

# The Epistemology of AI-Generated Images

## Context

Images have a central role in human knowledge practices, including science, journalism, and everyday communication. Photographs have served as epistemic artifacts that represent the world and are used as evidence for claims and beliefs. Here, “epistemic” refers to the quality of relating to knowledge; specifically, how beliefs are formed, justified, and judged to be true or false. Historically, photographs have been understood to have a unique value as carriers of knowledge. The specific epistemic value of photographs has been defended in virtue of their objects being necessarily real (Walton, 1984), their independence from mental states (Currie, 2004), and their basis in perception rather than testimony (Cavendon-Taylor, 2013). The information conferred by photographs has generally been taken to be trustworthy unless there are specific reasons for doubt. AI-generated images differ fundamentally from photographs. Diffusion models, generative adversarial networks, and related systems do not capture particular events; they generate outputs through statistical processes operating on large datasets. This does not align with existing interpretations of how we gain knowledge from images. Generative AI models and cameras are distinct in their mechanisms and operators, and they reflect completely different relationships between the technology and the objects that technology depicts.

Although it remains limited, work has been published in recent years that directly addresses the epistemology of AI-generated images (Scotti, 2024) (Zylinska, 2024). Sahebi and Formosa (2025) argue that AI-generated media poses a challenge to online communities because it forces a choice between maintaining normal levels of trust in images and risking vulnerability to false beliefs, or reducing general trust in images and lowering the ability of viewers to rely on them as sources of knowledge. Following from their assessment, a question arises of how digital interface design can support a balanced level of trust between the extremes of gullibility and excessive skepticism in the face of uncertainty about images.

The pressing importance of investigating this topic is revealed by research on false belief formation and the challenge of identifying AI-generated media. Lu et al. (2023) found that participants failed to distinguish real and AI-generated images 38.7% of the time. In Velásquez-Salamanca et al. (2025), about 40% of AI-generated images shown to participants were misidentified as real photographs. The increasing evidence that people are unable to reliably tell apart real and AI-generated images represents an urgent need for more work in epistemology, AI ethics, and HCI that addresses this issue.

## Objectives

The project will address the questions: *How do the technical properties of generative image models constrain philosophical claims about how their outputs can function as sources of knowledge? And what design guidelines should follow from a better understanding of the changing epistemic role of images caused by the proliferation of generative AI?* The project will incorporate empirical knowledge on the design and operation of GANs, diffusion models, and other computational models that can produce and alter images (Yazdani et al., 2025). A point of focus will be the probabilistic relationships between datasets, models, and their outputs, and what these relationships mean for the transformation of knowledge. By clarifying the differences between how photographs and AI-generated images constitute knowledge, this project aims to contribute to the fields of HCI, philosophy of AI, and epistemology. Specifically, the project aims to:

- Analyze the specific role of photographs as knowledge and contrast it with the possible role of AI-generated images as knowledge.
- Map specific epistemic properties of AI-generated images onto specific technical features, such as training data and system architectures.
- Investigate how users form and justify beliefs based on AI-generated images, focusing on trust and interpretative strategies in digital environments.
- Assess how and whether increased consumption of AI-generated images lowers global trust in images.
- Formulate guidelines for interface-level mechanisms that could assist online users and communities in assessing image origins and calibrating appropriate levels of trust versus skepticism in the visual media to which they are exposed.

