A Banal Account of a Safety-Creativity Tradeoff in Generative AI

KUSH R. VARSHNEY, IBM Research – Thomas J. Watson Research Center, USA
LAV R. VARSHNEY, University of Illinois Urbana-Champaign, USA

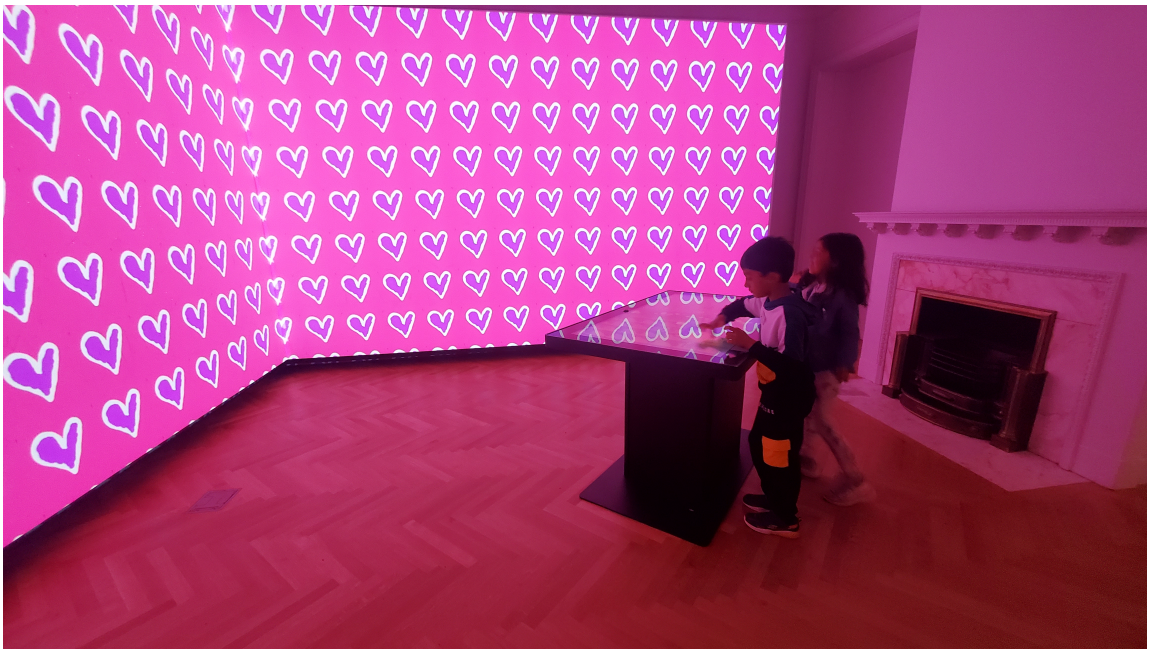

Fig. 1. Children creating wallpaper designs in the Immersion Room of the Cooper Hewitt, Smithsonian Design Museum.

Safety is banal.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: computational creativity, generative model, safety, information geometry

1 INTRODUCTION

DALL-E 2, Stable Diffusion, Midjourney, GPT-3, ChatGPT, YouChat and other generative artificial intelligence (AI) models may be used in a variety of tasks, some mundane and some creative. Their safety may be of concern.

2 SAFETY

Safety is defined in terms of harm, aleatoric uncertainty, and epistemic uncertainty [3]. Safe AI systems constrain the probability of expected harms and the possibility of unexpected harms [6]. Harms from generative AI may be representational, allocative, quality-of-service, interpersonal, or societal [5].

3 CREATIVITY

Creativity is the generation of an artifact that is high-quality and novel [9]. Quality metrics are specific to the application. Novelty is a more application-agnostic concept that may be measured using Bayesian surprise, the relative entropy

between the empirical distribution of an inspiration set and that set updated with the new artifact [2]. An inspiration set is a collection of previous artifacts in the creative domain.

Creativity by modern generative AI is implicitly or explicitly combinatorial. It generates unfamiliar combinations of familiar ideas [1]. Combinatorial creativity has precise information-theoretic limits on the tradeoff between quality and novelty [7]. On average, higher quality implies lower novelty and vice versa.

The more immature a creative domain is, the smaller the size of the inspiration set is. Creativity is easier because many concepts are unexplored. The feasible region bounded by the quality-novelty tradeoff curve is larger.

When creative artifacts are constrained, for example by requiring intentionality, the region becomes smaller and creativity becomes more difficult [8]. (This statistical phenomenon of optimal creativity systems contrasts the computational phenomenon of humans often being more creative with more constraints [4].)

4 SAFETY AND CREATIVITY

Safety is a constraint on artifacts. Like other constraints, safety makes the feasible region under the quality-novelty tradeoff curve smaller and creativity more difficult. Thus, banality, the lack of creativity, follows from safety. There is a tradeoff between safety and creativity.

5 IMPLICATIONS

Some applications of generative AI, like autonomously writing boilerplate, require safety whereas others, like inspiring a human poet, do not. Some applications of generative AI, like writing poetry, require creativity and others, like writing boilerplate do not. Applications requiring safety tend to also be ones not requiring creativity. Applications not requiring safety tend to also be ones requiring creativity.

6 CONCLUSION

Information theory tells us that most natural applications of combinatorial creativity with modern generative AI are feasible in terms of the safety-creativity tradeoff. Future work requires constructive algorithms for placing safety constraints on generative AI. The end.

REFERENCES

[1] Payel Das and Lav R. Varshney. 2022. Explaining Artificial Intelligence Generation and Creativity. *IEEE Signal Processing Magazine* 39, 4 (2022), 85–95.

[2] Laurent Itti and Pierre Baldi. 2005. Bayesian Surprise Attracts Human Attention. In *Advances in Neural Information Processing Systems*.

[3] Niklas Möller. 2012. The Concepts of Risk and Safety. In *Handbook of Risk Theory*. Springer, Dordrecht, Netherlands, 55–85.

[4] R. Keith Sawyer. 2012. *Explaining Creativity: The Science of Human Innovation*. Oxford University Press, New York, NY, USA.

[5] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791.

[6] Kush R. Varshney and Homa Alemzadeh. 2017. On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. *Big Data* 5, 3 (2017), 246–255.

[7] Lav R. Varshney. 2019. Mathematical Limit Theorems for Computational Creativity. *IBM Journal of Research and Development* 63, 1 (2019), 2.

[8] Lav R. Varshney. 2020. Limits Theorems for Creativity with Intentionality. In *Proceedings of the International Conference on Computational Creativity*. 390–393.

[9] Lav R. Varshney, Florian Pinel, Kush R. Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Yi-Min Chee. 2019. A Big Data Approach to Computational Creativity: The Curious Case of Chef Watson. *IBM Journal of Research and Development* 63, 1 (2019), 7.

Received 10 January 2023