To appear in *Connection Science*
Vol. 00, No. 00, Month 20XX, 1–35

ARTICLE

Has computational creativity successfully made it 'Beyond the Fence' in musical theatre?

Anna Jordanous*

School of Computing, University of Kent, Chatham Maritime, Medway, Kent, UK

(*Summer 2016*)

A significant test for software is when it is tasked with replicating human performance, as was the case recently with creative software and the commercial project *Beyond the Fence* (undertaken for a television documentary *Computer Says Show*). The remit of the *Beyond the Fence* project was to use computer software as much as possible to produce a musical show billed as 'the world's first computer-generated musical'[1]. Several computational creativity systems were used in the production of this musical, which was performed in London's West End for a two week run in 2016. This paper considers the challenge of evaluating this project. Current computational creativity evaluation methods are ill-suited to evaluating projects involving creative input from multiple systems and people. Following recent inspiration within computational creativity research from interaction design, here the DECIDE evaluation framework is used to evaluate the *Beyond the Fence* project along two questions: (1) To what extent is the project successful? and (2) To what extent does this project demonstrate computational creativity? Evaluation lets us conclude that this was a reasonably successful achievement of the task of using computational generation in producing a credible musical show. Lessons have been learned for future computational creativity projects though, particularly for affording creative software more agency and enabling software to interact with other creative partners. Upon reflection, the DECIDE framework emerges as a useful evaluation 'checklist' (if not a tangible operational methodology) for the evaluation of multiple creative systems participating in a single creative task.

Word count: 17000

Author biography: Dr. Anna Jordanous is a Lecturer in the School of Computing, at the Medway campus of the University of Kent. She is a member of the Computational Intelligence and Data Science research groups. Her research areas include computational creativity and its evaluation, music informatics, digital humanities, knowledge modelling, Semantic Web, and natural language processing. Primarily she works with computational creativity - the modelling, simulation or replication of creative activities and behaviour using computational means - with a focus on the question of how to evaluate claims of computer software being creative. As well as writing creative software to improvise music, Dr Jordanous has contributed a highly-cited standardised procedure for evaluating creative systems. She also uses music information retrieval and natural language processing in her work.

Keywords: computational creativity; musical theatre; beyond the fence; creativity evaluation; human-computer interactive creativity

*Corresponding author. Email: a.k.jordanous@kent.ac.uk

[1] `http://www.wingspanproductions.co.uk/news-and-awards/read/48/Beyond-the-Fence-the-world-s-first-computer-generated-musical` (last accessed August 2016).

1. Introduction

Beyond the Fence (Till et al., 2016) was billed as 'the world's first computer generated musical'.[2] With several computational creativity software packages and computational data analyses providing data, frameworks and content for the musical, in collaboration with (human) musical theatre experts, *Beyond the Fence* tests whether computational creativity can be used to create a musical. This work was undertaken by the television production company Wingspan Productions (led by Archie Baron and Dr Catherine Gale) for a two-part documentary about the process; *Computer Says Show* was commissioned by a UK satellite channel for the arts (Sky Arts) with support from Wellcome Trust. The musical had a two week run in London's West End, an area of London with an extremely vibrant theatre and musical theatre scene (to the extent that this part of London is colloquially referred to as 'Theatreland'). Section 2.1 gives some details about the *Beyond the Fence* project, how it was undertaken and what creative entities have been involved.

In musical theatre, those involved in shows are accustomed to being reviewed. In research, evaluation is a crucial part of how we learn from our efforts. Essentially this paper combines the two: an evaluative review of the 2016 project *Beyond the Fence* as a computational creativity project. The project combines two main outputs: the musical show *Beyond the Fence* (Till et al., 2016) and *Computer Says Show*, a documentary on the process of creating the musical (Gale, Baron, Lomax, & Wingspan Productions, 2016).

Having been involved in *Beyond the Fence* as an informed commentator for the *Computer Says Show* documentary (Gale et al., 2016), but not having been involved in creating any of the software for the musical (Till et al., 2016) itself, I have had the opportunity to gather and discuss information about this project with a variety of different sources, from people behind the software to the cast performing the show. From this perspective, I am interested in whether *Beyond the Fence* was a success? To what extent? And: what contributions does this project make to computational creativity research and knowledge?

