Using AI-Generated Videos and Storyboards as Elicitation Prompts in Human Factors Research

Catherine Harvey, David R Large, Benjamin Bowers & Iris Jestin

University of Nottingham

SUMMARY

This paper presents a novel methodological approach, illustrated by two contrasting case studies – one situated in a safety-critical military context and the other in everyday driving – that utilise generative artificial intelligence (GenAI) to create, visualise and communicate concepts for use in participatory ergonomics. We reflect on the approach, providing a detailed four-stage methodology, and discuss the implications from a human factors perspective.

KEYWORDS

Generative AI, GenAI, ChatGPT, interviews, focus groups, participatory ergonomics

Introduction

Human factors research frequently explores how people perceive, interact with, and adapt to emerging technologies and systems. Such investigations often rely on scenario-based methods using visual prompts, including storyboards (Truong & Abowd, 2006) and video clips (Neubauer et al, 2021), or written vignettes (Farraro & Moulouna, 2022); these help participants imagine system interactions beyond their immediate experience and to articulate their tacit knowledge about factors such as usability, workload or trust. Traditional methods used to create visual prompts, such as, professionally illustrated storyboards or professionally produced film vignettes (Nørskov & Damholdt, 2020), demand considerable time, expertise, and cost, and once produced, lack flexibility: they are difficult or costly to modify in response to changes in study design or participant feedback. Recent advances in generative AI (GenAI) – particularly, text-to-image and text-to-video systems (for example, ChatGPT, Google Veo 3), enable the creation of visual artefacts from text-based descriptions. Researchers can describe a specific situation or context using written language (for example, *"interactions between submariners and an autonomous agent in a Royal Navy nuclear-powered fleet submarine"*) and generate tailored visual materials using publicly available software within minutes. This overcomes the constraints of time and cost but also allows flexibility by enabling rapid redesign as needs change. Moreover, as the aforementioned example intimates, GenAI content allows researchers to depict complex or speculative technologies operating within sensitive or safety-critical scenarios.

We used GenAI to create elicitation prompts for interviews and focus group studies. This paper presents two of these as cases studies followed by an outline of the methodology (including our reflections on using it) and a discussion on its application and suitability, highlighting areas for further exploration and/or refinement.

Case Study One: Evaluating Human-Autonomy Teaming in Safety-Critical Operations

In the first case study, we wanted to explore Human-Autonomy Teaming (HAT) in a safety-critical operation. To achieve this, we needed to recreate a genuine maritime event, in which a Royal Navy Trafalgar-class submarine collided with a fishing boat (*Karen incident;* MAIB, 2016). Using the incident investigation report, three scenarios were created as vignettes – the first with a human-only crew and two human-autonomy variants, each depicting the same unfolding events. The latter variations introduced a voice-based AI agent with either a high or low level of autonomy, allowing us to vary team composition, command structure and function allocation that represent central features of hybrid teams (Parasuraman et al, 2000; 2007). The vignettes were used to explore the applicability and sensitivity of a battery of rating scales commonly used in HAT research, and elicit interview responses on perceived effectiveness across conditions, preferred AI roles in teams, and evaluation-relevant criteria not captured by surveys. This mixed approach allowed us to uncover failure attribution differences in HAT evaluation and novel considerations for future military test and evaluation.

Figure 1: A snapshot from the AI-generated vignette used in Case Study 1

An AI-based video generation tool (Google Veo 3, September 2025) allowed us to generate authentic video-based vignettes, without needing to create a working prototype of an autonomous agent, enter restricted facilities or expose human actors to unnecessary risk. A future in which autonomous technology works seamlessly alongside human operators is frequently discussed (O'Neill et al, 2020), but difficult to empirically examine in situ given (1) uncertainty around future AI capability, (2) how such systems will ultimately manifest in practice, and (3) the cost and complexity of high-fidelity prototyping and HAT simulation. AI-based video generation therefore provided a comparatively simple and low-cost tool to visualise future safety-critical operations through carefully curated prompts, such as the one used to generate the vignette depicted above:

> *"Generate an interior scene inside a Royal Navy nuclear-powered fleet submarine, specifically the Trafalgar-class. The setting is the control room. It is a compact, metallic interior with grey and green surfaces, dimly lit with red lighting. Multiple sonar operators in dark Royal Navy uniforms monitor bearing lines and acoustic data. The AI Tactical Overlay fades in across the main sonar repeater screen, filling the screen edge-to-edge with translucent cyan colour. A calm, synthetic voice with a slight digital timbre, audible to everyone, announces: 'Thirty-six contacts detected within ten nautical miles. Recommend*

