FEATURE

Reflections on Explainable AI for the Arts (XAIxArts)

Nick Bryan-Kinns, University of the Arts London

Insights

- → XAIxArts offers fresh viewpoints and challenges on what makes an AI explainable.
- → Time plays an important role in XAIxArts, from real-time explanations to creative practice with AI that spans years.
- → XAI helps artists identify and re-create unexpected AI errors.
- → Explanation of an AI artwork may reduce its meaning.

It is difficult—dare I say impossible—for us to understand how a deep-learning model works and how it produces humanlike outputs, especially when it generates outputs that appear to be creative and artistic. This is troubling for us as human-computer interaction researchers seeking to make computers less perplexing and more intuitive to use. In recent years, the fields of explainable AI (XAI) [1] and human-centered AI (HCAI) [2] more broadly have started to explore how the decisions complex AI models make can be explained or made more transparent to humans. For example, there are an increasing number of papers, sessions, and workshops about AI and, more specifically, XAI at HCI conferences such as ACM CHI.

Current XAI research mostly examines functional or technical explanations of what an AI is doing, for example, providing an explanation of how an image classifier works to help debug it when misclassifications are made. In these settings, there is typically a right answer, or correct set of outputs, that we are trying to train the AI to arrive at. In the arts there are no right or wrong answers, no correct set of outputs. In the arts we are often interested in outcomes that are surprising or unusual, or even disturbing and disruptive. Furthermore, in creative practice the focus is usually on the output itself rather than detailed explanations of how it was produced. What then does it mean to explain AI models in an artistic context? How are such explanations different from more functional explanations and context? And what insights might exploring these questions provide for XAI research more broadly? To begin exploring these questions, I brought together an international team of researchers to host the first international workshop on explainable AI for the arts (XAIxArts) at the 2023 ACM Creativity & Cognition conference [3]. In the following discussion, I'll reflect on the key themes that emerged in the workshop, what we learned about XAI and the arts, and how that might relate to XAI more broadly.

Our workshop kicked off in traditional style with introductions from the organizers about their research and XAIxArts focus. We then jumped into short position paper presentations by participants, which you can see at the workshop website (xaixarts.github.io). These sessions were interleaved with discussions and brainstorming about the nature of XAIxArts. Throughout the workshop, a spirited debate unfolded about XAIxArts as a way to examine current XAI and AI research and what is needed for the two to contribute to creative practice. Perhaps XAIxArts could be seen as an approach to exploring forms and views of explanation not considered in current XAI research.

REFLECTING ON THE NATURE OF XAIXARTS

Reflecting on our workshop debates and brainstorming sessions, we identified emergent themes to help frame XAIxArts. These built on our initial XAIxArts themes and features [4]: the nature of explanation for the arts and how it might be different from more technical explanations of AI models; thinking about the appropriateness of AI models and training sets for use in the arts; how to increase the user-centered design of XAI; and what features of XAI interaction design might be similar or different for XAIxArts. Below I outline the additional themes we identified in the workshop. The themes are based on my reflections and are reported here as a way to provoke further debate and discussion.

Temporal aspects of explanations. The feature of XAIxArts models that really struck me in the workshop was the temporal nature of XAIxArts, especially the liveness of XAI in many creative settings, from music-making to sketching and performance, where moment-by-moment feedback is essential for the creative process. For example, Gabriel Vigliensoni and Rebecca Fiebrink presented work that explored real-time AI audio generation in live music performance, requiring immediate audio generation and feedback as they navigated an AI model's latent space. This raises questions about the explainability of the latent space itself, as each latent space needed to be mapped by the musician to a human-performance space, essentially composing an explanatory map of latent space, which could then be navigated in real time. These real-time, in-the-moment explanations contrast with the post hoc explanations often found in XAI research. The role of time in the explanation itself is also important. For example, there are open questions about how an explanation might help us roll back to previous elements or stages of a generative art process. In other words, XAI might help us go back in time in our creative process. It is important to recognize that creative practice may take place over seconds, minutes, days, weeks, or even years, leaving open questions about what the role and form of explanations might be over such radically different time spans. This final point is particularly interesting, as it contrasts with the typical single-shot use of XAI in more functional domains, where an explanation might be used to understand an AI's decision or to help identify omissions in training, and is then discarded.

