International Conference on Computational Creativity: What If the Emperor's New Clothes Are A Straitjacket?

Pablo Gervás
Facultad de Informática
Universidad Complutense de Madrid
pgervas@ucm.es

Carlos León
Facultad de Informática
Universidad Complutense de Madrid
cleon@ucm.es

Dan Ventura
Department of Computer Science
Brigham Young University
ventura@cs.byu.edu

Abstract

The *International Conference on Computational Creativity* was founded to provide an interdisciplinary home for research that falls outside the evaluation criteria of mainstream AI conferences—work that is exploratory, speculative, and by definition difficult to measure. Over time, the conference has suffered from a set of structural tensions that have gradually shifted the conference away from the role it was designed to play. As part of this shift, the conference has drifted toward the very norms it was meant to complement: quantitative evaluation, reproducibility requirements, and a reviewing culture that may be systematically filtering out the most genuinely creative work. This paper describes some of those problems, examines the forces driving them, and suggests a set of concrete measures that, in our view, could help restore and renew the community's health. We propose a re-calibration of our reviewing heuristics to protect the "gray areas" of creativity research, suggesting that ICCC must prioritize being a unique discussion forum over becoming a "poor relative" to mainstream AI venues.

Introduction

Computational Creativity (CC) as a discipline emerged, in large part, to carve out intellectual space for questions that more established venues could not comfortably host. How can a computational system produce artifacts or ideas that are novel, surprising, and valuable? What does creativity even mean when the agent is a machine? These questions resist the tidy experimental protocols that serve well in other corners of AI and cognitive science, and the field's founding community knew it. The *International Conference on Computational Creativity* (ICCC) was established to provide a home for research that often falls between the cracks of traditional disciplines. It was designed to be a space where the philosophical, the artistic, and the computational could coexist without the restrictive "guardrails" of the classical STEM disciplines upon which it is built: computer science, mathematics, cognitive science, etc. To achieve this, ICCC was designed as a meeting place where unconventional methodology, provocative ideas, and cross-disciplinary exchange were not merely tolerated but actively cultivated. What made it distinctive was not merely the subject matter—though the questions were and remain genuinely fascinating— but the way the community approached those questions: with intellectual openness, methodological pluralism, and a willingness to sit with problems that resist easy formalisation.

Over many years of participation: reviewing papers, serving on programme committees, attending and running sessions, watching the community grow and change—we have observed a gradual drift that concerns us. ICCC is now an established conference, but are we happy with the Emperor's new clothes? Conversations within the community reveal a shared unease: that ICCC is becoming harder to distinguish from the conferences it was not supposed to resemble.

The purpose of this paper is to express a shared view on the evolution of the conference over time, offer our best understanding of how it may have shifted away from its original purpose, and suggest some practical paths forward that may resurrect some of the redeeming characteristics it had at its birth. We argue that these symptoms are neither coincidental nor inevitable. They share common structural causes, and they call for deliberate structural responses. We offer this as a provocation in the spirit of the ICCC short-paper format: not a finished programme of reform, but a set of issues and possible directions that we believe deserve community-wide debate.

We want to be clear about the epistemic status of what follows. This is not an empirical study; it is a practitioner's account, shaped by our particular vantage points and inevitably partial. We offer it in that spirit, and we welcome correction and disagreement from others whose experience differs.

Concerning Symptoms

The evolution and growth of any conference will naturally surface various concerns; we have identified two in particular that seem critical to the health of ICCC and thus to the thread of our argument: *field specific reviewing criteria becoming sidelined by standard expectations* and *the conference becoming a venue rather than a meeting place*.

Reviewing Criteria

Over the last few years, we have observed a shared sentiment discussed by an ever widening subset of members of the community. As the conference grows and evolves it is receiving a significant injection of new researchers. This

initially started at the level of researchers submitting their work to the conference. When the number of papers to be reviewed each year grew, the size of the Program Committee had to increase as well. This involved an additional injection of new researchers as reviewers. These new reviewers volunteered to undertake a very significant workload, usually on top of similar work they do for other conferences: they receive the papers assigned; they find a review template with which they are familiar; and they do their best to fill it in with arguments that are recognizably academic. We are forever grateful to all of them.

However, a natural side-effect is that reviewing is increasingly governed by heuristics imported from adjacent disciplines—heuristics that are appropriate there and inappropriate here. This is not necessarily the reviewer's fault, but rather it reflects the difficulty of finding enough qualified reviewers as the conference has grown. As a result, reviews apply default disciplinary heuristics rather than the field-specific standards that the work requires, which causes a damaging feedback loop: work that would energise the conference is rejected; researchers who would contribute most do not submit; the pool of relevant reviewers shrinks further.

