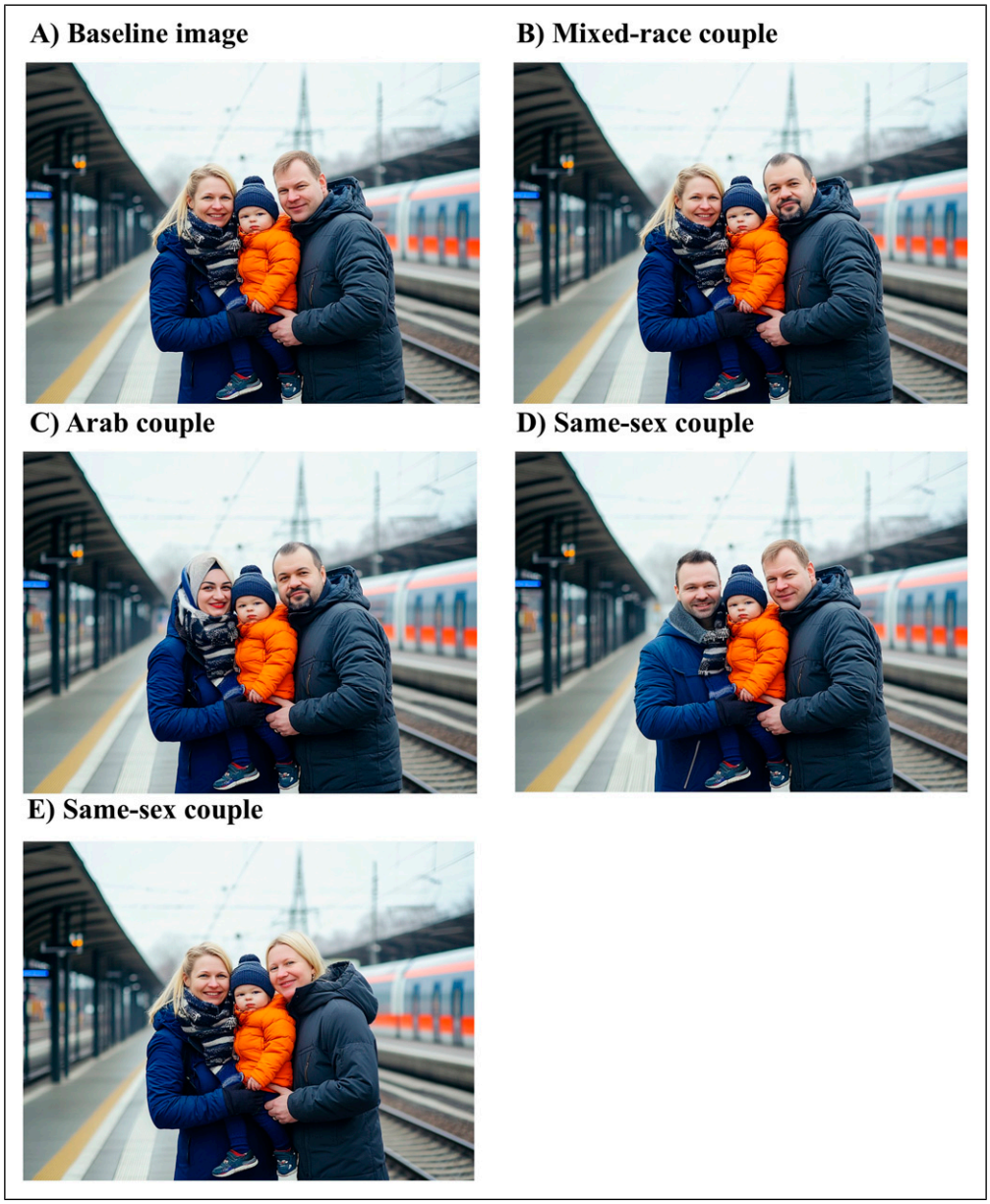

Figure 4. Images of a Family, Created by Midjourney. (A) Baseline Image, see Figure 1. (B) Selectively Edited Image. Face of the Person on the Right in Panel [A] was Highlighted and Edited With the Prompt: 'arab looking man'. (C) Selectively Edited Image. Face of the Person on the Left in Panel [B] was Highlighted and Edited With the Prompt: 'arab looking woman wearing a headscarf'. (D) Selectively Edited Image. Face of the Person on the Left in Panel [A] was Highlighted and Edited With the Prompt: 'gay german man'. (E) Selectively Edited Image. Face of the Person on the Right in Panel [A] was Highlighted and Edited With the Prompt: 'german woman. She is a lesbian'.

Figure 5. Images of a Speaker, Created by Midjourney. (A) Baseline Image, see Figure 3. (B) Selectively Edited Image. Clothes of the Speaker in Panel [A] Were Highlighted and Edited With the Prompt: 'person in black suit and tie'. (C) Selectively Edited Image. Clothes of the Speaker in Panel [A] Were Highlighted and Edited With the Prompt: 'homeless person with dirty and ripped clothes https://s.mj.run/ShhlXo6swOw'. (D) Selectively Edited Image. Clothes of the Speaker in Panel [A] Were Highlighted and Edited With the Prompt: 'person in german military uniform. With Germany flag on sleeve --iw 3 https://s.mj.run/oJdjylJ0oPM'. (E) Selectively Edited Image. Clothes of the Speaker in Panel [F] Were Highlighted and Edited With the Prompt: 'business casual attire. woman wearing jeans, blouse tucked into the jeans with a belt'. Baseline Image [A] was Also Passed to the Prompt. (F) Selectively Edited Image. Clothes of the Speaker in Panel [H] Were Highlighted and Edited With the Prompt: 'woman in formal business attire'. (G) Selectively Edited Image. Face of the Speaker in Panel [C] was Highlighted and Edited With the Prompt: 'woman with glasses'. Parts of Body Were Then Highlighted and Edited Using the Prompt 'female' to Change Bodily Features. The Then Created Image Was Manually Combined With the Face From Image [E] to Make it Consistent (Using Paint). (H) Selectively Edited Image. Face of the Speaker in Panel [D] was Highlighted and Edited With the Prompt: 'woman with black glasses'.

