Axes of Characterizing Generative Systems: A Taxonomy of Approaches to Expressive Range Analysis

Josiah Boucher[1,*], Gillian Smith[1,†] and Yunus Doğan Telliel[1,†]

[1] Worcester Polytechnic Institute (WPI)), 100 Institute Road, Worcester MA, U.S.A.

Abstract

Seeking to leverage Expressive Range Analysis (ERA)- a method of characterizing generative systems- for analysis of generative AI (GAI) systems and their outputs, this paper categorizes the approaches that have been used for ERA. We present a taxonomy with five axes of characterization that may be applied to ERA methodologies: content agnostic vs semantic; quantitative vs qualitative; product vs process; objective vs subjective; and automated vs manual. While ERA has traditionally been limited to the domain of Procedural Content Generation (PCG) in video game development, we recognize parallels between PCG and GAI and hope to see an expansion of the application of ERA into the domain of GAI. Serving this goal, these axes provide metrics through which to categorize, compare, and explore approaches.

Keywords

Expressive Range Analysis, Procedural Content Generation, Generative AI, Evaluating Generative Systems, Taxonomy

1. Introduction

Expressive Range Analysis (ERA) is a method of characterizing the nature and shape of a generative model in terms of its outputs. What sorts of outputs is a generator capable of? What impact does changing inputs of a generator have on its outputs? How do biases manifest in the generative outputs, and how do the inputs influence these biases? ERA is well suited for answering these questions [1]. ERA comes from the domain of Procedural Content Generation (PCG) in video game development: the practice of leveraging algorithmic methods to design and produce artifacts for use in games across a variety of contexts, such as levels or maps [2].

We recognize common threads between PCG and generative AI (GAI) systems- defined here as the process of using generative technologies such as large language models to produce content, often using natural language prompts as human-provided input- including high-level functionality, motivations, use cases, and shortcomings of both domains (for details, see Section 2). Because of these similarities, we seek to expand and leverage ERA—which has not only proven useful for recognizing bias within, and categorizing generative outputs of PCG systems, but also resembles some existing approaches for GAI analysis [3]—for analysis of GAI system outputs, especially text-to-X, large language model (LLM)-based applications (such as the text-to-text Chat GPT [4], text-to-image Stable Diffusion [5], and text-to-speech ElevenLabs [6]).

The application of ERA in the domain of GAI is not a one-to-one translation from how the method is used in PCG. Use of GAI tools commonly requires more complicated, varied, and linguistic inputs than PCG, which tends to operate from numerical randomization as a starting point for much of its generation. Because these inputs have a major impact on the outputs of GAI systems [7], selection of these inputs is an important consideration for applying ERA to GAI—e.g. some decision points for input selection could include whether to handcraft inputs or sample the latent space, and whether the inputs should be restricted in a way that targets a specific domain of outputs (i.e. a prompt to output a haiku would be very different from one that might produce a cooking recipe). Furthermore, ERA requires the application of metrics to large quantities of content outputted from the system in question. What metrics are used and how they are determined is a major component of ERA, often directly determining the value provided by the analysis. Because GAI outputs tend to occupy a broad domain of applications and mediums, unique challenges arise when considering these metrics, further complicating the application of ERA in this context.

Responding to 1) these parallels between PCG and GAI, and 2) the massive increase of scope presented by GAI, this paper presents a taxonomy for categorizing approaches to ERA. This taxonomy includes vocabulary to better distinguish examples from existing work in PCG, as well as a compass to guide further exploration of using this method in the rapidly expanding domain of GAI. The goal of this taxonomy is to push the boundaries of what ERA may be used for and how it may be applied.

2. Related Work

This work operates in the intersection of procedural content generation and generative AI. Guzdial provides a valuable bridge between these domains with the lens of human-AI interactive generation [8]. While Guzdial uses this lens to frame PCG and GAI as essentially the same process, we view PCG as a broader term describing the process of content generation at its highest level, and GAI as a more descriptive term identifying a subset of generative practices that use specific technologies. Framing GAI as a subset of PCG—or using Guzdial's lens of human-AI interactive generation [8]—broadens the scope of understanding for both domains and allows the application of evaluation methods of PCG for analysis of GAI systems.