Evaluating the creativity of software comes with many challenges, given the subjectivity and the lack of a 'right answer' to be achieved by creative systems. This particular project poses an interesting evaluative challenge which is arguably unique and certainly rare: several stand-alone computational systems contributing parts of a relatively large-scale commercial creative endeavour, with human participants taking key parts of the creative process of creating the musical. Section 2.2 considers the issues arising in evaluating this project, in the context of computational creativity evaluation. Much computational creativity evaluation work has focused recently on developing evaluation methodologies for measuring the success with which computationally 'creative' software actually achieves creativity, and to what extent. These methods tend to focus on aspects of a single system working alone, rather than the combination of several systems working on different parts of a creative task. Hence, evaluation of projects such as *Beyond the Fence* are tricky to tackle with these methods. This project offers a particular set of evaluation issues to consider, as a *co-creativity* project (where the creativity arises from collaboration between creative agents). We do not as yet have an accepted standard evaluative tool for evaluating a project of this complexity (in terms of participants and magnitude), though some proposals have been made towards this. The contribution of

[2] `https://www.theguardian.com/stage/2015/dec/01/beyond-the-fence-computer-generated-musical-greenham-common` (last accessed August 2016) and `http://beyondthefencemusical.com/` (last accessed March 2016). This latter website is now inactive, but the content as of April 2016 can be retrieved from `http://web.archive.org/web/20160409095721/http://beyondthefencemusical.com/` .

this paper, therefore, is greater than merely a routine end-of-project evaluation, simply because of the challenges in evaluating this type of project.

Recent work in computational creativity has suggested interaction design based evaluation strategies as an alternative for evaluating computational creativity work where multiple creative partners (computational and/or human) are involved (Kantosalo, Toivanen, & Toivonen, 2015). Kantosalo et al. proposed and applied the DECIDE evaluative framework (Rogers, Sharp, & Preece, 2011) as a suitable strategy for approaching evaluation of such computational creativity projects as we have with *Beyond the Fence*. Section 2.3 summarises this framework and considers its application in the context of computational creativity evaluation. Following this proposal, the DECIDE framework is used to evaluate *Beyond the Fence*, allowing us to consider in our conclusions the extent to which the *Beyond the Fence* project has been successful (see Section 3). This evaluation exercise also allows us to gather and generate knowledge on how applicable the DECIDE framework is as an evaluation tool for evaluating multi-participant collaborative computational creativity projects. Essentially, as much as we use DECIDE to evaluate *Beyond the Fence*, we also use *Beyond the Fence* as a case study to help us evaluate the DECIDE framework as an evaluative tool for this type of project. As long as we remain mindful of the clear interdependence of both analyses upon each other, various types of methodological and practical gains in knowledge can be made from reflecting upon the content in this paper from these two different perspectives.

Section 3.6.3 concludes the evaluation undertaken in this paper with a discussion of what the field of computational creativity can learn from the *Beyond the Fence* project. Where has computational creativity successfully contributed to *Beyond the Fence*? What has not worked so well in terms of computational creativity's application to this problem of creating a new musical? And where would future work in this direction be most usefully directed?

Section 4 examines this evaluative application of the DECIDE framework as a case study, considering what gains the framework affords and what further information is necessary for DECIDE to be applicable in this way. Much can be learned from the practical application of a theoretical model, as is the case here. The overall appropriateness and applicability of DECIDE is considered in this Section, as a result of the case study application to the task of *Beyond the Fence*.

Hence the contribution of this paper is two-fold; firstly, we examine and evaluate the success of the *Beyond the Fence* project as a computational creativity project and more generally. Secondly, we use the evaluation of *Beyond the Fence* as a case study to examine how useful the DECIDE framework is as an evaluative tool for computational creativity researchers. In concluding (Section 5), the paper summarises the results of evaluation and reflects upon both these perspectives.