Figure 6. Images of Environments, Created by Midjourney. (A) Prompt: 'poor neighbourhood in a german suburb. there is graffiti on walls'. Selectively Edited Image by Highlighting Part of the Images and Using the Prompt: 'german teenager throwing a bottle on the ground'. (B) Selectively Edited Image. Walkway and Buildings in Panel [A] Were Highlighted and Edited With the Prompt: 'pretty walkway no graffiti'. (C) Prompt: 'photo of residential street in germany. features a church in the background'. (D) Selectively Edited Image. Church in Panel [C] was Highlighted and Edited With the Prompt: 'mosque'.

image editor (Midjourney, 2025a), allowing researchers to modify existing photos – such as original images of actors or local scenes – offering greater control over visual stimuli.

Validation of Images

As with any experimental stimulus, AI-generated visual vignettes must be validated to ensure they align with research objectives and are perceived by the participants as intended by the researcher.

First, images should meet basic quality standards – appearing natural, free of artifacts (e.g. distorted hands, unnatural facial expressions), and clearly depicting relevant visual elements and scenarios. This applies to AI-generated images as well as those edited manually or by designers. Given these basic qualities are met, images should be pretested with a sample from the target population to assess perception. Validation should focus on dimensions relevant to the research question, such as the recognisability of demographic traits or clarity of the shown context. Comparative ratings across images can help detect intended or unintended differences in a set.

Based on the findings from this pretest, images may need further refinement through prompt adjustments, regeneration, or targeted edits. Once validated, the final images can be administered in a factorial survey experiment following established procedures (Auspurg & Hinz, 2015).

What Do Respondents Think? Empirical Pretest Application of AI-Generated Images

The previous section has outlined how realistic visual vignettes can be created using Midjourney. However, it remains an open question how respondents perceive AI-generated content. As part of a broader study on decision-making and stereotypes, I created 18 ambiguous scenarios involving male individuals who may (or may not) be engaging in criminal activities (i.e. shoplifting, car theft, and burglary). The images varied by age (young, old) and race/ethnicity (White, Black, Arab) of the offender. Figure 6 shows three young men potentially shoplifting.

To select the most appropriate 18 images (from an initial pool of 45 meeting quality standards), I conducted a pretest with 300 German-speaking participants recruited via Prolific (three waves; wave 1: $n = 240$; wave 2: $n = 40$; wave 3: $n = 20$), where each participant rated up to three images (one from each scenario) in terms of perceived ethnicity, age, attractiveness, and social status of the individual depicted, as well as the perceived realism of the image. Respondents were debriefed at the end and informed that the images were AI-generated. The sample was gender-balanced, with a mean age of 33 years (SD = 10.6, range 18–72). Most participants (73%) were born in Germany and 83% identified as White, broadly matching the target population. The three survey waves enabled iterative refinement, ensuring the final set was unambiguous in ethnicity and comparable in age. As expected, perceived attractiveness and social status followed known ethnic hierarchies and stereotypes (see e.g. Bjornsdottir & Beacon, 2024).

To assess perceived realism, which was a relevant dimension in this application, participants answered a seven-point scale question ('Some participants in this survey were shown AI-generated images. How real or how artificial did the image you just saw appear'?) supplemented by an open-text response. Across 880 ratings (2–3 per person), the average realism score was 3.8 (SD = 2.0), with no significant gender differences and a weak negative correlation with age ($\beta = -0.01$, $p < 0.1$), indicating slightly higher realism ratings among older respondents. A significant order effect emerged: The first image viewed was rated as more realistic than subsequent ones (mean: 3.41 vs. 4.07; $\beta = -0.66$, $p < .001$). A mixed-effects model revealed that 18% of variance was between participants, while 82% reflected within-participant variation.

Many participants provided detailed open-text comments on what made images seem artificial or realistic. A total of 822 responses were manually coded, with multiple aspects noted per image (see Table 1; all quotes were originally in German and are my own translation). Common indicators of AI generation included unnatural facial features, inconsistent lighting, overly smooth skin, sharp edges, unreadable text, and exaggerated colours. Participants also looked for familiar AI artifacts, such as distorted hands or unusual eye positions – though convincing details could sway perceptions: '*In this image, the eyeball looks towards the camera, which makes me believe that the image was not AI generated*'.