Creative practice may take place over seconds, minutes, days, weeks, or even years, leaving open questions about what the role and form of explanations might be over such radically different time spans.

Tailored explanations. There is growing recognition in XAI that explanations need to be tailored to individual users based on their skills and background. Michael Clemens proposed tailoring explanations to individual AI literacy and artistic domain expertise, while Marianne Bossema and colleagues tailored cocreative AI explanations to the needs and abilities of individual elderly people. We also need to consider the audience for the art produced with AI and what their values and explanation needs are. In the arts, we would need to consider different forms of explanation based on context and value of the piece, in addition to individual skills and background currently considered by XAI. Moreover, as highlighted by Nicola Privato and Jack Armitage, there can be no single appropriate explanation—explanations are necessarily relative to a specific audience or group. Indeed, we need to consider the form of explanation that would be suited to different sizes of groups and their role in the artistic experience, from audience members to ensemble performers and individual performers. Cheshta Arora and Debarun Sarkar critiqued the potential for explanations to reduce the sense-making value of AI art and emphasized the need in the arts to balance explanation with artistic intent and the necessary effort of sense-making on the part of the audience. In other words, thorough explanation of an AI artwork may in effect reduce the value and (constructed) meaning of the piece.

AI as material. In AI arts the AI model itself may be an artwork. In XAIxArts, traversing a generative model may be both an artwork and a form of explanation of the model. For example, Luís Arandas presented work with colleagues navigating a generative image model to generate