At least two aspects of ICCC offer significant challenges in this regard: *different paper types* and *different evaluation focus*. We discuss both here.

Different Types of Computational Creativity Papers The Call for Papers for this year's edition lists five different types of full papers, and ten different types of short papers. Both for full and short papers there is only one type that corresponds closely to the traditional technical paper as considered in other STEM conferences. Most of the others correspond to additional categories of paper developed over the years in an attempt to capture other possible ways of making progress in the understanding of computational creativity. In that sense, they are very likely to require significantly different criteria for acceptance. This has been the case historically for ICCC events from the beginning.

Yet it is very frequently true that reviewers newly come to the conference look at one of these paper types, apply wholesale the criteria they have developed for the average STEM conference, and reject them outright because they do not qualify under the generally accepted standard for technical papers. This leads to situations where papers that would have fit in well in the conference being left out: e.g., position papers rejected for not including quantitative evaluation; or papers reporting field events rejected for not including sufficient technical detail.

Different Focus for Quantitative Evaluation In addition to the difficulties presented by the large set of different paper types, even the standard for technical papers is different from that applied in other conferences. Most conferences impose baseline requirements on paper acceptance, including both quantitative evaluation and reproducibility. This is consistent with sound scientific methodology. But historically, both of these goals have been shown to present difficulties when applied to creativity.

The application of quantitative evaluation ensures that papers report results that are measurably correct. However, in the field of creativity there is no easy definition of what is correct. A system that scores well against a gold standard of prior creative work is, by the field's own definition, not creative. A system whose outputs are indistinguishable from prior work is, by the field's own criteria, not creative. If a system's primary goal is to reproduce a human-produced outcome to a high degree of statistical accuracy, it is performing a task of mimicry, not creativity. Historically, the field of computational creativity did go through a phase in which the community struggled with how to address the concept of evaluation. Slowly a body of work on evaluation of creativity arose. But this approach focused on trying to evaluate—sometimes in quantitative terms—how creative a system could be considered. Importantly, the characteristic that was being evaluated here was not how close to existing outputs a system could get but rather how different from them it could get while still producing valuable outputs itself—an extremely difficult balance to find. If this creativity-specific approach to quantitative evaluation is ignored, papers that explore the conceptual boundaries of creativity—areas where consistent quantitative metrics may not yet exist or may even be counter-theoretical—may be rejected for a lack of "rigor." This creates a paradox in which a system designed to reproduce a "Gold Standard" is rewarded, while a system that challenges the very nature of that standard is dismissed. If these important points are ignored, and reviewers systematically insist on metrics borrowed from, for example, mainstream ML, a systematic bias naturally arises against the most creative submissions.

The requirement of reproducibility also presents important challenges. Creativity research frequently involves processes that are intentionally singular, exploratory, or context-dependent. To require that results be straightforwardly reproducible is, in many cases, to require that they not be creative. Reproducibility is a genuine scientific virtue—but it sits in tension with novelty. If a result can be straightforwardly reproduced, it may indicate that the process is well-understood and parametric, which is at odds with what we typically mean by creative surprise. When reviewers demand strict reproducibility as a gate-keeping metric for ICCC, they create a "guardrail" that accidentally blocks novel, "one-off," or truly emergent creative phenomena. Requiring reproducibility as a blanket condition risks filtering out precisely the exploratory, one-off, or process-oriented work that ICCC was meant to host. The irony is that this requirement most severely penalises the work that is most genuinely experimental in character.

Venue vs. Meeting Place

This is perhaps the hardest thing to articulate precisely, but it is the change that concerns us most. ICCC began as, and for many years remained, a place where a community went to think together. Papers were a vehicle for conversation, not the point in themselves. The informal exchanges—at lunch, between sessions, over dinner—were where ideas were tested and collaborations formed. Researchers attending the conference discussed issues, collective decisions were made and the field's shared values were renegotiated.