Interestingly, features typical of professional photography, such as a blurred backgrounds or smooth textures, often triggered suspicion, whereas minor imperfections increased perceived realism. One respondent noted, '*The background seems lifelike and not too perfect*', while another commented, '*the image seems real because the person shown has flaws*'. Others flagged small visual discrepancies: '*the image looked very realistic – however, the eyebrows were very thinly plucked, the hairline appeared unnaturally straight, and the socks seemed overly smooth. But these are also characteristics one might see in a professional photoshoot*'. Overall, findings suggest that images perceived as 'too perfect' often appear artificial, while small flaws enhance

Table 1. Image Aspects and Their Frequencies in Open-Text Answers. Sorted by Frequencies of Being an Indication of Artificialness

Feature	Indication of..	Example statements	*n*
Facial features of person	Artificialness	'Weird eyes', 'the facial features looked AI-generated'	91
	Realness	'The young man had little birth marks on his face, making it real-looking', 'the eyeball looks towards the camera, which makes me believe that the image was not AI generated'	12
Sharpness, focus, edges	Artificialness	'Blurriness around the person', 'it looks hazy', 'the edges are too sharp'	80
	Realness	'AI-generated images are not this sharp'	10
Background setting	Artificialness	'The background seemed artificial'	49
	Realness	'The background looked very realistic'	22
Pose of person and their proportions	Artificialness	'The head does not fit to the body', 'the pose looks a bit unrealistic'	42
	Realness	'It is unlikely that AI would choose this pose'	5
Texture and quality	Artificialness	'Too much scrim diffuser'	44
	Realness	'AI-generated images often seem washed-out but this one does not'	4
Believability of situation	Artificialness	'The person does not fit to the scene', 'no one would behave like this'	31
	Realness	'Real everyday situation'	27
Colours	Artificialness	'Looks almost real, but the colours look edited', 'there seem to be a weird filter on the image'	26
	Realness	'The colours look real'	5
Hands	Artificialness	'The hand does not look real'	22
	Realness	'The hands looked normal'	3
Light, reflections	Artificialness	'The lightning conditions do not look real'	20
	Realness	'The shadows look natural'	2
Text	Artificialness	'The text on the label is not readable'	20
Person, general	Artificialness	'He does not look human-like'	19
	Realness	'The person looks real'	41
Specific other details	Artificialness	'The license plate on the car is too high', 'the wine bottle is too short'	19
	Realness	'The reflections on the car looked normal'	1
Composition, layout, perspective	Artificialness	'The image composition suggests AI'	14
	Realness	'With this angle, it does not look like an artificial image'	1
Person does not fit to background	Artificial	'Looks like a photomontage'	10
AI artifacts	Artificialness	'Obvious artifacts'	9
	Realness	'I have not notices anything that looks like AI'	99
Occurrence of flaws	Artificialness	'The face has too few flaws to belong to a real person'	7
	Realness	'The background looks life-like and not too perfect how it often is with AI-generated images'	11
Details, general	Artificialness	'It is low on details which is typical for AI'	7
	Realness	'With this level of detail, it does not look artificial'	1

realism. That said, similar critiques have been made of original, unedited images in other studies (see e.g. Eberl et al., 2022), highlighting general scepticism.

Overall, while many respondents generally acknowledged that the images appeared realistic (37% of the answers mentioned this explicitly, $n = 302/822$), they showed notable scepticism when directly prompted about authenticity. Many echoed sentiments as follows: '*It looks like a real photo, but I cannot be sure, as AI-generated images sometimes look just as real*' or '*I chose the middle category because I can no longer distinguish AI-generated images from real photos*'. Some revised their views after being prompted: '*The photo looked real – now I'm not so sure anymore*'. This aligns with the finding that later images were rated as less realistic, as several participants said they had been *'warned'* and thus approached subsequent images more critically. One noted: '*before reading this question, it never even occurred to me that the image could be AI-generated*'. These responses highlight how prompting participants to consider authenticity can heighten scepticism, even toward images they initially perceived as real.

In summary, participants responded positively to the AI-generated images, generally accepting them as credible stimuli. However, respondents were attuned to common AI-related inaccuracies. Many of the specific inaccuracies mentioned in this pretest (such as distorted hands, lighting inconsistencies) have already and are likely to diminish further as generative models improve, making AI-generated images increasingly realistic and visually credible. However, certain challenges are likely to persist: AI outputs remain shaped by their training data, which can overrepresent majority groups or common scenarios and underrepresent less typical contexts. This might even become amplified when AI learns from AI-generated content.

It is important to note that these perceptions are shaped by the current cultural context and form a snapshot in time (a snapshot taken in November 2024): Awareness of AI-generated media has been growing rapidly, and public reactions may change as AI images become more ubiquitous or politically charged. Social media discussions around 'AI slop' (see also Madsen & Puyt, 2025) suggest that some participants might respond negatively if they feel misled. This might expand to research contexts, making it more important to highlight that stimuli are hypothetical.

Limitations, Pitfalls and Recommendations

Generative AI provides powerful capabilities for creating visual stimuli for research purposes, but several limitations must be considered. While some can be managed through careful design and pretesting (see the next two sections), others – such as reliance on for-profit platforms and broader ethical concerns – require collective reflection within the research community (see the two following sections). The subsequent section offers practical recommendations.

Quality of and Biases in Images

While text-to-image models have made substantial progress, perfect high-resolution results cannot yet be fully expected. Manual human oversight is (at the moment) still needed to manage biases and ensure quality control. AI-generated outputs may sometimes appear unrealistic or inconsistent, and it is therefore necessary to inspect generated images and regenerate distorted areas until acceptable outputs are achieved. Acceptable does not necessarily mean fully photorealistic: Depending on research goals, impressionistic or stylised images may suffice as long as they clearly communicate the intended variation in experimental dimensions and maintain internal consistency. Whiles this manual supervision limits full, unsupervised automation, the relatively small number of distinct images needed for most experiments keeps the process scalable and practical.

A common concern is whether AI-generated images are recognisable as such. While this may or may not matter depending on research goals, differences in realism across image dimensions

could introduce bias. Evidence suggests people struggle to detect AI-generated content (Evsyukova et al., 2025; Nightingale & Farid, 2022). Results from the pretest of images created with Midjourney suggest that many respondents are now suspicious of any content when directly asked about its origin. To put this into context, comparing AI-generated images to real photos can help calibrate expectations.