> *merchant-only collision-avoidance protocol." The commanding officer says to the team of sonar operators in a male British accent: "Suspend close quarters procedures."*

In this example, the prompt language used for the control room description was iteratively refined and then held constant across the different conditions. Refinement involved achieving the required level of detail sufficient to constrain the model while avoiding prompt overload that can cause output mistakes. The human/AI agent appearance and voice sections of the prompt employed condition-dependent framing, allowing us to manipulate agent role and command structure while keeping the operational timeline and key events aligned with the MAIB (2016) report, using directly reported language, where feasible.

Case Study Two: Uncovering Drivers' Tacit Knowledge of Driver Monitoring Systems

In the second case study, we wanted to uncover drivers' tacit knowledge and current understanding of driver monitoring systems (DMS) and therefore needed to communicate several driving-related scenarios for a focus group study. We required visual prompts that could demonstrate a series of unfolding events (e.g. as a storyboard), to show the driving context, the driver state or behaviour warranting intervention, and the proposed intervention. However, we wanted to avoid depicting specific technologies or a particular vehicle make or model. In addition, we wanted to depict possible future applications of driver monitoring systems, for which no systems currently exist.

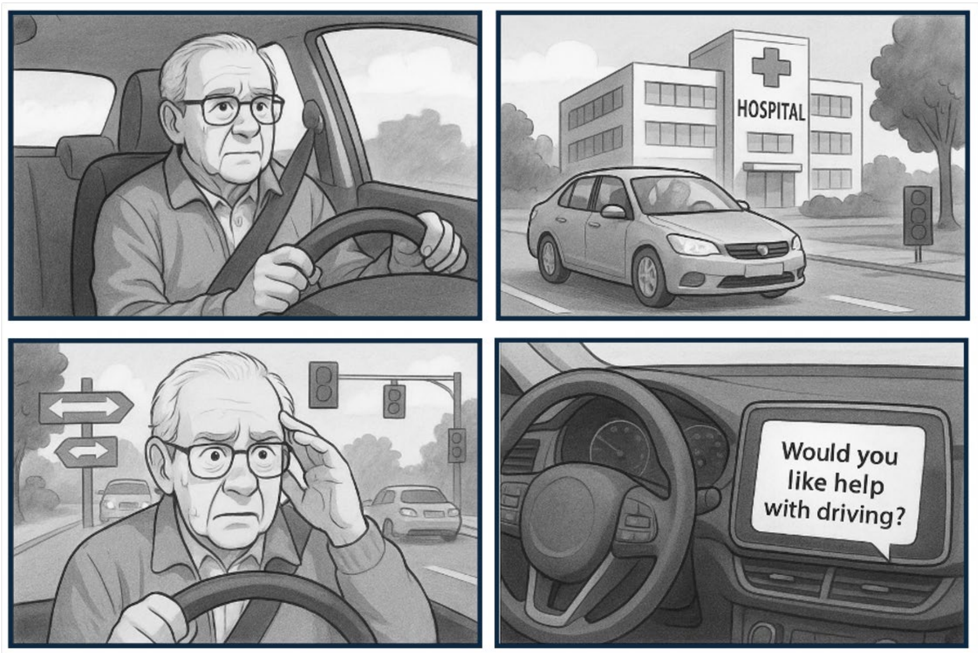

Figure 2: An example of a storyboard used in Case Study 2

In the example in Figure 2, an older driver appears to be confused, which in this instance we imagined to be due to cognitive decline. Thus, the storyboard depicts a future scenario in which the car has detected aspects of driving behaviour that may be indicative of cognitive decline and offers support with the driving task.

Using GenAI to create the storyboards ensured that the examples lacked technical specificity but were still capable of communicating the desired aspects of driver monitoring. Chat-GPT was used to generate the storyboards in July 2025. Text prompts were designed to contextually situate the scenario. They specifically avoided the use of the terms 'monitor' and 'diver monitoring system', and avoided referencing any specific technology, such as a 'driver facing camera'. All descriptions were presented in the third person to avoid superiority and social desirability bias, with names initially selected from common gender-neutral options. The final prompt used to generate the storyboard in Figure 2 is below.