Particularly, we are interested in leveraging expressive range analysis for evaluation of GAI systems [1]. Withington et al. list expressive range analysis as one of 12 "features" used for comparison of PCG systems [9]. We consider ERA a promising method of GAI analysis compared to alternative PCG-evaluation practices due to its flexibility in applying

EXAG 2024: AIIDE Workshop on Experimental AI in Games, November 19, 2024, Lexington, KY

[*] Corresponding author.

[†] These authors contributed equally.

jdboucher@wpi.edu (J. Boucher); gmsmith@wpi.edu (G. Smith); ydtelliel@wpi.edu (Y. D. Telliel)

0009-0002-6865-9031 (J. Boucher); 0000-0002-2765-7702 (G. Smith); 0000-0001-5651-7349 (Y. D. Telliel)

a broad range of analysis metrics to its evaluated systems, particularly highlighting its strength of identifying bias and system tendencies- including as influenced by user input [1]. Withington et al. identifies weaknesses with current evaluation methods for PCG systems, suggesting they could be mitigated by diverse research frameworks and promoting the reuse of methodology where possible [9]. This paper seeks to answer this call by broadening the applicability of ERA.

We seek to do so by presenting a taxonomic framework for categorizing approaches used for ERA, drawing from similar work in PCG such as Togelius et. al. [10] and Smith's [2] taxonomies of PCG, as well as Withington et al.'s modern taxonomy of available evaluation approaches [9].

Further outlining the connection between GAI and PCG, we consider the motivations and use-cases for such systems. PCG and GAI are both used to automatically generate large amounts of content, and to increase variety of content [11, 12]. PCG is also used for assistive tools, helping with tasks such as level creation [13], and support tools have emerged to make these systems easier to understand [14]. LLM applications like ChatGPT [4] and Stable Diffusion [5] are promoted for their potential to reduce labor costs, enable new business models, and increase access to content creation– Cook claims all of these as common motivations for studying AI in games [15]. We also acknowledge the use of machine learning in PCG [16, 17] as a point that highlights the connection between PCG and GAI.

We also recognize parallel claims of GAI and PCG capabilities—and shortcomings thereof—that may provide useful points of interest for continuing this investigation. Developing a useful PCG generator often takes just as much, or more, effort than hand-crafted alternatives [18], despite the allure of rapid content production. Furthermore, the variety content generators are able to produce can be limited, as players are often capable of identifying patterns between generated content [19]. While Generative AI offers potential to expand the type of content that can be generated—such as more complex generative audio and music used to increase variety and interactivity of game music compared to non-generative methods [20], as well as the range of users that can make use of generators—we also view the drawbacks of PCG as rich sites of investigation in GAI. ERA has proven a useful method for identifying such weaknesses in the domain of PCG, and we hope to see its effectiveness applied to GAI as well.

GAI technologies have prompted significant concern—i.e. regarding its social and environmental impact, underlying politics of AI design, implementation, and advocacy [21]. Additional concerns include embedded systemic biases, risk of plagiarism and misinformation, and human and environmental costs [22]. Jiang et al. find that artists, in particular, identify harms to professional reputation, intellectual property, and financial risk [23]. Because of GAI's recognized potential for harm, we value tools and methods for understanding and analyzing these systems and their outputs.

3. The Problem Space

Summarizing the process of Expressive Range Analysis, this form of inquiry operates in three steps [1]:

- 1) Start with a set of images or similar data (i.e. levels, maps, etc.) produced by the generator being analyzed.
- 2) Determine useful metrics for the goal of analysis (i.e. linearity or leniency of a level) and apply them to the data.
- 3) Produce a visualization of the metered data space.