Generative AI models are trained on existing content, which can make it difficult to emulate content that is less represented or absent from the web. The unknown training data can induce biases into the output, and it can be difficult to evaluate how pretraining shapes these outputs (Bianchi et al., 2023; Davidson, 2024). For example, Davidson (2024) shows that AI-generated images of Black Lives Matter protests often depict masked protestors – likely due to training on pandemic-era imagery. Researchers must be mindful of such learned biases, either by adjusting prompts or manually editing images (e.g. removing masks). Some biases may be subtle and harder to detect, underscoring the value of using stimulus sampling and multiple image variants to reduce dependency on any single image (Fong & Grimmer, 2023; Sambanis & Kyrkopoulou, 2025) – a process which is strongly simplified through the use of AI due to better scalability.

Consistency across manipulated images also poses challenges. Altering a person's race or ethnicity may change skin tone and facial structure in ways that improve plausibility but reduce researcher control (see Figure 7). Multiple iterations and selective edits – such as adjusting gaze or facial hair – are often necessary to achieve visual parity. Importantly, since AI learns from real-world images, it may reproduce common stereotypes. As an alternative, researchers might use AI images in tandem with morphing tools that offer finer control over facial features (see e.g. Evsyukova et al., 2025).

Figure 7. Images of Adolescents Shopping, Created by Midjourney. (A) Prompt: 'photo of a black shoplifter standing between the aisles. he is wearing a black oversized bomber jacket and it is unzipped. the jacket is open. his left hand is in his jacket pocket, with the right hand he grabs a bottle of red wine and hides:5 it under his jacket. he looks at the shelf:1 https://s.mj.run/JY_IJeh0wKA'. Selectively Edited Image in Several Rounds by Highlighting Part of the Images to Change the Look of the Person. (B) Selectively Edited Image. Face and Hands of Person in Panel [A] Were Highlighted and Edited With the Prompt: 'arab teenager, age 15. he looks muslim. he looks suspicious as he is stealing something. his head is slightly tilted looking to the right'. (C) Selectively Edited Image. Face and Hands of Person in Panel [A] Were Highlighted and Edited With the Prompt: 'white teenager, age 17. blond hair and bright skin. a little bit of beard. looking suspiciously'.

Use of Visual Stimuli in Experimental Social Science Research

Beyond the specific constraints of AI-generated images, visual stimuli pose broader challenges in experimental research. A key issue is their context-dependency and interpretative ambiguity. Like text, the interpretation of visual stimuli can vary based on cultural, contextual, or individual factors, but they may introduce more unintended variance than precisely worded text. Therefore, images must be designed with a clear awareness of the target audience and need to be pretested accordingly.

Another limitation lies in the restricted number of manipulable dimensions. Unlike textual vignettes, which can convey both visible and abstract traits (e.g. education and personality), visual stimuli are confined to directly observable features. Representing more abstract attributes visually often requires reliance on stereotypes, which can introduce bias or misinterpretation.

Further, visual stimuli are often better suited to between-subject designs, where each participant sees only one version of a scenario. Within-subject designs – where participants view several versions of the same scene with small variations (e.g. ethnicity, attire, or background) – can appear unnatural and confuse respondents. In such cases, researchers should design stimuli with greater variation between images to mask experimental manipulations and avoid detection of the study's intent.

Dependency on Corporations: Closed-Source and Open-Source Software

This article used the closed-source software Midjourney to create images. There are important trade-offs between open-source and closed-source software. Currently, the most powerful and easy-to-use models with remarkable out-of-the-box quality are developed by for-profit companies. They are easy to access with minimal setup requirements, allowing researchers to produce high-quality images quickly without the technical overhead of self-hosting or fine-tuning a model. This makes them a straight-forward choice for short-term projects with limited timelines, such as many vignette experiments.

However, this introduces a dependence on corporations which have their own guidelines in place and can modify models and access options without notice. For instance, when using Midjourney, there are content guidelines in place which restrict the creation of certain scenes (e.g. violent scenes or those containing nudity). Some image and text prompts are automatically blocked. This can hinder researchers aiming to study biases or behaviours related to sensitive topics.

Open-source models (e.g. StableDiffusion) can rival these proprietary solutions, but their use requires access to expensive technical infrastructure and more advanced computational training, especially when using features like the selective variation of areas. While closed-source models are currently a more viable solution for most social scientists, as argued by Davidson (2024), in the long run, academic researchers will be best served by generative AI developed and maintained specifically for research purposes (Bail, 2024; Grossmann et al., 2023). Looking forward, hybrid approaches might appear and provide a preferable soluation: robust open-source models with cloud-based, user-friendly interfaces, or consortia-driven platforms tailored to research. Such developments could allow research teams to generate, validate, and share stimuli more reliably, while reducing dependency on proprietary systems and enabling longer-term replication and cumulative research.

Ethical Concerns

A key ethical question is whether it is morally sound to present research participants with AI-generated content. Eberl et al. (2022) argue that respondents are inherently deceived if deepfakes are used without explicit disclosure. In the social sciences, the use of deception is debated (Barrera & Simpson, 2012), but visual vignettes – like textual ones – are typically framed as hypothetical scenarios. When no claim is made that an image is real, and the scenario is clearly identified as fictional, I would argue that no deception occurs. Still, researchers should critically reflect on their intent when generating AI-based imagery (de Ruiter, 2021).

