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

Kush R. Varshney[1], Lav R. Varshney[2]

[1]IBM Research – Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York, USA 10598
[2]University of Illinois Urbana-Champaign, 1308 West Main Street, Urbana, Illinois, USA 61801

Abstract
Safety is banal.

Keywords
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 [1]. Safe AI systems constrain the probability of expected harms and the possibility of unexpected harms [2]. Harms from generative AI may be representational, allocative, quality-of-service, interpersonal, or societal [3].

3. Creativity

Creativity is the generation of an artifact that is high-quality and novel [4]. 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 [5]. 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 [6]. 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 [9].)

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

Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia
krvarshn@us.ibm.com (K. R. Varshney); varshney@illinois.edu (L. R. Varshney)
https://krvarshney.github.io (K. R. Varshney); http://www.varshney.csl.illinois.edu (L. R. Varshney)
0000-0002-7376-5536 (K. R. Varshney); 0000-0003-2798-5308 (L. R. Varshney)

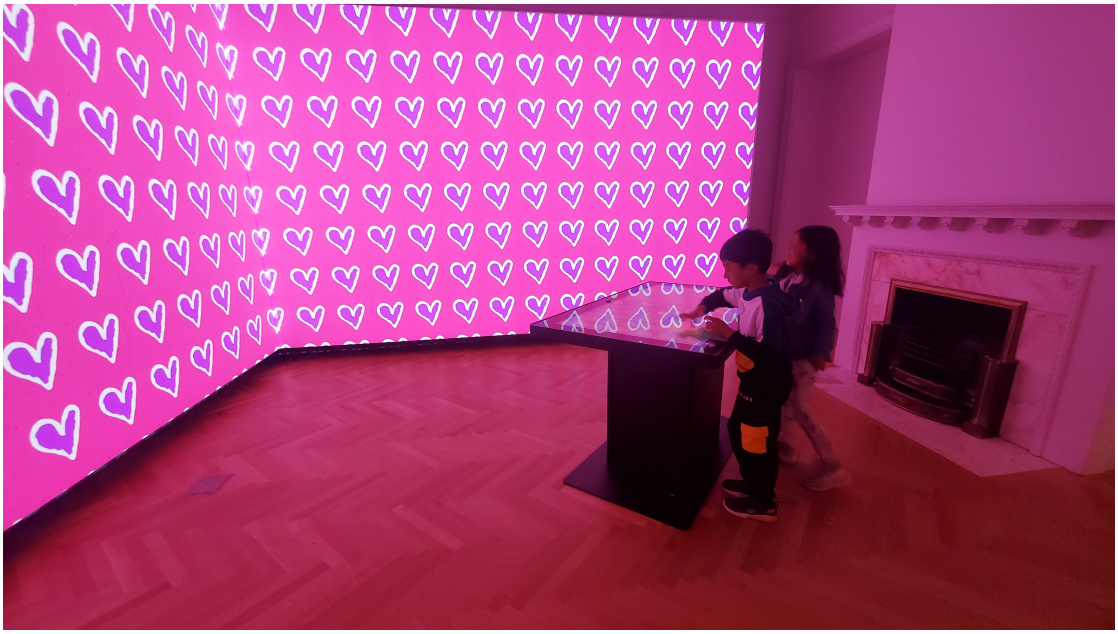

Figure 1: Children creating wallpaper designs in the Immersion Room of the Cooper Hewitt, Smithsonian Design Museum.

work requires constructive algorithms for placing safety constraints on generative AI. The end.

References

[1] N. Möller, The concepts of risk and safety, in: Handbook of Risk Theory, Springer, Dordrecht, Netherlands, 2012, pp. 55–85.

[2] K. R. Varshney, H. Alemzadeh, On the safety of machine learning: Cyber-physical systems, decision sciences, and data products, Big Data 5 (2017) 246–255.

[3] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, G. Virk, Sociotechnical harms: Scoping a taxonomy for harm reduction, arXiv:2210.05791, 2022.

[4] L. R. Varshney, F. Pinel, K. R. Varshney, D. Bhattacharjya, A. Schörgendorfer, Y.-M. Chee, A big data approach to computational creativity: The curious case of Chef Watson, IBM Journal of Research and Development 63 (2019) 7.

[5] L. Itti, P. Baldi, Bayesian surprise attracts human attention, in: Advances in Neural Information Processing Systems, 2005.

[6] P. Das, L. R. Varshney, Explaining artificial intelligence generation and creativity, IEEE Signal Processing Magazine 39 (2022) 85–95.

[7] L. R. Varshney, Mathematical limit theorems for computational creativity, IBM Journal of Research and Development 63 (2019) 2.

[8] L. R. Varshney, Limits theorems for creativity with intentionality, in: Proceedings of the International Conference on Computational Creativity, 2020, pp. 390–393.

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