The Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness

Michael P. Clemens

University of Utah, 50 Central Campus Drive, Salt Lake City, 84112, USA

Abstract

While explainable computational creativity (XCC) seeks to create and sustain computational models of creativity that foster a collaboratively creative process through explainability, there remains no way of quantitatively measuring these models. Through this research, we propose *The Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness* to quantitatively assess the artists' experience within the system concerning this communication paradigm.

Keywords

computational creativity, explainable artificial intelligence,

1. Introduction

With the recent explosion of work using neural networks and deep learning for co-creative applications, researchers are calling for explainability within these computational models [1]. The work surrounding this effort is called Explainable Computational Creativity (XCC). Although there have been efforts to explore this area—e.g. the work by Zhu et al. 2018 for video games, Bryan-Kinns et al. 2022 for music—none to date have defined a framework for evaluating a computational model's *explainability* within collaborative co-creative applications. This research introduces the *Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness* and elaborates on each E related to its creative application.

Success in computational modeling has led to a surge of research that promises autonomous systems to learn, decide, and act on their own volition. Although these systems have produced tremendous results within their respective contexts, their effectiveness is often hindered by the lack of transparency from the model itself [4]. From this lack of transparency, humans are often reluctant to implement techniques that are not interpretable, tractable, and trustworthy [2].

XAI explores how computational models such as ensembles, neural networks, or deep-learning methods can be made more understandable to humans. The motivation behind this field of

---

*Corresponding author.

✉ michael.clemens@utah.edu (M. P. Clemens)

🌐 http://mclem.in/ (M. P. Clemens)

🆔 0000-0002-4507-8421 (M. P. Clemens)

research is to increase the usability and accessibility for non-experts to utilize these models intuitively in their respective domains [5].

XCC has been presented as a sub-field of XAI, emphasizing the construction of models that foster bi-directional communication between the user and the system. Within this same context—and grounded in HCI and creativity literature—Bryan-Kinns et al. (2022) argued that AI in the interactive arts can be systematically analyzed in three ways:

1. Per the role of AI - ranging from models that perform generative tasks without engaging with humans to AI models that serve as collaborators in creative partnerships.
2. Per the interaction with the AI - the more interactive and responsive an AI is, the more likely people would grasp what it is doing now and in the future.
3. Per the common ground with the AI - classify what a person might be able to infer about an AI's output state.

They argue—and we agree—that a creative AI's explainability depends on all these facets. For instance, more explanation seems necessary as a process becomes more collaborative, demanding more engagement and grounding. Further, increased contact with the agent supports individuals in learning about and inferring knowledge and understanding of the co-creative AI. We complement their case-study based work by arguing that explainability can *also* depend on the creative AI's status relative the creativity literature surrounding the Four P's [6].

2. Research Plan

The Three E's is a framework we created to evaluate a system's explainability, specifically in the context of co-creative applications. This framework does not seek to find the optimal solution for every creative domain within computational creativity (CC). Rather, this framework provides a tool for members of the CC community to use while building explainability into their computational models as it is needed. Although our Three E's should *not* be used as a way to justify whether the system is inherently explainable, this tool can be used to guide designers in curating the type of experience they want to create between the co-creative agent and the artist.

2.1. Research Objectives

The main research objective is to create a framework that can be used to quantitatively assess the explainability of a model employed within a co-creative application. Researchers have converged on evaluating a system's creative potential relative to *The Four P's: Person, Process, Product, and Press* [6]. That is, modern evaluations use at least one P to describe the work's type of impact on the field [7]. We therefore plan to discuss each P as it relates to explainability.

At the same time, we note that the need for explainability is (at least) culturally determined. Different cultures have idiosyncratic information needs and processes that demand unique information architectures [8]. In turn, the systems are impacted by the environment in which they are deployed. Hence, culture has a profound bi-directional influence on co-creative system design [9]. Our current measures of E's within P's have the caveat that we do so from a global Western cultural perspective and that other cultures may require different levels of emotionality, effectiveness, and explicitness more appropriate for their explainability needs.