> *"Please create a black and white storyboard in comic-strip style with 4 images to show the following situation. Roger is an older driver. Roger is driving to an appointment at the hospital. At a junction, Roger looks very confused and has difficulty concentrating. After a long time waiting, the car asks if Roger would like help with his driving."*

In this example, only minor changes to the text prompt were made between three iterations before achieving the desired result. This included changing the name and description of the driver state and behaviour. The use of gender-neutral names also caused some confusion for the AI model, with the first depiction confusing genders between panels. No further manual edits were made to the storyboard – although it is noted that in the third panel, Roger has lost the top of his car.

Methodological Approach

The methodological approach consists of four key stages:

1. Prompt Design and Generation of Visual Artefacts

GenAI uses a text prompt (which could be captured through speech) to create storyboard images or video clips. However, crafting the correct prompt is critical to success and works best as an iterative process. Prompts must have enough detail to guide the AI model clearly, but not so much detail that the prompt becomes restrictive or contradictory. Too little detail often leads to vague or generic outputs, or hallucinations, because the model must fill in gaps using general statistical patterns. This can result in content that misses the author's intent, lacks coherence, or defaults to common clichés. Too much detail, however, can also reduce output quality. Overly long or restrictive prompts may include conflicting instructions, unnecessary constraints, or stylistic overload, which can confuse the model or lead to outputs that feel forced, fragmented, or less creative. Excessive detail can also make prompts harder to reuse and adapt.

We therefore recommend that the prompt describes the environmental conditions, user activities, emotional tone and desired outcome in sufficient detail, including any constraints. Potential practitioners could also consider allowing flexibility (i.e. limit text instructions) where GenAI creativity or interpretation is beneficial, or where elements of the scenario may be unknown, complex or highly speculative. In addition, we recommend starting with a clear but moderate level of detail, reviewing the output, and then refining the prompt as needed. This balance maximises both control and quality in GenAI outputs.

2. Validation and Curation

Generated materials must be reviewed for visual coherence, cultural representation and absence of bias. Images may also contain persistent unrequired artefacts, even after multiple iterations. Manual editing may therefore be required to correct any deficiencies, improve realism or clarity, and ensure alignment with specific design requirements and compliance with legal or institutional norms. From a creative point of view, manual editing supports human authorship and intentionality. While GenAI can rapidly explore visual possibilities, human intervention enables selective refinement, contextual judgment and aesthetic decision-making. Editing thus transforms AI outputs from automated suggestions into collaboratively authored works, preserving the role of human creativity rather than replacing it.

For videos, although GenAI technologies have progressed rapidly, outputs may be limited in duration (for example, ours were restricted to 8s) or exhibit anomalies such as temporal inconsistencies, visual artefacts, unstable motion and inaccuracies in object continuity across frames. Issues including flickering, distorted anatomy or inconsistent lighting can be common due to the complexity of modelling time-dependent visual data. Manual editing of videos may therefore be required but this can be inherently more complicated than editing a still image. In our example, the videos were created as multiple short clips, partly due to the time limitations of GenAI video generation, but also to ensure that all elements of the required narrative were covered. As a consequence, the clips need to be joined or spliced together. A further element to consider is the diegetic sound, or soundtrack. To avoid complication, sound can be generated independently and overlaid on the combined video.

In practice, we recommend that in an exploratory context – such as the current applications – speed and accessibility of generative AI images and videos may be acceptable without extensive manual editing. Consequently, although some unusual artefacts may remain, these may be an acceptable compromise.

3. Deployment in Data Collection Activity

During the data-collection activity, participants view the AI-generated materials as elicitation prompts to discuss factors of interest. The visuals allow participants to externalise experiences and expectations, rather than acting as prototypes to be evaluated. Materials should therefore be explicitly framed as representations rather than perfected designs to avoid participants focusing excessively on visual fidelity instead of functional meaning. Transparency about the artificial origin of the images and videos is also important to maintain research integrity, and discussions on how AI might influence the future technologies under examination could therefore also be explored.

In our case studies, participants treated AI-generated prompts as reference points that accelerated scenario comprehension. Video vignettes supported immersion and temporal reasoning (e.g. anchoring comments to moments in the unfolding timeline) in Case Study 1, while storyboards supported structured elicitation in Case Study 2. Participants occasionally commented on ambiguities, such as minor visual inconsistencies, but these did not prevent substantial discussion, instead becoming prompts to surface assumptions and interpretations. To minimise distraction, we found it helpful to explicitly frame the visuals as representations of a situation (not a prototype to be judged on fidelity), and to be transparent that the materials were AI-generated.