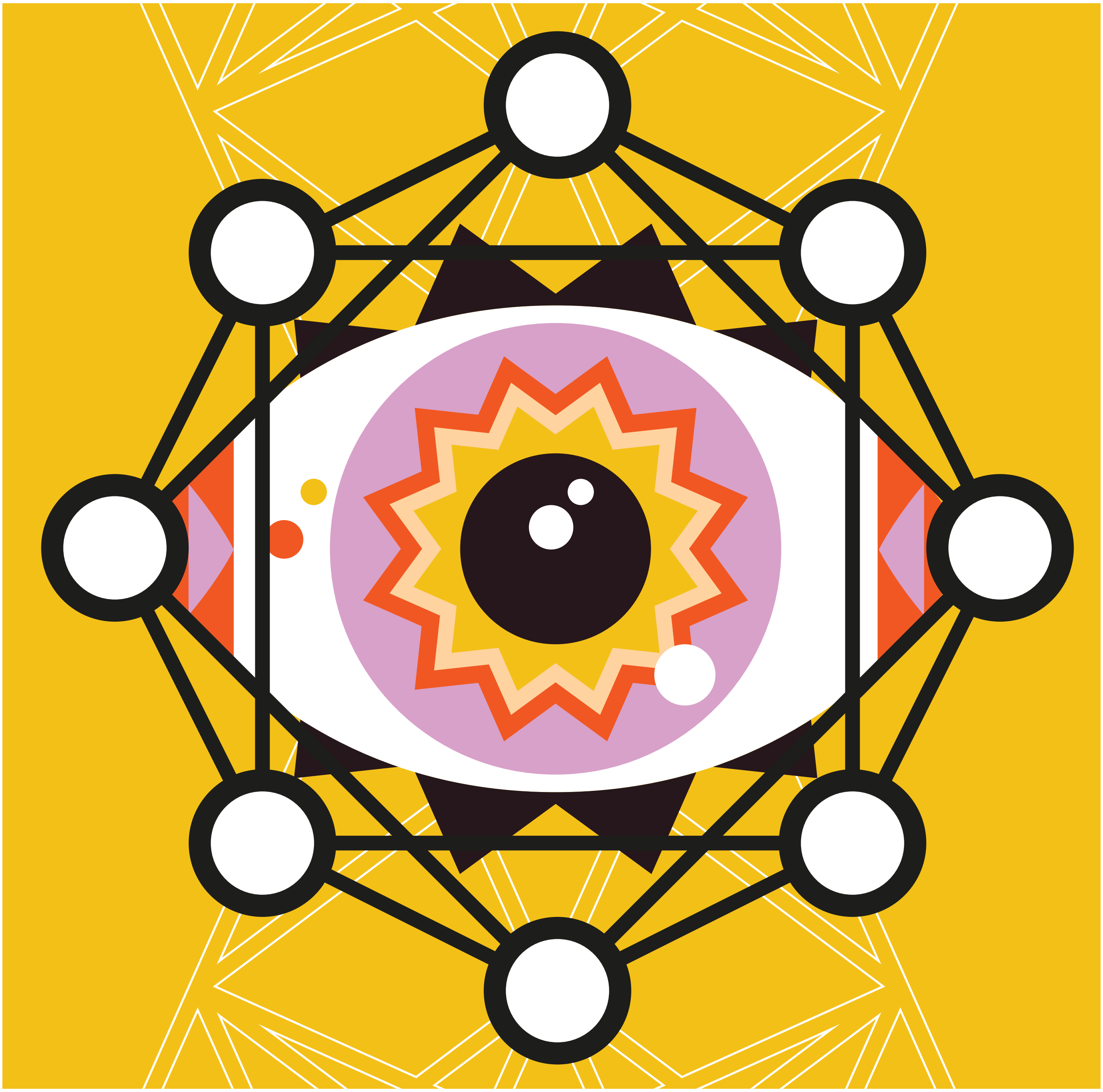


IMAGE BY RIO RITA / SHUTTERSTOCK.COM

filmic output to expose the bias of the model's training. Similarly, Ashley Noel-Hirst and I navigated a generative music model to expose and explain the shape and limits of the latent space of the model itself. In these ways, the navigation of an AI model through creative practice offers artist-led rather than model-led explanations of features of the model such as training set bias.

In these cases, we might ask who has authorial intent of the AI output and the explanation. If the explanation of a model is artist led, then we must question where the creative agency resides—is it with the model that generates the images or the artist who navigates the space? Moreover, if the explanation is an artistic piece that relies on sense-making to interpret the artistic intent, then who is responsible for the explanation: the AI model, the artist, or the audience?

The artist-led navigation and exploration of AI models to produce artistic output also potentially offers ways to explain AI that are more open to engagement by a wider demographic. For example, Drew Hemment and Dave Murray-Rust presented work with colleagues on explorations of how artists could define latent space dimensions for generative art models. These could then be explored and navigated by audiences in public exhibitions as a form of sense-making about the models and their workings. In this way, creative methods from the arts augment XAI approaches and offer practice-based methods for engaging people with XAI. This contrasts with design-led explanations of AI models, as illustrated by Lanxi Xiao and colleagues, who created interactive artworks to explain image misclassification in museum and gallery settings to engage the public in explanations of AI.

XAI in generative tools. Hanjie Yu and colleagues noted that current AI generative content (AIGC) tools expect users (artists) to be able to specify their requirements concretely. This does not fit well with a typical iterative, trial-and-error creative process where alternatives are routinely applied, modified, and removed from artistic efforts. From an XAI perspective, the question here is not how to better explain the AIGC generative process, but rather how to use XAI to reduce barriers to trial and error given that it is inherent in many artistic practices. Supporting trial and error also asks how to use XAI to help identify differences between an artist's expectation and an AIGC's output to help iteratively bring the two closer together. We might explore XAI approaches such as providing contrastive examples or offering more conversational interaction. In this view, we are thinking of how to use XAI to make the relationship between AI and artist more transparent and balanced, as noted by practicing artist Makayla Lewis. We could go further and explore how the experiences of artists in artist-AI collaboration could be explained and explored by the artists themselves as a form of reflective practice, and by others as a window into the creative practice behind artworks.

XAI is useful for identifying errors in AI models, typically leading to debugging and revision of the model to improve its performance. In the arts, Jamal Knight and colleagues highlighted that in addition to identifying errors in AI models to remove them, we might also want to know how to re-create and reuse those errors in creative practice. In Jamal's case, that would be identifying glitches in motion tracking for performance arts. This is an intriguing differentiator between XAI and XAIxArts; instead of thinking of XAI as a way to identify errors in AI generation and then correct or mitigate for them, we could consider ways for XAI to help us re-create those glitches (somewhat) reliably. Moreover, the kinds of explanations we are seeking are different—I mean, "Why is this glitching?" is quite a different question from "Why did you make that decision?" Such a perspective also questions how to balance the surprise of the error and the explanation of how to re-create it—again, emphasizing that surprising and unexpected outputs are often welcome in XAIxArts.