2. Background

2.1. *Details of the* Beyond the Fence *project*

The musical *Beyond the Fence* (Till et al., 2016) ran for two weeks, being performed in London's West End in February/March 2016. In total, there were 15 performances, including three preview performances, one gala performance and two matinee (daytime) performances as well as twelve regular evening performances. The plot of the musical was as follows:

> 'September 1982. Mary and her daughter George are celebrating one year of living at the

> Greenham Common peace camp. The group of women they have joined are all committed to stopping the arrival of US cruise missiles through non-violent protest. When Mary is faced with losing her child to the authorities, an unlikely ally is found in US Airman Jim Meadow. How can she continue to do what is best for her daughter while staying true to her ideals? Beyond the Fence is a powerful new musical about hope, defiance, unity and love.' [3]

Another story of interest to computational research is not that portrayed within the end product itself, but the process and interactions that took place between different creative entities to create this musical. This process was driven by a team from Wingspan Productions, who describe the engagement with computational creativity as follows:[4]

> 'The process began with a predictive, big data analysis of success in musical theatre, conducted by Dr James Robert Lloyd, Dr Alex Davies and Prof Sir David Spiegelhalter (Cambridge University). They interrogated everything from cast size, to backdrop, emotional structure to the importance of someone falling [in] love, dying (or both!) - in more and less successful shows - to create a set of constraints to which the musical had to conform, to theoretically optimise chances of success.
>
> Next, the team visited what's known as the What-If Machine at Goldsmiths, University of London. With Prof Simon Colton, Dr Maria Teresa Llano and Dr Rose Hepworth at the helm, the machine generated multiple central premises, featuring key characters, for the new show. The team selected this as the starting point and the original idea for the musical: What if a wounded soldier had to learn how to understand a child in order to find true love?
>
> A plot structure for the musical was also generated computationally, thanks to work led by Dr Pablo Gervás (Complutense University of Madrid). A brand new analysis of musical theatre narratives enabled him to adapt an existing story telling computer system, called PropperWryter, to turn its hand to musicals and build the core narrative arc of the new show.
>
> Taken together, all of the above enabled the precinct for the emerging story to be identified: Greenham Common. The team then wrote a book and lyrics (with the assistance of some other computational tools) that fitted all these constraints.
>
> Finally, the music material has been provided by Dr Nick Collins (Durham University), who has created a computer composition system based on a machine listening analysis of musical theatre music, conducted by Dr Bob Sturm (QMUL) and Dr Tillman Weyde (City University). Additional computer music material [was] generated using the FlowComposer system created by Dr Pierre Roy and Dr Francois Pachet (SonyCSL, Paris).'[5]

In the credits for the musical, the 'creative team' listing includes the software programs involved and key researchers on each piece of software, plus two human musical theatre experts (Benjamin Till and Nathan Taylor), who curated the software outputs into its final musical format. The accompanying documentary shows how human members of the creative team took care to adhere to the spirit of this project: using as much computer-generated material as possible even when this caused difficulties. 'Ground rules' were circulated across the team involved, setting out how situations should be handled to keep the project on track (see Appendix A). Every member of the team promised to adhere to these ground rules and more generally to the principle of the project of using as much computer-generated material as possible. As Neil Laidlaw (producer of the stage

[3] `http://beyondthefencemusical.com/about-the-show` (last accessed March 2016, now retrievable via `http://web.archive.org/web/20160418004951/http://beyondthefencemusical.com/about-the-show`).

[4] The teams involved have published fuller details of how Beyond the Fence was constructed (Colton et al., 2016). This current paper is not intended to duplicate these descriptions, but to critique the project from an independent perspective. Please consult Colton et al. (2016) for further details about how the project was realised computationally.

[5] `http://beyondthefencemusical.com/the-science` (last accessed March 2016, now retrievable via `http://web.archive.org/web/20160418004951/http://beyondthefencemusical.com/the-science`).

performance) says, 'we have to honour what we've signed up for' (Gale et al., 2016).