4. Interpretation and Analysis

As with traditional participatory ergonomics studies, discussions should be transcribed and coded thematically but should also pay specific attention to how participants referenced or interpreted the visual elements. It is also important to consider how the AI-generated visuals, including any ambiguity or imperfections (whether inherent or deliberately tolerated), contributed to the discussions.

In our case studies, the visuals influenced not only what participants discussed but also how they reasoned about the scenario. Participants frequently pointed to specific frames or visual elements to negotiate interpretation and make counterfactual proposals. Thus, ambiguity and minor imperfections sometimes functioned as productive gaps that could be filled with participant expectations (e.g. from domain expertise in defence contexts or from driving experience). We treated references to artefacts as analytically meaningful during analysis.

Discussion

GenAI has evolved from early rule-based systems and simple neural networks into a transformative technology. This has been enabled by advances in deep learning, particularly using models based on vast data sets and large language models, such as Generative Pretrained Transformers (GPTs), that allow the generation of fluent, coherent text, realistic images and multimodal outputs including art, video and 3D designs. Now widely integrated across industries, GenAI supports automated content creation and enhances efficiency and creativity for writers, designers, developers and researchers. At the same time, its use raises significant concerns, including the production of biased or inaccurate information, risks of plagiarism and misinformation, potential job displacement, erosion of critical thinking and unresolved ethical issues surrounding data use and ownership. Consequently, the use of AI has emerged as a prominent, and often divisive, topic of discussion across multiple domains. This study contributes to the debate by examining the use of GenAI to create visual elicitation prompts for human factors research, highlighting both the benefits and challenges encountered.

Our case studies show that GenAI can be a valuable addition to the human factors research toolkit, offering practical advantages such as generating context-specific materials without specialist skills or additional funding, improving accessibility, and allowing text prompts to be shared verbatim to support methodological transparency. It is, however, noted that, although the prompting method itself is reproducible, AI-generated images and videos are not inherently reproducible, even when identical prompts are reused; this is because modern generative models rely on stochastic processes that introduce randomness into outputs. Consequently, identical prompts tend to produce results that are broadly similar but differ in specific details, a limitation observed in case study 2 where requests for minor image changes often resulted in the generation of entirely different scenes. Reproducibility is typically even lower for video generation due to increased temporal complexity. Although exact reproduction of images and videos is theoretically possible under tightly controlled conditions – requiring fixed model versions, training states, prompts, generation parameters, and random seeds – publicly available systems rarely preserve these variables, making exact replication impractical for end users. As a result, GenAI outputs are theoretically but functionally non-reproducible in most real-world settings. This could have important implications for scientific reliability, credibility, creative ownership and the use of GenAI in academic and legal contexts.

AI-generated visuals can, however, circumvent privacy and consent issues associated with photographing or filming human actors (as well as the inherent cost and logistical complications of doing so), and this also allows researchers to maintain control over the diversity and inclusivity of depicted personas and environments, aligning with ethical guidelines in human factors and ergonomics. The GenAI approach also extends traditional visual elicitation techniques by introducing real-time adaptability: AI-generated visuals can be modified rapidly and repeatedly, making it possible to regenerate scenes based on feedback during study piloting or even within a single study session with participants present.

It is worth re-iterating that prompt design is non-trivial, not least because the wording of AI prompts implicitly embeds the writer's assumptions about human behaviour, task design and socio-technical contexts. Generated materials should therefore be checked and edited to ensure that they are appropriate and align with the theoretical constructs under evaluation (e.g. workload, situational awareness, trust etc.) in an objective and unbiased manner – although, arguably, there are similar concerns associated with the creation of images and videos using established techniques, for example, with human actors.

An interesting point to note is that the perception of images and videos (i.e. whether they were AI-generated or human-made) can itself impact on realism and credibility (Velásquez-Salamanca et al., 2025). Humans may even struggle to distinguish real images or videos from AI-generated ones, particularly in situations where these are intended to deceive, such as 'fake news' (Lu et al., 2023), and this can cause confusion and diminish trust. It is therefore recommended that, in the context of human factors research, participants are briefed on the synthetic nature of the videos and images to contextualise their purpose as exploratory tools, and data is captured to document responses to the disclosure or any abnormalities. Furthermore, it is unclear how AI-generated elicitation prompts affect participant engagement compared with conventional materials, and future work should explore this, as well as examining cross-cultural differences in interpretation, to highlight any further dangers or pitfalls associated with the proposed method.