Generative AI models can consume huge amounts of energy for both their training and use. Petra Jääskeläinen argued that XAI could be used to expose and explain to creative practitioners the environmental impact of generative AI art models. In this way, environmental impact might become a design constraint in AIGC tools. For example, artists and designers might use low-energy generation and explanations for exploration, and high-energy consumption for high-quality final production.

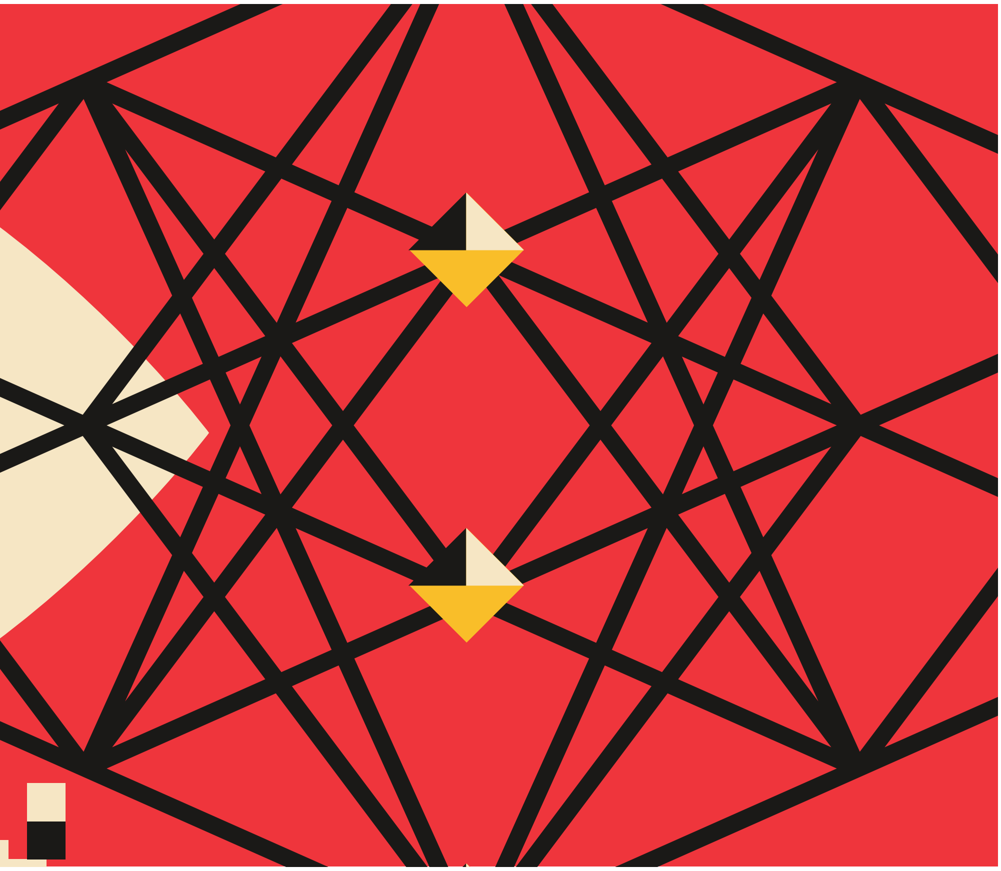
Ethics and responsibility. A key concern in the workshop was how XAI might support and also disrupt trust in AI arts. Makayla highlighted the need for transparency of attribution in generative AI models to build trust with artists and the proactive role that XAI could take in explaining whose creative content was used to generate new outputs. As mentioned earlier, XAI could also be used to expose bias in datasets and models through artistic practice, building trust through sense-making. Finally, Cheshta and Debarun highlighted the ethical concerns of using XAI to value and revalue artworks for different audiences, and the ethical challenge of explanations potentially devaluing artworks if they then can't be valued in themselves without the XAI.