2.2. Approach / Methodology

In this section, we explore each of the E's in detail and justify our rationale behind using them to deepen our understanding of explainability within collaborative computational co-creative applications.

Emotionality   In collaborative systems, it's imperative to discern how affect will be demonstrated as it has a significant influence on how the interaction between the participants unfolds and, subsequently, what kind of sensemaking is derived from this interaction [10]. Participants' reactions are triggered by emotions induced by stimulus events, allowing them to adjust to an ongoing collaboration [11]. During a collaborative session, participants can be aware of their collaborators' feelings, which helps them verify their actions from their collaborator's perspective and use this awareness to continue with participatory sensemaking [12]. As Leite et al. (2013) have demonstrated, reifying these characteristics is essential: meaningful human-robot relationships are *shaped* by the robot's ability to communicate emotions.

Inspired by this line of work, we present *emotionality* as our first key dimension of explainability. We define *emotionality* as the agent's capacity to (a) understand the user's emotions and (b) offer feedback that predictably elicits targeted emotion(s). While working alongside co-creative agents, artists will ascribe certain beliefs and values to that agent [14]. To meet artists' expectations, the system's interaction framework should include the articulation of emotional input from the user and provide the appropriate feedback for the user to observe emotional output.

This dimension begs the question: "*How do you measure the emotionality of the explainability within a system?*" While it is not obvious how to measure this property of a model, we propose using a *high*, *medium*, and *low* scale for the amount of emotional feedback observed by the artist from the co-creative agent, as well as for the amount of emotional input the system affords the artist to articulate. We might imagine using this same scale to quantify *effectiveness* and *explicitness*. For example, high emotionality might be described as a co-creative agent receiving the articulation of emotional input by the user and presenting emotionally relevant feedback that is observable. Medium emotionality might be a co-creative agent that can receive emotional input from the user yet the system presentation lacks emotionally relevant feedback. An agent that cannot engage emotionally with the artist might have low emotionality.

Effectiveness   Designing creative systems ultimately requires evaluating their underlying computational models of creativity [15]. When evaluating these systems, the term *effective* is used to express whether the system was successful in accomplishing its intended goal.

Although we as a community wish to have a framework that supports a standardized way of evaluating a creative system's effectiveness, we lack an agreed-upon metric that can be used across domains [16]. Further, the term *effective* itself is subject to interpretation. For example, Hartson et al. (2001) use it to denote the *thoroughness* and *validity* of a system, per quantitative *usability* evaluations—a system is effective (i.e. achieves its intended goal) to the degree it is thorough and valid.

In our Three E framework, we reformulate the term *effectiveness* to describe how well the user's mental model of the creative system corresponds to its exhibited behavior. In the design

sciences, Gero and Kannengiesser (2004) argue that designers perform an *evaluation* to identify whether the user behavior a designed artifact *should* elicit corresponds to the behavior that *actually manifests* from the designed artifact's use. Here, we generalize this notion to include all steps of the design process, not solely the artifact's evaluation. That is, *effectiveness* reflects the user's capacity to *predict* (and thereby *direct*) how the co-creative system will act based on their mental model of the system.

High effectiveness might describe a frictionless match between the agent's behavior and the expected behavior. Medium effectiveness might describe varying discontinuities between the system behavior and the user's mental model of the agent's behavior. Low effectiveness represents almost no match between how the system behaves and the expected behavior derived by the structure from the user.

Explicitness   Interpretability and explainability have become conflated in the AI literature [19]. To situate our work, we rely on the use of these terms within XAI. Recently, Alvarez Melis and Jaakkola (2018) proposed that explanations should meet three general characteristics: explicitness, faithfulness, and stability. To them, explicitness addresses the question: "*Are the explanations immediate and understandable?*" Relatedly, Palacio et al. (2021) define explicitness as "*how understandable are the explanations*" relative to the ease of a person's interpretation for given explanations.