Finally, the use of GenAI raises concerns regarding sustainability – the United Nations Environment Programme (UNEP) (2024) estimated that a request made through ChatGPT consumes 10 times the electricity of a Google Search, and the water used by data centres – of which there are reportedly now more than 8 million according to UNEP (2024), compared to 500,000 in 2012 – is significantly more. While UNEP suggest that improving the efficiency of AI algorithms could reduce their demand for energy, and recycling water (used for both the construction of data centres and to cool electrical components) could reduce consumption, neither of these appears to be a substantive solution.

Conclusion

We present and reflect upon a novel methodological approach that uses GenAI to create visual elicitation materials for participatory ergonomics and human factors research. Through two contrasting case studies – one situated in a safety-critical military context and the other in everyday driving – we demonstrate how AI-generated images and videos can be used to externalise complex, futuristic, or otherwise inaccessible scenarios, and support rich discussion in interviews and focus groups. The four-stage methodology outlined herein provides a structured way to design, curate, deploy and analyse such materials while retaining human judgement and theoretical grounding. Across both use cases, GenAI offered clear practical benefits: rapid production of context-specific visual prompts, reduced cost and logistical burden, avoidance of privacy and safety risks, and flexibility to iteratively adapt materials as research needs evolved. Importantly, the visuals functioned not as prototypes to be evaluated, but as representational artefacts that supported sense-making, elicitation of tacit knowledge and negotiation of meaning among participants. Minor ambiguities and imperfections – whether inherent to the technology or deliberately tolerated – were often analytically productive rather than detrimental.

At the same time, the work highlights important limitations and responsibilities. Prompt design embeds assumptions and therefore requires reflexivity; outputs are functionally non-reproducible in most real-world settings; and ethical, epistemic and sustainability concerns accompany the use of GenAI. Transparency with participants about the synthetic nature of materials, careful validation and curation and explicit alignment with human factors constructs are therefore essential. GenAI should be understood as an enabling tool rather than a substitute for established participatory methods, human creativity or critical analysis.

Overall, we argue that GenAI can meaningfully extend the repertoire of visual elicitation techniques in human factors research when used deliberately and critically. Future work should empirically examine how AI-generated prompts influence participant engagement, interpretation and trust compared with conventional materials, explore cross-cultural effects and further interrogate the environmental and ethical trade-offs involved. Used with care, GenAI has the potential to support more accessible, adaptive and imaginative participatory ergonomics, while keeping humans firmly in the loop.

References

Ferraro, J. C., & Mouloua, M. (2022). Assessing How Driving Self-Efficacy Influences Situational Trust in Driver Assist Technologies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1260–1264. https://doi.org/10.1177/1071181322661476

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X. and Ouyang, W., 2023. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in neural information processing systems*, *36*, pp.25435-25447.

Marine Accident Investigation Branch (MAIB), 2016. *Report on the investigation of the collision between the stern trawler Karen (B317) and a dived Royal Navy submarine in the Irish Sea on 15 April 2015* [online] Available at: https://assets.publishing.service.gov.uk/media/57fe2e2ee5274a496200000a/MAIBInvReport20_2016.pdf

Neubauer, C., Forster, D. E., Blackman, J., Lakhmani, S., Fitzhugh, S. M., Krausman, A., Schaefer, K. E., Berg, S., & Rovira, E. (2021). Developing a new human-autonomy team cohesion Scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 801–806. https://doi.org/10.1177/1071181321651324

Nørskov, S., Damholdt, M. F., Ulhøi, J. P., Jensen, M. B., Ess, C., & Seibt, J. (2020). Applicant fairness perceptions of a robot-mediated job interview: A video vignette-based experimental survey. *Frontiers in Robotics and AI*, *7*, 586263.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Parasuraman, R., Barnes, M., Cosenzo, K., & Mulgund, S. (2007). *Adaptive automation for human-robot teaming in future command and control systems*. https://apps.dtic.mil/sti/html/tr/ADA503770/

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286–297.

Truong, K. N., Hayes, G. R., & Abowd, G. D. (2006). Storyboarding: An empirical determination of best practices and effective guidelines. *Proceedings of the 6th Conference on Designing Interactive Systems*, 12–21. https://doi.org/10.1145/1142405.1142410

UNEP (2024). AI Has an Environmental problem. Here's What the World Can Do about that. [online] UNEP. Available at: https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about.

Velásquez-Salamanca, D., Martín-Pascual, M.Á. and Andreu-Sánchez, C., 2025. Interpretation of AI-generated vs. human-made images. *Journal of Imaging*, *11*(7), p.227.