Beyond deception, broader ethical concerns surround the use of generative AI. Developers have faced accusations of violating copyright laws, with lawsuits from content creators arguing that AI systems exploit their work without compensation. From this perspective, using generative AI may be seen as endorsing unethical practices, benefiting from the unpaid labour and creativity of others (Goetze, 2024). There are also further labour and environmental concerns. Generative AI tools may displace professionals such as models, photographers, and designers, reducing opportunities in creative industries.[5] Furthermore, AI models require substantial computational resources during training and generation, resulting in high energy consumption and a significant environmental footprint (Bourzac, 2024).

These issues raise important questions for researchers, as the decision to use such tools is both an individual and collective ethical choice – one that warrants active discussion within the research community. As of now, no established best practices guide these decisions.

Recommendations

Based on the insights presented in this paper and the limitations discussed, I propose the following key recommendations for effectively integrating AI-generated images into vignette experiments:

Ensure Thematic and Methodological Fit. Before incorporating AI-generated images, researchers should carefully assess whether visual stimuli meaningfully enhance the vignette experiment. Not all research questions benefit equally from images, and their use should align with both theoretical and methodological considerations. In addition, not all research questions benefit equally from AI-generated images. Real images might be preferable when the recognisability of specific people or places is important. In contrast, AI-generated images are particularly advantageous when researchers need to depict ethically sensitive or logistically difficult scenarios. A hybrid approach – starting with real images and selectively editing them with AI – can combine the authenticity of photographs with the flexibility of generative tools.

Consider Ethical and Data Protection Implications. The use of AI-generated human representations, especially for sensitive topics, can raise ethical concerns. Researchers should adhere to ethical guidelines, consider the potential impact on participants, and ensure compliance with data protection regulations.

Select Appropriate Software. A variety of proprietary and open-source software is available for generating images. Researchers should stay informed about current developments to choose the most suitable tool based on their needs. However, it is not always necessary or even advisable to chase the latest model. Researchers might benefit more from selecting a model which offers all the features, mastering its capabilities, and maintaining consistency across experiments than from constantly switching to newer versions. Technical developments in generative AI typically outpace the timelines of social science research projects, and frequent changes can introduce unnecessary variability, compromise reproducibility, and steepen the learning curve for research

teams. In many cases, older models already produce images that are sufficiently realistic and flexible for experimental purposes, especially when paired with careful prompting, selective editing, and pretesting. Generally, when possible, open-source, self-hosted, or potential research-focused versions should be prioritised.

Generate Suitable Images, Standardise Image Variation and Control for Bias. Images can be generated using specific and clear prompts. Experimentation in phrasing, the careful reference towards stereotypes, and refinements using images and prompts can be used to achieve appropriate baseline images. The use of seed parameters can further aid reproducibility. These AI-generated images – or pre-existing baseline images – should then be systematically varied while ensuring that irrelevant factors do not unintentionally influence responses.

Pretest Images for Validity and Comprehension. To ensure that participants interpret and perceive AI-generated images as intended, pretesting and validation is essential. Small-scale pilot studies can help identify ambiguities or unintended associations and ensure comparable image quality across image variations. Participants should rate the images in terms of the vignette dimensions and relevant confounders (when varying persons, this might include aspects like age, trustworthiness, reliability, etc.).

By following these practical recommendations, researchers can maximise the methodological rigour and validity of AI-assisted vignette experiments while minimising potential pitfalls.

Conclusions

Photos and images are everywhere – appearing in news articles, social media posts, or advertisements. Despite their ubiquity, visual stimuli are rarely used in factorial survey experiments due to the difficulty of controlling image variations. This article discussed how generative AI offers a powerful new way to create customised visual stimuli for research, making the process more accessible and efficient. Unlike traditional methods, AI can produce custom images based on specific criteria without the need for recruiting models, staging complex scenes, or relying on expensive graphic design services. Using the image generator Midjourney as an example, I demonstrated how generative AI can be used to selectively vary specific dimensions of images for factorial survey experiments.

AI-generated images can find applications as stimuli in other experiments, both within surveys and in the field. For example, some correspondence audit studies (field experiments in which researchers send fictious applications to employers to study discrimination) have attached images or videos to applicants' CVs – instead of relying on textual information – to study the effect of attractiveness (e.g. Kühn & Wolbring, 2024; Rooth, 2009) or to signal ethnicity/race (e.g. Polavieja et al., 2023) or physical disability (e.g. Stone & Wright, 2013). Additionally, field experiments which take place on online platforms have also made use of variations in images to compare different scenarios (e.g. Doleac & Stein, 2013; Evsyukova et al., 2025). While images cannot do everything – only directly observable characteristics can be included in images – they offer an engaging new stimulus format for participants which, with the use of generative AI, does not lack controllability anymore and can also be implemented at scale. Further survey methodological research is still needed to systematically compare and contrast different formats of vignettes (Sauer et al., 2020; Shamon et al., 2022).

Generative AI can expand social scientists' methodological horizons, but they need to be approached with some caution. Text-to-image generators are, as of now, able to produce high-quality images, but they are not perfect – however, as the technology is going forward, these models will only improve, making them even better suited for experimental social science research.

Acknowledgements

The author thanks Ulf Liebe, Didier Ruedin, and Juliane Kühn for their constructive comments which greatly improved the article. The author also thanks participants at several workshops/research seminars, including the *Workshop on Experimental Sociology* and the *MZES Workshop on Novel Data Sources and Methods for Understanding Discrimination and Bias* for helpful feedback and suggestions. Financial support from the Postdoc Career Academy of the University of Mannheim is gratefully acknowledged.

ORCID iD

Nicole Schwitter https://orcid.org/0000-0002-3837-680X

Ethical Considerations

The Ethics Review Committee at University of Mannheim approved the pretest surveys (approval: EK Mannheim 65/2024) on October 22, 2024.