We have watched that function erode. Attendance decisions are increasingly driven by publication needs rather than intellectual engagement. The number of researchers genuinely interested in establishing specific rules to drive research on creativity forward beyond the standard approaches to science has declined. The number of papers accepted at the conference that discuss issues important to creativity but which would not have been accepted at other STEM conferences has also declined. The conference has gradually come to resemble what it was not supposed to be—another ranked venue where researchers deposit output rather than a community where researchers think together. Once this dynamic takes hold, it becomes self-reinforcing: the less the conference functions as a community hub, the fewer reasons the people that would have contributed the most have to attend it in that spirit. We have reached a point where many researchers seem to feel that they must play by rules that value numbers over nuance.

Sources of Pressure

The problem is compounded by two sources of pressure on the participants of the conference: *institutional pressure* and *the rankings trap*.

Institutional Pressure Funding agencies, research assessment exercises and the logic of bibliometric ranking all impose real costs on researchers, especially those early in their careers, for investing in a venue that does not appear on the approved lists used by the institutions that evaluate them. We are not unsympathetic to this pressure. We have watched talented younger colleagues navigate it with less support than they deserved.

The Ranking Trap Conferences assigned a low ranking by any body seen as credible suffer a structural disincentive that affects who submits, who reviews, and who attends. Trying to rectify this by chasing a higher ranking requires behaving more like venues ranked highly, which can introduce untenable compromises. This dynamic tends to reward conformity and potentially penalise distinctiveness. At the same time, it is often equally untenable to ignore ranking information—many researchers face an environment in which, at best, they are discouraged from investing time in venues that are not highly ranked or that lack "objective" scientific markers.

Impact on Conference Management These two additional factors put pressure on conference organizers to take measures that further complicate the situation. Attempts have been made over the last few years to improve the scores assigned to the conference by recognised ranking agencies by making efforts to identify the criteria employed by such agencies. This effort led to the identification of two options that supposedly impact a conference's ranking. The first one involves ensuring that the set of conference organizers includes researchers that already have an established record in conferences ranked highly. This runs the risk of working against the idea of the existence of researchers specializing on computational creativity and mainly publishing their work in this conference. The second one involves lowering the acceptance rate for papers submitted to the conference. This runs the risk of further reducing the number of experimental papers accepted at the conference.

Under such circumstances, one reaction might be to despair of recovering the original spirit of the conference on computational creativity; another would be to suggest some drastic changes and possibly some sacrifices.

Suggested Remedies

Having identified symptoms and sources of pressure that might be contributing to the observed drift, we undertake the intellectual effort of suggesting some measures that may help reduce the worst issues. These should not be considered as prescriptive but as a listing of possibilities that the community may want to discuss. We propose the following: *development of reviewing rubrics*, *training for new reviewers*, *finetuning the conference format* and *engagement with ranking questions*. The proposals emerge from our experiences and from our observations of how similar dynamics have played out in other communities. We offer them as serious prompts for discussion, while acknowledging that none of them is straightforward to implement.

Develop Reviewing Rubrics As it stands, the conference is lacking rubrics that are specific to paper type and honest about the field's methodology. Ideally, what is wanted is a (possibly controversial) shift in our reviewing philosophy that explicitly makes available in the reviewing templates some of the tenets of computational creativity research that over time have been identified as characteristics of the field. We need suggestions that may allow any reviewer, from whatever discipline, to write a good review: by making the review template guide them towards the type of reviewing they should be doing. If the right questions are included in the review template, even reviewers new to the field can provide valuable feedback, assuming that they have good disposition and that they are willing to consider the indications in the templates. Many conferences have gone this way before us. If we are looking for papers that make the conference *lively*, we should put that as a specific question in the review form. If we value papers that are *provocative*, we should include a question about that explicitly in the review form. If we want the papers to go beyond simply reporting on working systems that undertake supposedly creative tasks, we should include in the form explicit questions that require some kind of evaluation. If we want the evaluation to focus on creativity rather than correctness, we should make it explicit in the forms. If we consider that the most interesting systems are often the hardest to evaluate quantitatively, we should put that as a caveat right next to the explicit questions about evaluation in the review form.

To address the problem of the many different types of paper described in the call for papers, we should develop explicit, concrete rubrics for different paper types that are calibrated to what constitutes good work in each category. For each type, the criteria for "interesting and worth discussing" should be made explicit, so that reviewers are not left to apply default disciplinary heuristics. Further, these rubrics should state clearly, for reviewers and authors alike, that

quantitative evaluation and reproducibility are not universal requirements, rather explain explicitly how they might be detrimental to achieving creativity in a computational system. To summarize, the set of proposed rubrics should clearly describe what constitutes compelling evidence for work that does not lend itself to such "traditional standards".