Two challenges arise when considering this method for GAI-generators as opposed to PCG-generators. First, producing a set of data to analyze tends to be more complicated. GAI text-to-X applications require linguistic input, leading to larger variance of the input space in terms of both quantity and meaning. Further complicating the issue, PCG tools are typically custom-made for specific use-cases [18], where GAI tools (such as ChatGPT [4] or Google's Gemini [24]) are commonly presented as general-purpose—while random inputs could be provided in place of handcrafted ones or the latent space could be randomly sampled, these approaches would provide undesirable outputs, lacking the targeted focus of a particular use-case. The general-purpose nature of GAI systems adds noise to the output space when looking at these tools for specific use-cases, since outputs that don't fit the specific use-case become irrelevant. Guzdial summarizes this aspect as GAI tool developers seeking to expand the possible valid outputs for *a particular tool* to include all possible valid outputs for *every tool* [8]. This noise presents additional challenges: how do you narrow the scope of the output to only match the relevant use-case? What prompts do you provide as an input to produce a suitable set of data to categorize the generator as a whole? Guzdial's human-centered input alignment [8] may prove a useful attribute of consideration to respond to this challenge.

The second challenge that arises from considering ERA for use in analyzing GAI systems comes from the metrics of analysis. ERA can be used alongside any metrics, depending on what a generative system is being analyzed for. These metrics must be applied to large quantities of data, in some cases encompassing the entire generative output of a system [1]. Furthermore, many examples of ERA employ automatic processing of data, but not all data and research goals benefit from automatable metrics- which tend to be quantitative in nature. The challenge here is twofold: how do you determine semantically meaningful metrics for increasingly variable data? And, accounting for practical workload and feasibility, how do you apply these metrics to increasingly multitudinous data?

We treat *producing an intended result* as the motivation behind creating and using generators. While some generators offer promises of increased speed and efficiency, those claims are not always accurate [18]—despite this, generators are valued by many as useful tools. We recognize the purpose of ERA as evaluating the success of that motivation. Using ERA is akin to asking: has this this generator successfully captured what it was designed to capture? With this framing in mind, our motivation for this paper is to facilitate the answering of that question in a greater variety of contexts. The challenge in the case of GAI is, in short: how?

4. Framework

Here, we present a taxonomy of ERA methods, seeking to provide language and framing to better address the challenges of applying ERA to a greater variety of contexts—especially GAI systems. We have defined five axes for characterizing ERA methodologies: quantitative vs qualitative, product vs process, objective vs subjective, content agnostic vs semantic, and automated vs manual.

We considered three factors as a basis for defining these axes: the origin and definition of ERA, the adopted practice of this method, and its potential for further expansion and refinement. We considered the origin of ERA [1], identifying decision points in how the method may be applied. We also considered how ERA has been adopted in research endeavors and sought to challenge assumptions that have become commonplace. Finally, this taxonomy is also a result of our efforts to address challenges we faced in applying ERA to GAI. We hope to further refine these axes and their definitions in future work.

4.1. Quantitative vs Qualitative

The metrics that are applied to the generator-produced data may be quantitative, qualitative, or varying degrees of mixed-method. Many traditional ERA approaches utilize quantitative metrics, as this approach is often more suitable for automatically applying metrics to data and producing a descriptive visualization. Smith and Whitehead use two metrics for applying ERA to Launchpad- linearity and leniency [1]. Both of these metrics are quantitative, because they are described, measured, and depicted numerically. Interestingly, Kreminski et al. [25] identify determining *quantitative* metrics as an essential step of ERA, which is consistent with the examples of the method's initial presentation [1]- it is therefore unsurprising that Qualitative metrics are relatively under-utilized in ERA. This taxonomy challenges this assumption—which is present across many ERA-focused research endeavors—instead suggesting an expansion of valid data collection efforts.