Consent to Participate

Respondents gave written consent before starting the survey.

Funding

The author disclosed receipt of the following financial support for the research of this article: This work was supported by the Postdoc Career Academy of the University of Mannheim and the German Research Foundation (DFG)'s Emmy-Noether Program (ZH 613/1-1).

Declaration of Conflicting Interests

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

Data Availability Statement

The data and code underlying this study are available here: https://github.com/nschwitter/AI_vignettes_SSCR.

Notes

1. A somewhat related development is the presentation of textual information embedded into visuals to provide a more realistic presentation of the vignette, for example, by showing hypothetical newspaper articles or social media posts (Smetana et al., 2024).
2. https://stable-diffusion-art.com/controlnet/.
3. Since the writing and revising of this manuscript, the available software has further evolved. As of October 2025, OpenAI's Sora has gained traction (especially due to its praised prompt adherence, meaning that it tends to understand prompts better) and Google's Gemini by now also features powerful image editing options.
4. https://www.midjourney.com/explore.
5. In the case of creating visual vignettes, these models, photographers, and graphic designers are often (student) volunteers or members of the research team themselves, making this point less applicable.

References

Arseniev-Koehler, A., & Foster, J. G. (2022). Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research*, *51*(4), 1484–1539. https://doi.org/10.1177/00491241221122603

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. Sage Publications, Inc. https://doi.org/10.4135/9781483398075

Auspurg, K., Hinz, T., & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden, Daten, Analysen*, *3*(1), 59–96. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-126659

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences of the United States of America*, *121*(21), Article e2314021121. https://doi.org/10.1073/pnas.2314021121

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021). Conjoint survey experiments. In D. P. Green & J. N. Druckman (Eds.), *Advances in experimental political science* (pp. 19–41). Cambridge University Press. https://doi.org/10.1017/9781108777919.004

Baptista, I., Spence, C., Shimizu, R., Ferreira, E., & Behrens, J. (2023). Color is to flavor as shape is to texture: A choice-based conjoint study of visual cues on chocolate packaging. *Journal of Sensory Studies*, *38*(1), Article e12793. https://doi.org/10.1111/joss.12793

Barabas, J., & Jerit, J. (2010). Are survey experiments externally valid? *American Political Science Review*, *104*(2), 226–242. https://doi.org/10.1017/S0003055410000092

Baron, S. W., Forde, D. R., & Kennedy, L. W. (2001). Rough justice: Street youth and violence. *Journal of Interpersonal Violence*, *16*(7), 662–678. https://doi.org/10.1177/088626001016007003

Barrera, D., & Simpson, B. (2012). Much ado about deception: Consequences of deceiving research participants in the social sciences. *Sociological Methods & Research*, *41*(3), 383–413. https://doi.org/10.1177/0049124112452526

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June, 2023, pp. 1493–1504. https://doi.org/10.1145/3593013.3594095

Bjornsdottir, R. T., & Beacon, E. (2024). Stereotypes bias social class perception from faces: The roles of race, gender, affect, and attractiveness. *Quarterly Journal of Experimental Psychology*, *77*(11), 2339–2353. https://doi.org/10.1177/17470218241230469

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings (arXiv:1607.06520)*. arXiv. https://doi.org/10.48550/arXiv.1607.06520

Bourzac, K. (2024). Fixing AI's energy crisis. *Nature*. Online ahead of print. https://doi.org/10.1038/d41586-024-03408-z

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 529–533. https://doi.org/10.1016/S0022-5371(70)80096-2

Burt, C. D. B., Halloumis, S. A., McIntyre, S., & Blackmore, H. S. (2010). Using colleague and team photographs in recruitment advertisements: Effects on applicant attraction. *Asia Pacific Journal of Human Resources*, *48*(2), 233–250. https://doi.org/10.1177/1038411109355358

Caro, F. G., Ho, T., McFadden, D., Gottlieb, A. S., Yee, C., Chan, T., & Winter, J. (2012). Using the internet to administer more realistic vignette experiments. *Social Science Computer Review*, *30*(2), 184–201. https://doi.org/10.1177/0894439310391376

Caruso, E. M., Mead, N. L., & Balcetis, E. (2009). Political partisanship influences perception of biracial candidates' skin tone. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20168–20173. https://doi.org/10.1073/pnas.0905362106

Davidson, T. (2024). Start generating: Harnessing generative artificial intelligence for sociological research. Socius: Sociological Research for a Dynamic World, 10, Article 23780231241259651. https://doi.org/10.1177/23780231241259651

DeLong, K. L., Syrengelas, K. G., Grebitus, C., & Nayga, R. M. (2021). Visual versus text attribute representation in choice experiments. *Journal of Behavioral and Experimental Economics*, *94*, Article 101729. https://doi.org/10.1016/j.socec.2021.101729

de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

Doleac, J. L., & Stein, L. C. D. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469–F492. https://doi.org/10.1111/ecoj.12082

Don-Yehiya, S., Choshen, L., & Abend, O. (2023). *Human learning by model feedback: The dynamics of iterative prompting with Midjourney (arXiv:2311.12131)*. arXiv. https://doi.org/10.48550/arXiv.2311.12131

Eberl, A., Kühn, J., & Wolbring, T. (2022). Using deepfakes for experiments in the social sciences—A pilot study. *Frontiers in Sociology*, *7*, Article 907199. https://doi.org/10.3389/fsoc.2022.907199

Eifler, S., & Petzold, K. (2022). Fear of the dark? A systematic comparison of written vignettes and photo vignettes in a factorial survey experiment on fear of crime. *Methods, Data, Analyses*, *16*(2), 201–234. https://doi.org/10.12758/mda.2022.01

