Should Machines Evaluate Us? Opportunities and Challenges

Kuakou Bossou and Margareta Ackerman
Computer Science and Engineering Department
Santa Clara University
Santa Clara, CA 95053 USA
{cbossou,mackerman}@scu.edu

Abstract

Generation alone does not make a Computational Creativity system. But, what about machines that *only* evaluate? When it comes to co-creative systems, humans often take on the primary evaluative role, while machines assist with the generation of creative artifacts. In this paper, we propose flipping the paradigm, envisioning machines that (only) evaluate humans. Challenges and opportunities in this new direction are discussed.

Introduction

Evaluation is a critical aspect of Computational Creativity (CC). In fact, systems that generate without evaluating have been called "mere generation" (Ventura 2016), suggesting that without evaluation, a program should not be considered a CC system. In the co-creative paradigm, the system often takes on the role of the generator, while the human evaluates. In fact, while it can be argued that computers are better at generating, humans retain an advantage in our evaluative capacities (Karimi et al. 2018).[1]

In contrast to generative systems, here we propose the study of *evaluative systems*. To differentiate from evaluators in other spaces, we also refer to such systems as "creative evaluators." We ask a daring new question: What happens when a system only evaluates, and does not generate? What role can such systems play in Computational Creativity? From a co-creative standpoint, we explore an extreme point on the continuum, asking: What if we had a co-creative system that *only* evaluates?

The concept of a machine evaluating humans has been explored in other spaces, primarily in contexts where a convergent solution is desired. For example, machines have been used to evaluate mortgage applicants (Chen, Guo, and Zhao 2021), grade essays (Santos, Verspoor, and Nerbonne 2012; Ramalingam et al. 2018), judge startup pitches (Hu and Ma 2020), and inform investment decisions (Wu and Gnanasambandam 2017; Bento 2018). In these cases, the problems were formulated as machine learning models through a convergent lens, assuming that there is a single ground truth and the aim is to accurately score or classify.

In contrast to these types of evaluative systems, creative evaluators face a divergent problem, assessing the quality of creative artifacts where there often is not, and typically should not be, a ground truth solution. In this context, the creative evaluator can provide value to a human creative partner by not only evaluating the result, but also by providing helpful feedback. This iterative process can consequently form a meaningful co-creative experience between the human and machine, without the machine ever (directly) engaging in generation.

In this paper, we present our vision for creative evaluators. We contrast creative evaluators against evaluators in other spaces, and discuss the particular challenges of building evaluators for interaction with humans on creative tasks.

Previous work

The Computational Creativity literature stressed the importance of evaluation. Veale and Pérez y Pérez (2020) write that evaluation is a staple of Computational Creativity. Similarly, Ventura (2016) proposes that mere generation systems may not be considered creative. Ventura (2016) writes that "the Computational Creativity community (rightfully) takes a dim view of supposedly creative systems that operate by mere generation". He further argues that the question of whether a system is beyond mere generation is very closely related to the question of system evaluation, making evaluation part and parcel of creative systems.

Due to the centrality of evaluation in Computational Creativity, CC systems often consist of both a generative and evaluative component (see, for example (Toivonen and Gross 2015), for a discussion of generative and evaluative components of creative systems that utilize machine learning and data mining methods). Note that the emphasis is typically on incorporating evaluation into creative machines that also generate, rather than considering machines who sole purpose is to evaluate.

A number of prominent evaluation methods have gained recognition in the CC community. Ritchie (2007) suggested empirical criteria for evaluating the relative value and novelty of a system output. Colton et al. (2011) propose two formal complementary models (FACE and IDEA) for evaluating creative acts of CC programs. The FACE model

[1] More balanced frameworks have also been proposed (Kantosalo and Toivonen 2016), challenging us to develop machines that engage more deeply in a co-creative process with humans through a combination of generation and evaluation.

describes a program creative act whereas the IDEA model embodies notions related to the impact of the creative act. Jordanous (2012) proposed a three-step Standardised Procedure for Evaluating Creative Systems (SPECS). Pérez y Pérez (Pérez y Pérez 2014) proposed a three-layer evaluation model for computer-generated plots. Ventura (2016) presented a spectrum of abstract prototype systems that can be used as benchmarks for evaluating relative creative ability of CC systems. More recently, evaluation of co-creative systems has also been explored (Karimi et al. 2018; Kantosalo and Toivonen 2016).