Train Reviewers Along with suggestions above for overhauling and improving review rubrics, reviewers themselves must also be "overhauled"; they must come to understand that because ICCC is a different sort of conference, a different sort of reviewing is required, one that both safeguards that uniqueness and that ensures quality contributions. First, reviewers must attend the conference; if they do not, it is unlikely that they will ever comprehend what it is that they are tasked with. Second, they must appreciate that the study of (computational) creativity is by definition, and for the many reasons mentioned in previous sections, somewhat unique when compared with traditional STEM fields. Third, they must have an open mind about what makes a valuable contribution and be willing to accept that such a contribution might look very different than what they are used to seeing in other conferences they attend. Accomplishing this level of education is a daunting task, and any solutions must be lightweight, given that conference organizers are overburdened. Senior PC members must be willing to take stronger positions during review discussions—this in itself can be a significant source of training, and meta-reviews should be conducted by an SPC member that did not also perform a normal review of the paper.

Finetune the Programme Format While it is true that the standard conference format is the standard for a reason, it is also true that attendees of ICCC universally agree that all the best discussion/learning/creativity happens outside the formal conference sessions: at coffee breaks, in the halls, over lunch, etc. Presenting and consuming traditional paper talks is probably here to stay and does serves some important purposes; however, we can re-emphasize the uniqueness of ICCC as a meeting place rather than a venue with subtle design tweaks that provide more time for the quality informal discussions that attendees value most: reemphasizing the social program, coffee breaks and communal meals; contemplating new kinds of question/answer formats (what about submitting questions in advance? Or the speaker asking questions for the audience to answer? Or...? Even subtle programme design choices can likely have an outsized impact on how the conference affects the experience of attendees and provides the sense that attending ICCC is valuable *beyond* just gaining a(nother) publication.

Engage with the Ranking Question Rather than either ignoring external evaluation pressures or simply submitting to their logic, the community could invest in investigating what options actually exist. What ranking bodies exist? Are some more qualified for understanding the goals and importance of a conference like ICCC? Are some more amenable to alternative forms of making one's case? In what sort of conference category does ICCC belong? AI, for example, might seem like an obvious suggestion here, but in fact, upon careful introspection, it likely does not share nearly as much of a platform with ICCC as one might initially guess. How can we find a balance between preserving the "fun" and "finesse" of CC research and having the value of that research recognized by agencies whose opinion matters to those outside our community?

Conclusions

We have put forth concerns suggesting that the conference may be drifting away from its original reason for existence. However, we believe the current trend is not inevitable but that it is possible to redirect the future of ICCC so that it can provide a unique, fertile space for discovering where computational creativity can take us. We have tried to condense a set of issues that we have identified over the years, but these are not exclusively ours, and they are not necessarily complete nor agreed upon by everyone. Different interests and sensitivities can improve this proposal and complement it with different views. However, we do believe these concerns speak to real issues that must be openly debated and addressed, even when opposing positions arise.

Other disciplines and communities face similar problems. Balancing identity and content quality against research evaluation agencies is not an issue exclusive to ICCC. However, the very spirit of the conference is achieving creativity (in whatever form) and this strictly *requires* the community to break free from the straitjacket of traditional rubrics for evaluating contribution. If we simply apply traditional norms, we face the risk of becoming a(nother) subordinate venue for machine learning/artificial intelligence research. By definition, creativity lives in the spaces which cannot be simply reached by producing one more output. In order to reach those spaces, we need to leap into the gray areas in which right or wrong, good or bad are not clearly definable.

However, one possible consequence to the various considerations outlined in the paper arises as a possible corollary. Given the conflicting tensions between the expectation of funding agencies and ranking bodies on one hand, and the valuable features that once made the conference on computational creativity a desirable venue for researchers interested in modelling creativity on the other, we may be faced with a future in which:

1. Researchers are encouraged to publish their computationally heavy, metric-driven work in mainstream AI/ML venues
2. ICCC explicitly re-positions itself as a forum for interdisciplinary, high-risk ideas that would not fit elsewhere

Would this be desirable for the conference? Is it tenable in today's academic environment? Might such a dichotomization signal a rebirth for the conference? Or, a death-knell? These are questions we urgently need to consider.

In any case, we believe that the hurdles we now face are not insurmountable. As a community, we have the tools to define ourselves and our practices, be it with clearer goals for the conference or more investment of effort. We need to "renew our vows" and, against the new landscape of artificial intelligence and data science, create a space in which mere numerical evaluation does not constrain our progress.