Evsyukova, Y., Rusche, F., & Mill, W. (2025). LinkedOut? A field experiment on discrimination in job network formation. *The Quarterly Journal of Economics*, *140*(1), 283–334. https://doi.org/10.1093/qje/qjae035

Facciani, M., Brashears, M. E., & Zhong, J. (2022). Visual vignettes for cross-national research. *International Journal of Social Research Methodology*, *25*(1), 29–43. https://doi.org/10.1080/13645579.2020.1844897

Finch, J. (1987). The vignette technique in survey research. *Sociology*, *21*(1), 105–114. https://doi.org/10.1177/0038038587021001008

Fong, C., & Grimmer, J. (2023). Causal inference with latent treatments. *American Journal of Political Science*, *67*(2), 374–389. https://doi.org/10.1111/ajps.12649

Gaddis, S. M. (2017). How Black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science*, *4*, 469–489. https://doi.org/10.15195/v4.a19

Goetze, T. S. (2024). AI art is theft: Labour, extraction, and exploitation: Or, on the dangers of stochastic pollocks. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June, 2024, pp. 186–196. https://doi.org/10.1145/3630106.3658898

Golden, J. H., Johnson, C. A., & Lopez, R. A. (2001). Sexual harassment in the workplace: Exploring the effects of attractiveness on perception of harassment. *Sex Roles*, *45*(11), 767–784. https://doi.org/10.1023/A:1015688303023

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual Studies*, *17*(1), 13–26. https://doi.org/10.1080/14725860220137345

Haut, K., Wohn, C., Antony, V., Goldfarb, A., Welsh, M., Sumanthiran, D., Jang, J., Ali, M. R., & Hoque, E. (2021). *Could you become more credible by being White? Assessing impact of race on credibility with deepfakes (arXiv:2102.08054)*. arXiv. https://doi.org/10.48550/arXiv.2102.08054

Havekes, E., Coenders, M., & van der Lippe, T. (2013). Positive or negative ethnic encounters in urban neighbourhoods? A photo experiment on the net impact of ethnicity and neighbourhood context on attitudes towards minority and majority residents. *Social Science Research*, *42*(4), 1077–1091. https://doi.org/10.1016/j.ssresearch.2013.02.002

Hu, M., Lee, S., Xu, H., Melipillán, R., Smith, J., & Kapteyn, A. (2022). Improving anchoring vignette methodology in health surveys with image vignettes. *Methoden, Daten, Analysen*, *16*(2), 273–314. https://doi.org/10.12758/mda.2022.02

Jahani, E., Manning, B. S., Zhang, J., TuYe, H.-Y., Alsobay, M., Nicolaides, C., Suri, S., & Holtz, D. (2025). *Prompt adaptation as a dynamic complement in generative AI systems (arXiv:2407.14333)*. arXiv. https://doi.org/10.48550/arXiv.2407.14333

Jiwa, M., Halkett, G., Meng, X., Pillai, V., Berg, M., & Shaw, T. (2014). Supporting patients treated for prostate cancer: A video vignette study with an email-based educational program in general practice. *Journal of Medical Internet Research*, *16*(2), Article e63. https://doi.org/10.2196/jmir.3003

Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics*, *22*(1), 1–13. https://doi.org/10.1002/hec.1806

Kabaya, K., Tajima, K., Ichinose, D., & Asano, M. (2024). Do different visual presentation formats encourage different choice behaviors? Discrete choice experiment on urban park landscapes. *Environmental Economics and Policy Studies*, *27*(1), 23–41. https://doi.org/10.1007/s10018-024-00405-4

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405

Krakowski, K., Kyrkopoulou, E., Reinhardt, L., & Sambanis, N. (2024). Good citizenship and native-immigrant conflict: Experimental evidence from Europe. *Comparative Political Studies*, *58*(13), 2939–2972. https://doi.org/10.1177/00104140241306944

Krysan, M., Couper, M. P., Farley, R., & Forman, T. A. (2009). Does race matter in neighborhood preferences? Results from a video experiment. *American Journal of Sociology*, *115*(2), 527–559. https://doi.org/10.1086/599248

Kühn, J., & Wolbring, T. (2024). Beauty pays, but not under all circumstances: Evidence on gendered hiring discrimination from a novel experimental treatment using deepfakes. *Research in Social Stratification and Mobility*, *94*, Article 100992. https://doi.org/10.1016/j.rssm.2024.100992

Loosschilder, G. H., Rosbergen, E., Vriens, M., & Wittink, D. R. (1995). Pictorial stimuli in conjoint analysis. *Market Research Society Journal*, *37*(1), 1–15. https://doi.org/10.1177/147078539503700104

López Ortega, A., & Radojevic, M. (2024). Visual conjoint vs. text conjoint and the differential discriminatory effect of (visible) social categories. *Political Behavior*, *47*(1), 335–353. https://doi.org/10.1007/s11109-024-09953-7

Madsen, D. Ø., & Puyt, R. W. (2025). When AI turns culture into slop. *AI & SOCIETY*, Onlinefirst. https://doi.org/10.1007/s00146-025-02630-1

Manghani, S. (2012). *Image studies: Theory and practice*. Routledge. https://doi.org/10.4324/9780203134917

Midjourney. (2025a). External editor. External Editor. https://docs.midjourney.com/docs/external-editor (last accessed 2025-10-24).

Midjourney. (2025b). Midjourney parameter list. https://docs.midjourney.com/docs/parameter-list (last accessed 2025-10-24).

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. *Social Psychology*, *45*(4), 299–300.