The arts provide a refreshing and insightful counterpoint to more functional explanations of AI.

TAKEAWAYS

Exploring XAI for the arts offers us as HCI researchers fresh viewpoints and challenges on what makes an AI

IMAGE BY NEO GEOMETRIC / SHUTTERSTOCK.COM



explainable. Indeed, the arts provide a refreshing and insightful counterpoint to more functional explanations of AI. A key takeaway for me from our workshop was that time is an often overlooked aspect in XAI, whether it is the need for real-time in-the-moment explanations or explanations that resonate through years of creative practice. Tailoring explanations beyond the current XAI recognition of individual users to consider audiences, artists, and ensemble practitioners also stood out to me alongside questions of what values and context our explanations might need to respond to beyond work-oriented XAI concerns. Another key aspect the workshop stressed was the use of XAI to reveal the ethical implications of our AI systems—from bias to energy consumption—to be balanced with the thorny challenge of explanations potentially devaluing the very artwork that they attempt to explain. Here the role of sense-making in XAIxArts challenges current pragmatic approaches to XAI explanation. In terms of producing AI arts, the use of XAI to help understand and re-create glitches was another key takeaway, along with the need for generative tools to integrate better into a trial-and-error iterative creative process. Finally, the role of AI and XAI as an artistic material and the use of arts practice as a way of explaining AI through practice-based exploration and navigation really stood out to me as a key differentiator for XAIxArts compared to XAI.

Unfortunately, we encountered the age-old problem of terminological ambiguity: The term *explanation* is used differently by different groups of XAI researchers. For the machine learning (ML) research community, explanation is often narrowly defined as explaining the reasons behind ML decisions such as image classifications or predictions. For other researchers such as Q. Vera Liao and colleagues [5], explainability is more broadly understood as encompassing "everything that makes ML models transparent and understandable, also including information about the data, performance, etc." Personally, I follow Liao's broader definition, and given the discussion in the workshop on the nature of explanation in the arts, I suggest that this view is more suited to XAIxArts.

One final thought to close with: For me there are parallels between the shift from XAI to XAIxArts and the shift from second-wave HCI to third-wave HCI [6]. There is a similar shift from a work-oriented focus of tasks, goals, and cognition to embracing "experience and meaning-making" [6] in understanding our interaction with computers. This is an exciting parallel for XAIxArts, offering a vision of how an artistic approach to XAI could become a new wave or even a new paradigm of AI research.

ACKNOWLEDGMENTS

With many thanks to the workshop participants, authors, and the organizing team including: Corey Ford, Alan Chamberlain, Steven David Benford, Helen Kennedy, Zijin Li, Wu Qiong, Gus G. Xia, and Jeba Rezwana. Thanks also to Corey Ford for his comments on this text.

Endnotes

1. Gunning, G. *Explainable Artificial Intelligence (XAI).* DARPA/I2O Proposers Day (Aug. 2016).
2. Shneiderman, B. *Human-Centered AI.* Oxford Univ. Press, 2022.
3. Bryan-Kinns, N. et al. Explainable AI for the arts: XAIxArts. *Proc. of the 15th Conference on Creativity & Cognition.* ACM, New York, 2023, 1–7; https://doi.org/10.1145/3591196.3593517
4. Bryan-Kinns, N. et al. Exploring XAI for the arts: Explaining latent space in generative music. *Proc. of the 1st Workshop on eXplainable AI Approaches for Debugging and Diagnosis.* 2021.
5. Liao, Q.V., Gruen, D., and Miller, S. Questioning the AI: Informing design practices for explainable AI user experiences. *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, New York, 2020, 1–15.
6. Bødker, S. Third-wave HCI, 10 years later—participation and sharing. *Interactions 22*, 5 (Sep.–Oct. 2015), 24–31.

Nick Bryan-Kinns is a professor of creative computing at the University of the Arts London. He is a fellow of the Royal Society of Arts and the British Computer Society and a senior member of the Association for Computing Machinery. He researches explainable AI, interaction design, mutual engagement, interactive art, and cross-cultural design.
→ nickbk@acm.org