Evaluation in CC has been studied from three distinct perspectives: A system evaluating its own artifacts (Ackerman and Loker 2017; Pérez y Pérez and Sharples 2001) as in internal process, evaluation of autonomous CC systems (Colton 2012; Pérez y Pérez and Sharples 2001), and evaluation of co-creative systems (Karimi et al. 2018; Kantosalo and Toivonen 2016; Ackerman and Loker 2017). We approach evaluation from a novel perspective, where a machine evaluates artifacts created solely by a human.

Considering machines evaluating humans, there is some relevant work outside of the context of creativity. Those systems include selecting mortgage applicants (Chen, Guo, and Zhao 2021; Thomas, Crook, and Edelman 2017), grading essays (Santos, Verspoor, and Nerbonne 2012; Ramalingam et al. 2018), and startup investment decisions (Wu and Gnanasambandam 2017; Bento 2018). There have even been attempts at making automatic paper reviewing systems (Leng, Yu, and Xiong 2019), although these are in their early stages.

This paper proposes the challenge of introduction evaluation-only systems to Computational Creativity. What role can systems that evaluates, but do not generate, can play in CC? What would it take to create such machines?

Creative Machine Evaluators

In the most common co-creative paradigm, the computer agent's primary contribution is on the generative side (even if the computer engages in some internal evaluation), while the human takes on the main evaluative responsibilities (even if the human also engages in generation).

For example, the Computational Creativity musical, "Beyond The Fence" was based on an original idea by Simon Colton's WhatIf machine (Colton et al. 2016). The WhatIf machine engine generates fictional plots using What-if scenarios. Many ideas were generated by the WhatIf machine, allowing the makers of the musical to select one of those ideas, which became the starting point for the musical's plot.

As another example, ALYSIA (Ackerman and Loker 2017) is a co-creative system that originally focused on assisting users with the creation of original vocal melodies. The machine suggests melody lines for user-provided text, which the user could select, alter, or ask for more options. This process captures role allocation based on the main strengths and weaknesses of human and machine agents.

While co-creative systems where the computer agent's contribution is largely on the side of generation are common, more balanced models have also been considered. For example, alternating co-creativity (Kantosalo and Toivonen 2016) puts humans and machines on more equal grounds. It is further suggested that the attained results should satisfy both parties.

Instead of placing humans and computers on an equal plane, we seek to study another under-explored interaction, which inverts the traditional co-creative paradigm: What happens if, instead of us evaluating machines, machines evaluate us? In this paper, we posit a new form of co-creativity where the human generates and the machine (only) evaluates.

Imagine, for instance, a machine that assists visual artists. Showing their art to the machine, the creative evaluator will provide meaningful feedback on the art - perhaps commenting on composition, color choices, or even how the art may relate to current events. The feedback may in turn help the artist to improve their work, much like feedback from a human domain expert. Similarly, we can envision a storytelling evaluator, which provides feedback on the story arc, character development, and overall quality of a user-provided story.

In the subsequent subsection, we discuss the landscape of machines evaluators as they exist today, following which we address challenges of building creative machine evaluators.

Taxonomy of Machine Evaluators

In Table 1, we classify machine evaluators in AI into three categories: (1) Convergent Evaluators, (2) Creative Machines that generate and evaluate, and (3) Creative Evaluators (that exclusively evaluate), which is the new category proposed in this paper.

Machine evaluators outside of CC tend to adhere to clear objective functions ("Convergent evaluators"). In contrast to traditional AI, however, CC systems are not inherently convergent. As Ventura (Ventura 2017) says, "There is no such thing as a best song, or best theorem or best design. One cannot maximize a piece of visual art or a recipe or a poem." As such, evaluators in creative domains must acknowledge the inherently divergent nature of creativity.

The remaining two classes of machine evaluators fit within CC and are consequently divergent in nature. Current CC machines that include evaluative capabilities also engage in generation. While it is not uncommon to find machine agents with creative aims that only generate ("mere generation"), we have not encountered creative machines that engage in evaluation without also engaging in generation. It has been argued that the former (machines that only generate) should not be considered CC machines (Ventura 2016). Our paper expands this dialog to ask what role the latter, machines that evaluate without generating ("Creative Evaluators"), can play in the CC space.

Creative Evaluators Opportunities

Before delving further into the challenges of developing creative evaluators, we discuss the opportunities for CC in this domain.