Mugglin, L., Murahwa, B., & Ruedin, D. (2025). When politicians feel pressure to represent: Evidence from South Africa. *Parliamentary Affairs*, *78*(4), 789–815. https://doi.org/10.1093/pa/gsae046

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(8), Article e2120481119. https://doi.org/10.1073/pnas.2120481119

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press. https://doi.org/10.4324/9781315798868

Polavieja, J. G., Lancee, B., Ramos, M., Veit, S., & Yemane, R. (2023). In your face: A comparative field experiment on racial discrimination in Europe. *Socio-Economic Review*, *21*(3), 1551–1578. https://doi.org/10.1093/ser/mwad009

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception & Psychophysics*, *76*(2), 270–279. https://doi.org/10.3758/s13414-013-0605-z

Rooth, D.-O. (2009). Obesity, attractiveness, and differential treatment in hiring: A field experiment. *Journal of Human Resources*, *44*(3), 710–735. https://doi.org/10.3368/jhr.44.3.710

Sambanis, N., & Kyrkopoulou, E. (2025). O Xenos: Racial difference, colorism, & xenophobia in Greece (SSRN scholarly paper 5133525). *Social Science Research Network*. https://doi.org/10.2139/ssrn.5133525

Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing multi-factorial survey experiments: Effects of presentation style (text or table), answering scales, and vignette order. *Methods, Data, Analyses*, *14*(2), 195–214. https://doi.org/10.12758/MDA.2020.06

Schlochtermeier, L. H., Kuchinke, L., Pehrs, C., Urton, K., Kappelhoff, H., & Jacobs, A. M. (2013). Emotional picture and word processing: An fMRI study on effects of stimulus complexity. *PLOS ONE*, *8*(2), Article e55619. https://doi.org/10.1371/journal.pone.0055619

Shamon, H., Dülmer, H., & Giza, A. (2022). The factorial survey: The impact of the presentation format of vignettes on answer behavior and processing time. *Sociological Methods & Research*, *51*(1), 396–438. https://doi.org/10.1177/0049124119852382

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7

Shr, Y.-H., Ready, R., Orland, B., & Echols, S. (2019). How do visual representations influence survey responses? Evidence from a choice experiment on landscape attributes of green infrastructure. *Ecological Economics*, *156*, 375–386. https://doi.org/10.1016/j.ecolecon.2018.10.015

Sidler, P., Knotz, C., & Ruedin, D. (2024). How do people perceive immigrants? Relating perceptions to numbers. *The Journal of Race, Ethnicity, and Politics*, *9*(3), 689–708. https://doi.org/10.1017/rep.2024.18

Slattery Rashotte, L. (2003). Written versus visual stimuli in the study of impression formation. *Social Science Research*, *32*(2), 278–293. https://doi.org/10.1016/S0049-089X(02)00050-9

Smetana, M., Vranka, M., & Rosendorf, O. (2024). The "commitment trap" revisited: Experimental evidence on ambiguous nuclear threats. *Journal of Experimental Political Science*, *11*(1), 64–77. https://doi.org/10.1017/XPS.2023.8

Stone, A., & Wright, T. (2013). When your face doesn't fit: Employment discrimination against people with facial disfigurements: When your face doesn't fit. *Journal of Applied Social Psychology*, *43*(3), 515–526. https://doi.org/10.1111/j.1559-1816.2013.01032.x

Sturgis, P., & Luff, R. (2021). The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015. *International Journal of Social Research Methodology*, *24*(6), 691–696. https://doi.org/10.1080/13645579.2020.1844896

Teti, A., Gross, C., Knoll, N., & Blüher, S. (2016). Feasibility of the factorial survey method in aging research: Consistency effects among older respondents. *Research on Aging*, *38*(7), 715–741. https://doi.org/10.1177/0164027515600767

Thébaud, S., Kornrich, S., & Ruppanner, L. (2021). Good housekeeping, great expectations: Gender and housework norms. *Sociological Methods & Research*, *50*(3), 1186–1214. https://doi.org/10.1177/0049124119852395

Treischl, E., & Wolbring, T. (2022). The past, present and future of factorial survey experiments: A review for the social sciences. *Methods, Data, Analyses*, *16*(2), 141. https://doi.org/10.12758/mda.2021.07

Tuscher, M. (2022). Processing speed and comprehensibility of visualizations and texts. *Proceedings of CESCG*.

van Zelderen, A. P. A., Masters-Waage, T. C., Dries, N., Menges, J. I., & Sanchez, D. R. (2024). Simulating virtual organizations for research: A comparative empirical evaluation of text-based, video, and virtual reality video vignettes. *Organizational Research Methods*, *28*(3), 457–486. https://doi.org/10.1177/10944281241246770

Vecchiato, A., & Munger, K. (2024). Introducing the visual conjoint, with an application to candidate evaluation on social media. *Journal of Experimental Political Science*, *12*(1), 1–15. https://doi.org/10.1017/XPS.2024.15

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520. https://doi.org/10.1016/j.ssresearch.2009.03.004

Wouters, R., & Walgrave, S. (2017). Demonstrating power: How protest persuades political representatives. *American Sociological Review*, *82*(2), 361–383. https://doi.org/10.1177/0003122417690325

Xie, Y., Pan, Z., Ma, J., Jie, L., & Mei, Q. (2023). A prompt log analysis of text-to-image generation systems. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May, 2023, pp. 3892–3902. https://doi.org/10.1145/3543507.3587430

Author Biography

Nicole Schwitter is a postdoctoral researcher at the Mannheim Centre for European Social Research (University of Mannheim) and an honorary research fellow at the University of Warwick. Her research interests include computational social science, migration and discrimination