While there is not a strong foundation of qualitative metrics applied to ERA in PCG, some hypothetical qualitative metrics could come from methods like user surveys or interviews, such as those found in play-testing efforts to evaluate elements of a game like challenge or engagement. These metrics could be visually represented with tools such as word clouds, affinity charts, or heat-maps that highlight the frequency of thematic elements within the data, for example.

4.2. Product vs Process

Because our goal is to evaluate if a generator is successful in capturing what it was designed to do, it is interesting to consider both the product of a generator and the process of producing that output. The product of a generator is its output- an image, the answer to a question, or a video game level are all examples of this. Analyzing the product is useful for identifying what a generator is capable of- in terms of quality, variety, etc.- and what sort of biases are present in the outputs. Smith and Whitehead [1] present a product-focused approach, as their linearity and leniency metrics consider only the levels produced by the generative system. Withington's [26] approach is also product-focused, since this work considers the differences between outputs of a generator rather than considering the process of producing those outputs.

The process entails the experience of producing that output. Common narratives of generative systems sell them as faster and more efficient than manual alternatives. Looking at the process allows us to evaluate those claims, using metrics such as the time it takes to get a desirable output or how many iterations of prompts/generations it takes to produce such an output. Kreminski et al. [25] arguably present a process-focused approach, because their motivation for analysis seeks to evaluate the usability experience of a generative system. However, the metrics used are primarily focused on the product of the generator. A more process-centric example would include metrics that evaluate procedural aspects such as the time it takes to generate an output, or the number of generative attempts before a user finds a suitable option.

Shaker et. al. [11] highlight generator reliability as one important piece of PCG evaluation. Looking at this aspect provides a mixed-method approach that tends to use product to reveal something about the process.

4.3. Objective vs Subjective

This axis is concerned with the nature of the analysis metric—is a given metric verifiable based on factual evidence, or does it vary with perspective, based in emotion or opinion? Smith and Whitehead's two metrics for Launchpad again provide a useful example: Linearity, as described in their paper, is an objective metric that describes the factual ”profile” of platforming levels (how well the geometry of the level fits a straight line) in the evaluated system [1]. Leniency, however, is identified as a subjective score based on an intuitive sense of how lenient components of a level are towards a player [1]. Subjective metrics are relatively underrepresented in PCG, but may prove useful for evaluating GAI systems– especially as they are applied in creative domains. We consider user experience analysis a useful point of reference for how subjective metrics may be used in data analysis and visualization [27, 28].

4.4. Content Agnostic vs Semantic

This axis is concerned with the data itself and the value sought from analysis. Semantic approaches seek to find meaning within the context of the inputs and outputs of the generator. Smith and Whitehead [1] present a semantic approach, as their metrics are intended to allow comparison of the generated content. Lucas and Volz [29] provide another example of a semantic approach, as theirs is also intended to compare generated content.

Content agnostic approaches seek to find meaning in the generator, regardless of its inputs or outputs- though the inputs and outputs may be useful for analysis. While Kreminski et al. [25] evaluate the product of the generator, their interests lie in the users of the generative system-answering questions such as how thoroughly they explore the generative range- rather than finding meaning from the generative output itself, making their approach content-agnostic. Withington's exploration of quality-diversity algorithms [26] presents another content-agnostic example because the focus is not on the details of specific outputs, but rather measuring the differences between them.

4.5. Automated vs Manual

This axis is concerned with the process of applying metrics to the data. Automated processes are typically conducted computationally, making them especially suitable for quantitative analysis metrics and often desirable for the promise of reduced processing time or effort. Smith and Whitehead [1], Kreminski et al. [25], and Kybartas et al. [30], among others, all present automated approaches for applying ERA metrics to their data.

Manual processes require human labor for applying analysis metrics to each piece of data—while this tends to be more time-consuming, it also allows closer scrutiny of elements that are difficult to capture without direct human intervention. Manual processes are often unsuitable for the large quantities of data that ERA typically processes, but methods such as crowd sourced photogrammetry—e.g. as used for identifying information about wildlife populations [31]—may hypothetically be leveraged to manually process such data. Human subject experiments, such as those commonly used in narrative generation projects [32], are another example of a manual approach.