Inspired by this line of work, we propose explicitness ought to assess both the explainability of the model and the justification for the model's complexity. Explaining black-box models mathematically or computationally may not be appropriate for tackling explainability in CC applications [22]. Instead, justification and rationale for added complexity should be centered.

High explicitness might describe when non-specialists can readily understand the model's explanations without the intervention of an expert user. Medium explicitness might denote when the explanations require domain-specific knowledge but are still understandable within that context. Low explicitness might describe a model's explanations that are either completely absent or unintelligible by anyone other than a domain expert.

Computational Model   Although the proposed evaluation framework focuses on assessing and defining the explainability of an artifact, we have yet to discuss what an explanation is and how it manifests itself via a computational creativity system. In this manner, we proposed a computational model including George Abowd's [23] four agents in his *Framework for Discussing Interaction*: two explicit (*User* and *System*) and two implicit (*execution* and *evaluation*). *Execution* is articulation from the *User* of the problem and the *performance* metric from the *Input* to the *System* itself. *Evaluation* is the presentation from the system, creating the *Output* that is then observed by *User*.

Figure 1 demonstrates a computational model of an explanation in CC domains, including three stages, the User, the explainable user interface (XUI), and the Computational Creative Agent (CCA). The XUI is the interaction for the explanation between the CCA and the User. The User articulates the problem they want to explore (e.g., a *trigger*). The XUI will take this trigger, create a problem formulation based on the User's goal, and articulate that need to the CCA. The CCA will interpret this Input as the User's Goal State. The User's Goal and Initial

State will be assessed in the Planning Stage, where the CCA will produce a plan to take the User from the Initial State to the Goal State through a series of explanations. The Output from the Planning Stage will be the Input to the Plan Synthesis, which will focus on the XUI. This step in the process will determine from the list of explanations provided by the planning stage which will be most effective for the User. The Output of the Plan Synthesis will be both an Output to the User and an Input to the User's Initial State. This pathing is due to the long-term memory design principle within XCC systems, where the system will update over time based on what the User has learned. As the system explains more of the process, these explanations will be added to the memory bank as explanations that have been used previously and the determination of whether they were influential on that User. Let us take a User who has been presented with an explanation yet continuously asks the same trigger question from the XUI to the CCA. The CCA should be able to determine that the chosen explanation used was not a sufficient explanation for the User's intended goal state.
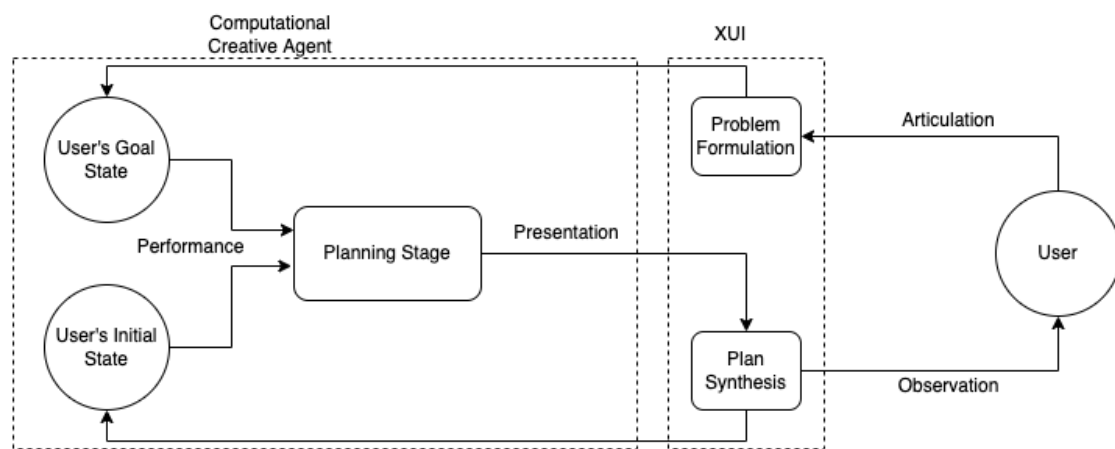

Figure 1: A computational model of an explanation using Abowd's Interaction Framework, concepts from XUI, and XCC.