- *Essay competitions* may be approached from both a convergent and divergent perspective. Prior work has focused

Machine evaluators in AI			
Convergent Evaluators	**Creative machines that generate & evaluate**		**Creative Evaluators**
Machines evaluating humans outside CC	Machine evaluating autonomous system's artifacts	Internal evaluation in co-creative systems	Machine evaluating human artifacts
• Startup investment • Essay • Mortgage	• Painting fool • MEXICA	• Impro-Visor • ALYSIA	**?**

Table 1: Machine Evaluators in AI and their applications. We introduce the new category of Creative Evaluators.

on correctness-based criteria (Santos, Verspoor, and Nerbonne 2012), such as the percentage of errors appearing in the writing. However, there are also opportunities to approach essay evaluation by viewing an essay as a creative artifact, and making a creative evaluator that would both judge and provide feedback on an essay through this broader lens.

- In *business*, there are applications such as pitch competitions and resume comparison applications (Roy, Chowdhary, and Bhatia 2020). The approaches taken for such evaluators typically fall under the convergent category. However, there are opportunities for building creative business-related human machine evaluators for creative tasks such as company names, logo creation, etc.
- In *mortgage and job applications* evaluation (Chen, Guo, and Zhao 2021; Thomas, Crook, and Edelman 2017), current machine evaluators are typically ML-based models that are prone to replication or even amplifying the biases found in the data on which they are trained. Creative evaluators may potentially offer an avenue for mitigating this problem if mortgage and job applications evaluation are viewed as a creative task, rather than a convergent one.
- *Education* offers another arena where creative evaluators can offer value. For instance, educational evaluators can provide personalized feedback to help art, music, or poetry students to improve - taking on a partial role of an educator. Further, the evaluators could provide assessment and grading for educational institutions. We can also envision creative evaluators tackling complex tasks such as curating gallery shows or casting actors.

Challenges with Creative Evaluators

In this section, we introduce several considerations and challenges when conceptualizing and developing creative evaluators.

Divergence

It is well-established that creativity is composed of quality/value and novelty (Ritchie 2007). Value encourages convergent thinking in seeking quality artifacts while novelty relies on divergent thinking to allow originality. As seen in Table 1, we already have convergent machine evaluators for a variety of applications.

When it comes to creative evaluators, we want these machines to go beyond mere convergent thinking. Primarily, the work that is being evaluated must be a creative task/artifact. More importantly, the machine evaluator must take into account novelty so as to avoid conformity while at the same time not lacking in quality.

Perhaps the primary challenge in the making of creative evaluators is to balance the need for providing concrete, justifiable feedback, while encouraging divergent thinking and pushing the human partner to explore profoundly novel possibilities. Creative evaluators may be modelled after the best educations, who seamlessly combine knowledge transfer with the fostering of divergent thinking, encouraging their students to take big risks into the unknown.

Explainability

We propose that having some level of explaniblity is a core feature of a Creative Evaluator. Recently, Explainable Computational Creativity has been proposed as

> the study of bidirectional explainable models in the context of computational creativity – where the term explainable is used with a broader sense to cover not only one shot-style explanations, but also for co-creative interventions that involve dialogue-style communications. (Llano et al. 2020)

In the context of machine evaluators in general and creative evaluators in particular, *explainability* is important consideration. The European AI Experts (High-Level Expert Group on AI) have been encouraging AI researchers to consider explainability as a core ethical AI principle.

We argue that a creative evaluator must be able to provide feedback to the human involved in the creative process. For example, after the user presents its artifact, the system may provide ideas to help the human to improve the artifact. While full explainability may not always be possible, particularly with ML-based models, higher degrees of transparency are desired.

Fairness

How can a Creative Evaluator be a fair judge/evaluator? The Journal of Computational Creativity defines Computational Creativity as "the art, science, philosophy and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative" (Journal of Computational Creativity). But, what is an unbiased observer?

Humans are known to exhibit biased behavior. Despite our best efforts, we may carry conscious or unconscious biases that impact how we evaluate (Fiarman 2016). We therefore ask if, in the place of human evaluators, machine evaluators may be less biased?

Machines don't run the risk of being impulsively subjective. Any degree of subjectivity on the part of a machine is likely to be more consistent. While eliminating bias from automated systems is a real challenge, it is more feasible for machine evaluators to avoid bias based on human agents' backgrounds (gender, race, etc) when such information is not explicitly provided to the machine evaluator. This form of discrimination may nevertheless happen if careful attention is not paid when utilizing data-driven approaches, algorithms or word embeddings that are being used (Bolukbasi et al. 2016), although there are certainly instances where one's identity impacts their art in a variety of ways.