5. Discussion

Framing GAI under the lens of PCG- or both as the same process, e.g. through the lens of human-AI interaction generation [8]- incorporates this new technology into an established domain of research that has ERA as a design-focused method for evaluation. In connecting GAI and PCG, though, we have identified a need for both an improved vocabulary for describing ERA and an expanded potential scope for ERA that challenges existing methods. Thus this taxonomy further expands the potential range of analytical applications for ERA, providing language to better describe and imagine its use-cases and identify its historical gaps. This expansion responds to existing weak-points in PCG evaluation—such as those identified by Withington et al. [9]—by increasing the diversity and re-usability of ERA as a framework for generative analysis.

Further, this taxonomy allows for better description of the flexibility and space occupied by broadly applicable aspects of evaluation, such as those presented in Guzdial's human-AI interaction generation [8]. For example, Guzdial's call for human-centered input alignment considers the relationship between valid system inputs and user-preferences regarding those inputs. Using the vocabulary from our taxonomy, this is a process-focused, content-agnostic, subjective approach, because meaning is found according to individual perceptions without concern for the generative output. Such a metric could be applied to data quantitatively or qualitatively, using either an automated or manual approach. Guzdial's adaptability similarly considers the process of generation and user perceptions [8], and has similar axis placement to human-centered input alignment– though it is objective rather than subjective, since adaptability was a predetermined aspect of the generative process, rather than a variable expressed by human interpretation. Novelty, however, is an objective, product-focused metric because it is based in the observed possible generative outputs.

It is our hope that using the vocabulary of this taxonomy can provide some clarity to the research community on how they are using ERA for evaluation, as well as identify potential new approaches for ERA.

6. Conclusions, Limitations, & Future Work

This paper explores an avenue for exploring the intersection of PCG and GAI, using expressive range analysis as a common method for analyzing generative systems and their outputs. We present a taxonomic framework for categorizing existing and imagined ERA inquiries, hoping to allow more effective navigation of this space and leverage PCG tools for study of GAI technologies and systems.

This taxonomy opens interesting possibilities for future applications of ERA. What do qualitative applications of ERA look like? How would manually applying metrics to data compare to more commonly applied automated processes? These are relatively underexplored areas of ERA, and the possibility of applying this method to generative systems makes these questions more compelling.

The scope of this paper is limited to theory. This thread of research would benefit from a more complete, systematic review of ERA projects that places existing work on the axes of this taxonomy and further inform the chosen axes. As an extension of this research, we also see value in performing ERA on GAI tools using different combinations of axis placement—especially including qualitative and manual approaches.

Acknowledgments

This material is based upon work partially supported by the National Science Foundation (NSF) under Grant No DGE-1922761. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] G. Smith, J. Whitehead, Analyzing the expressive range of a level generator, in: Proceedings of the 2010 workshop on procedural content generation in games, 2010, pp. 1–7.

[2] G. Smith, Understanding procedural content generation: a design-centric analysis of the role of pcg in games, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 917–926.

[3] N. Deckers, J. Peters, M. Potthast, Manipulating embeddings of stable diffusion prompts, arXiv preprint arXiv:2308.12059 (2023).

[4] OpenAI, Chatgpt, https://chat.openai.com/, 2024.

[5] S. AI, Stable diffusion, https://stability.ai/stable-image, 2024.

[6] ElevenLabs, Elevenlabs, https://elevenlabs.io/, 2024.

[7] P. Korzynski, G. Mazurek, P. Krzypkowska, A. Kurasinski, Artificial intelligence prompt engineering as a new digital competence: Analysis of generative ai technologies such as chatgpt, Entrepreneurial Business and Economics Review 11 (2023) 25–37.

[8] M. Guzdial, Human-AI interaction generation: A connective lens for generative AI and procedural content generation, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), 2024.