## Approach

The project combines philosophical analysis with empirical research in HCI. The first element of the approach will involve developing a philosophical framework of photographic knowledge, drawing on philosophy of photography and social epistemology. This framework will identify key epistemic features attributed to photographs, such as causal connection to depicted events and traceability of production processes. These features will then be systematically compared with the technical structure of generative image models. To explain how the properties of these models constrain the kinds of epistemic claims that can be made about their

outputs, we will draw on previous work by Martinez et al. (2025) concerning the impact of generative AI tools on the representational agency of people being photographed and taking photographs. A second element will be an investigation into how users interpret and evaluate AI-generated images. Possible specific methodologies include controlled online surveys (in the style of experimental philosophy) and qualitative semi-structured interviews to investigate users' experiences of trust and belief in images. Rather than focusing solely on detection accuracy, as in other studies on perception of AI-generated images, the emphasis will be on belief formation and knowledge production. Examples of topics that may be addressed include how users justify their trust or distrust in images and where they set thresholds for belief under uncertainty. The final phase will translate these findings into practical design insights (e.g. transparency mechanisms that are visible and comprehensible to non-expert users) that serve the goal of developing more trustworthy and responsible AI systems.

### **Skill Requirements**

The candidate will have a background in philosophy and cognitive science, with specific knowledge in epistemology, ethics, and AI.

### **Suitability for SCAI**

This doctoral project aligns with SCAI's mission to promote interdisciplinary research that bridges AI research with the humanities and social sciences. The project responds directly to SCAI's research strategy in two key ways. First, it addresses a core theme in digital humanities by examining how AI technology is reshaping a fundamental aspect of human culture and knowledge: our relationship with images. Second, it engages deeply with the field of human-machine interaction, as the project's conclusions will have direct implications for the design of trustworthy AI systems. This project would make a significant contribution to SCAI's goal of supporting interdisciplinary, technically advanced, and socially responsible AI research.

### **Supervision**

The project will be co-supervised by experts in HCI and philosophy to guarantee that the project is both philosophically rigorous and grounded in an empirical understanding of AI. The first co-supervisor will be *Dr. Baptiste Caramiaux*, a CNRS researcher who leads the HCI Sorbonne group at ISIR (L'Institut des Systèmes Intelligents et de Robotique). *Dr. Caramiaux* has expertise on human-AI trust and the political and social implications of generative AI usage in the cultural sector. His work on the societal impacts of generative models would be highly useful in developing the project, and his experience with design-oriented research would serve the empirical components of the proposal. The second co-supervisor will be *Dr. Anouk Barberousse*, a CNRS researcher in philosophy at the SND (Sciences Normes Démocratie) group at Sorbonne Université. She holds expertise in epistemology, the philosophy of science, and specifically, the epistemic value of computational modeling. Her extensive work on modeling also includes questions about the epistemological status of non-linguistic elements, including images. She will guide the philosophical side of the project, ensuring engagement with the relevant concepts and literature.

### **References**

- Cavedon-Taylor, D. (2013). Photographically based knowledge. *Episteme*, 10(3), 283–297.
- Currie, G. (2004). *Arts and Minds*. Oxford University Press.
- Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*.
- Martinez, L., Caramiaux, B., & Alaoui, S. F. (2025). *Generative AI in documentary photography: Exploring opportunities and challenges for visual storytelling*. CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan.
- O'Connor, C., Goldberg, S., & Goldman, A. (2024). Social epistemology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition). Metaphysics Research Lab, Stanford University.
- Sahebi, S., & Formosa, P. (2025). The AI-mediated communication dilemma: epistemic trust, social media, and the challenge of generative artificial intelligence. *Synthese*, 205(3).
- Scotti, A. (2024). What to do with AI images? Towards an epistemology of trust. *I Castelli Di Yale*, 12, 43–54.
- Velásquez-Salamanca, D., Martín-Pascual, M. Á., & Andreu-Sánchez, C. (2025). Interpretation of AI-Generated vs. Human-Made Images. *Journal of Imaging*, 11(7).
- Walton, K. L. (1984). Transparent pictures: On the nature of photographic realism. *Critical Inquiry*, 11.
- Yazdani, S., Singh, A., Saxena, N., Wang, Z., Pan, D., Pal, U., Yang, J., Zhang, W., & Palikhe, A. (2025). Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, 12(1).
- Zylinska, J. (2024). Diffused seeing: The epistemological challenge of generative AI. *Media Theory*, 8.