Biased evaluation is a serious concern when either human or machine evaluators (or both) are involved. The question is whether in some contexts, fairness may be more feasible to attain through a machine agent than a typical human agent.

Conclusions and Future Work

How can a machine evaluator encourage divergent thinking? How do we practically develop such machines? In this paper, we introduce the concept of creative machines whose only role is to evaluate human-made artifacts. This can be conceptualized as an inversion of the typical co-creative paradigm, where the human generates and the machine evaluates. On the educational front, creative evaluators may act much like teachers, offering ongoing feedback for improvement. In competitive contexts, creative evaluators open up intriguing opportunities for fair evaluation of artistic artifacts.

We hope that this exploration into evaluation-only creative machines will spark discussion in the computational creativity community. If evaluation is indeed central to CC, is it sufficient for a machine to do nothing but evaluate? Turning off generation leads to many interesting applications, may get us closer to true unbiased observers, and gets us to question the very foundation of CC: If evaluation is all we have, will we go out searching for generation?

Acknowledgments

The authors would also like to thank two anonymous reviewers for their helpful comments.

References

Ackerman, M., and Loker, D. 2017. Algorithmic songwriting with ALYSIA. In *International conference on evolutionary and biologically inspired music and art*, 1–16. Springer.

Bento, F. R. d. S. R. 2018. *Predicting start-up success with machine learning*. Ph.D. Dissertation.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Chen, S.; Guo, Z.; and Zhao, X. 2021. Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research* 290(1):358–372.

Colton, S.; Llano, T.; Hepworth, R.; Charnley, J.; Gale, C.; Baron, A.; Pachet, F.; Roy, P.; Gervás, P.; Collins, N.; et al. 2016. The beyond the fence musical and computer says show documentary.

Colton, S.; Charnley, J. W.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *ICCC*, 90–95.

Colton, S. 2012. The painting fool: Stories from building an automated painter. In *Computers and creativity*. Springer. 3–38.

Fiarman, S. E. 2016. Unconscious bias: When good intentions aren't enough. *Educational Leadership* 74(3):10–15.

High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

Hu, A., and Ma, S. 2020. Human interactions and financial investment: A video-based approach. *Available at SSRN*.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Journal of Computational Creativity. About this Journal. Association for Computational Creativity. https://jcc.computationalcreativity.net/.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the seventh international conference on computational creativity*, 77–84.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. *arXiv preprint arXiv:1807.09886*.

Leng, Y.; Yu, L.; and Xiong, J. 2019. Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In *2019 International Conference on Multimodal Interaction*, 395–403.

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational creativity. In *Proc. ICCC*.

Pérez y Pérez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Pérez y Pérez, R. 2014. The three layers evaluation model for computer-generated plots. In *ICCC*, 220–229.

Ramalingam, V.; Pandian, A.; Chetry, P.; and Nigam, H. 2018. Automated essay grading using machine learning algorithm. In *Journal of Physics: Conference Series*, volume 1000, 012030. IOP Publishing.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Roy, P. K.; Chowdhary, S. S.; and Bhatia, R. 2020. A machine learning approach for automation of resume recommendation system. *Procedia Computer Science* 167:2318–2327.

Santos, V. D.; Verspoor, M.; and Nerbonne, J. 2012. Identifying important factors in essay grading using machine learning. *Selected papers in memory of Dr Pavlos Pavlou—Language testing and assessment round the globe: Achievements and experiences. Frankfurt: Peter Lang. Available at http://urd. let. rug. nl/nerbonne/papers/Santos_et_al-2012-grading. pdf.*

Thomas, L.; Crook, J.; and Edelman, D. 2017. *Credit scoring and its applications*. SIAM.

Toivonen, H., and Gross, O. 2015. Data mining and machine learning in computational creativity. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5(6):265–275.

Veale, T., and Pérez y Pérez, R. 2020. Leaps and bounds: An introduction to the field of computational creativity. *New Generation Computing* 38(4):551–563.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24. Sony CSL, Paris.

Ventura, D. 2017. How to build a CC system. In *ICCC*, 253–260.

Wu, V., and Gnanasambandam, C. 2017. A machine-learning approach to venture capital. *The McKinsey Quarterly*.