[9] O. Withington, M. Cook, L. Tokarchuk, On the evaluation of procedural level generation systems, in: Proceedings of the 19th International Conference on the Foundations of Digital Games, 2024, pp. 1–10.

[10] J. Togelius, G. N. Yannakakis, K. O. Stanley, C. Browne, Search-based procedural content generation: A taxonomy and survey, IEEE Transactions on Computational Intelligence and AI in Games 3 (2011) 172–186.

[11] N. Shaker, J. Togelius, M. J. Nelson, Procedural content generation in games (2016).

[12] M. Hendrikx, S. Meijer, J. Van Der Velden, A. Iosup, Procedural content generation for games: A survey, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 9 (2013) 1–22.

[13] A. Liapis, G. N. Yannakakis, J. Togelius, Sentient sketchbook: computer-assisted game level authoring (2013).

[14] M. Cook, J. Gow, G. Smith, S. Colton, Danesh: Interactive tools for understanding procedural content generators, IEEE Transactions on Games 14 (2021) 329–338.

[15] M. Cook, Optimists at heart: Why do we research game ai?, in: 2022 IEEE Conference on Games (CoG), IEEE, 2022, pp. 560–567.

[16] A. Summerville, Expanding expressive range: Evaluation methodologies for procedural content generation, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 14, 2018, pp. 116–122.

[17] A. Summerville, S. Snodgrass, M. Guzdial, C. Holmgård, A. K. Hoover, A. Isaksen, A. Nealen, J. Togelius, Procedural content generation via machine learning (pcgml), IEEE Transactions on Games 10 (2018) 257–270.

[18] T. Short, T. Adams, Procedural generation in game design, CRC Press, 2017.

[19] E. Short, Bowls of Oatmeal and Text Generation, https://emshort.blog/2016/09/21/bowls-of-oatmeal-and-text-generation/, 2016.

[20] C. Plut, P. Pasquier, Generative music in video games: State of the art, challenges, and prospects, Entertainment Computing 33 (2020) 100337.

[21] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623.

[22] J. E. Fischer, Generative ai considered harmful, in: Proceedings of the 5th International Conference on Conversational User Interfaces, Association for Computing Machinery, New York, NY, USA, 2023.

[23] H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, T. Gebru, Ai art and its impact on artists, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2023, pp. 363–374.

[24] G. AI, Gemini, https://gemini.google.com/, 2024.

[25] M. Kreminski, I. Karth, M. Mateas, N. Wardrip-Fruin, Evaluating mixed-initiative creative interfaces via expressive range coverage analysis., in: IUI Workshops, 2022, pp. 34–45.

[26] O. Withington, Illuminating super mario bros: quality-diversity within platformer level generation, in: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, 2020, pp. 223–224.

[27] A. Bangor, P. T. Kortum, J. T. Miller, An empirical evaluation of the system usability scale, Intl. Journal of Human–Computer Interaction 24 (2008) 574–594.

[28] S. Djamasbi, Eye tracking and web experience, AIS Transactions on Human-Computer Interaction 6 (2014) 37–54.

[29] S. M. Lucas, V. Volz, Tile pattern kl-divergence for analysing and evolving game levels, in: Proceedings of the Genetic and Evolutionary Computation Conference, 2019, pp. 170–178.

[30] B. A. Kybartas, C. Verbrugge, J. Lessard, Tension space analysis for emergent narrative, IEEE Transactions on Games 13 (2020) 146–159.

[31] S. A. Wood, P. W. Robinson, D. P. Costa, R. S. Beltran, Accuracy and precision of citizen scientist animal counts from drone imagery, PloS one 16 (2021) e0244040.

[32] R. Sanghrajka, E. Lang, R. M. Young, Generating quest representations for narrative plans consisting of failed actions, in: Proceedings of the 16th International Conference on the Foundations of Digital Games, 2021, pp. 1–10.