3. Progress Summary

The first research project targeting this line of work used CBR within a co-creative agent to assist musicians in their aesthetic goals through a vocal audio plugin. Results showed that although participants were interested in using a co-creative agent throughout the production process, they acted against the vocal plugin parameter recommendations set by the agent. Participants showed frustration when the co-creative agent acted in a way that deviated from set expectations. From this research, we posit that explainability is an essential aspect of effective CBR models within co-creative agents.

Our next goal is to assess the various levels of explainability aforementioned in the context of the Four P's. This will involve at least four projects focusing on each P individually and evaluating the explainability based on user interactions.

References

[1] M. T. Llano, M. d'Inverno, M. Yee-King, J. McCormack, A. Ilsar, A. Pease, S. Colton, Explainable computational creativity., in: ICCC, 2020, pp. 334–341.

[2] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. M. Youngblood, Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation, in: IEEE CIG, 2018, pp. 1–8.

[3] N. Bryan-Kinns, B. Banar, C. Ford, C. Reed, Y. Zhang, S. Colton, J. Armitage, et al., Exploring xai for the arts: Explaining latent space in generative music (2022).

[4] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable ai, arXiv preprint arXiv:1810.00184 (2018).

[5] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE trans. on Neural Networks and Learning sys. 32 (2020) 4793–4813.

[6] M. Rhodes, An analysis of creativity, The Phi Delta Kappan 42 (1961) 305–310.

[7] C. Lamb, D. G. Brown, C. L. Clarke, Evaluating computational creativity: An interdisciplinary tutorial, ACM CSUR 51 (2018) 1–34.

[8] J.-m. Choe, The consideration of cultural differences in the design of information systems, Information & Management 41 (2004) 669–684.

[9] T.-F. Kummer, J. M. Leimeister, M. Bick, On the importance of national culture for the design of information systems, B & I Systems Engineering 4 (2012) 317–330.

[10] S. Abdellahi, M. L. Maher, S. Siddiqui, Arny: A co-creative system design based on emotional feedback., in: ICCC, 2020, pp. 81–84.

[11] R. K. Sawyer, Group creativity: Music, theater, collaboration, Psychology Pr., 2014.

[12] U. X. Eligio, S. E. Ainsworth, C. K. Crook, Emotion understanding and performance during computer-supported collaboration, Comp. in Human Beh. 28 (2012) 2046–2054.

[13] I. Leite, C. Martinho, A. Paiva, Social robots for long-term interaction: a survey, Int'l J. of Social Robotics 5 (2013) 291–308.

[14] L. Henrickson, Tool vs. agent: attributing agency to nlgs, Dig. Creativity 29 (2018) 182–190.

[15] A. Jordanous, A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative, Cog. Comp. 4 (2012) 246–279.

[16] P. Karimi, K. Grace, M. L. Maher, N. Davis, Evaluating creativity in computational co-creative systems, arXiv preprint arXiv:1807.09886 (2018).

[17] H. R. Hartson, T. S. Andre, R. C. Williges, Criteria for evaluating usability evaluation methods, Int'l J. of HCI 13 (2001) 373–410.

[18] J. S. Gero, U. Kannengiesser, The situated function–behaviour–structure framework, Design stud. 25 (2004) 373–391.

[19] T. Miller, P. Howe, L. Sonenberg, Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences, arXiv (2017).

[20] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, NeurIPS 31 (2018).

[21] S. Palacio, A. Lucieri, M. Munir, S. Ahmed, J. Hees, A. Dengel, Xai handbook: Towards a unified framework for explainable ai, in: IEEE/CVF, 2021, pp. 3766–3775.

[22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Mach. Intel. 1 (2019) 206–215.

[23] G. D. Abowd, Formal aspects of HCI, University of Oxford Oxford, 1991.