2.2. *Evaluation issues*

The purpose of this paper is to evaluate to what extent *computational creativity* has been applied to realise the *Beyond the Fence* project. Practical issues in the *Beyond the Fence* project pose interesting challenges in terms of evaluation of the creativity of the software involved. Multiple software was used during this project, as well as interactions with creative people. Hence we could either evaluate the creativity of each constituent software-based part of the creative team, or focus on evaluating the overall collective project. As this paper's aim is to evaluate the multiple parts of the project as a whole, it focuses on the latter aim; the task of evaluating individual software falls better to papers that report on individual software.

Though we now have several evaluation methods/frameworks available for the evaluation computational creativity software (Jordanous, 2017; Ritchie, 2017), it is difficult to apply these to the evaluation of *Beyond the Fence*. This is because the methods and frameworks resulting from research into computational creativity evaluation tend to focus on the evaluation of a single system, rather than the combination of several systems with human intervention.

The Standardised Procedure for Evaluating Creative Systems (SPECS) (Jordanous, 2012b) is close to being applicable to evaluate *Beyond the Fence*; it asks researchers to (1) identify a characterisation of creativity by which to evaluate creative systems (2) derive standards or benchmarks by which to measure our systems, and (3) devise suitable tests to evaluate our systems against these benchmarks. For our base characterisation of creativity for Step 1, we could use the 14 components of creativity representing a general basis for creativity as derived in Jordanous and Keller (2016). The task of creating a musical, however, combines many different domains of creativity including music, book (the script and lines), plot, lyrics, story-telling. Different types of creativity appear to prioritise different aspects of creativity as important (Jordanous, 2012a; Plucker & Beghetto, 2004); for example in musical improvisation creativity the ability to communicate and interact socially is considered much more important for creativity than arriving at perfect musical results (Jordanous & Keller, 2012). Given this diversity of creative tasks involved in musical theatre, it would be difficult to use the fourteen components of creativity to accurately represent creativity across all the creative tasks involved in the musical. We could alternatively use as a base model for Step 1 the several musical theatre elements that were identified from the big data analysis of features of successful musicals performed by the Cambridge team (see Section 2.1 and Colton et al. (2016)). Practically, though, these results should not be used for evaluation as (a) these results were directly used to influence the creative process so their use in evaluation would skew results; and (b) the big data analysis measured features associated with high *quality* musicals rather than highly *creative* musicals - though creativity and quality have a high degree of overlap, they are different concepts (Jordanous, 2011). It seems that tackling the question of what is the most faithful model of creativity in this particular scenario would be interesting - and challenging - work to carry out as a follow-up project in its own right, though possibly with no single answer to reward the researchers.

The FACE model (Colton, Charnley, & Pease, 2011) offers an approach for analysing creative systems on their ability to use and/or generate new methods for using Frames (natural language descriptions of what the software has done), Aesthetics (measures of aesthetic quality), Concepts (underlying theory/ies that guide the creative process) and

Example outputs. This is very focused on a single system though, whereas we wish to evaluate the entire project (a combination of systems' and people's creative work).

The Creative Tripod (Colton, 2008) asks whether the system under question can be considered as a candidate for a creative system, through identifying the system's ability to demonstrate skill, imagination and appreciation. Again, though, this would vary between individual systems and we would not get much depth of feedback from applying this framework to the combination of systems.

Ritchie has also devised empirical criteria for evaluating creative systems (Ritchie, 2007), though the criteria are again less applicable for this evaluation as they are based upon ratings of the typicality and value of multiple artefacts produced by a creative system, whereas we only have one output to judge. They also give only very limited options for evaluating the creativity of the *process* as well as the end *product* (Jordanous, 2012a, 2016).

2.3. *Evaluation of co-creativity*

Considering the interactive nature of the *Beyond the Fence* project, it is difficult to generalise our currently available models across multiple systems acting collaboratively (co-creativity). There is also the question of how to include human members of the creative team; this is important to include, given the extensive curation, editing and extra creative work that people did to contribute to this project. These issues make the above-mentioned existing evaluation methodologies and methods difficult to apply in this evaluation.

In collaborative, co-creative scenarios, Maher has investigated how to tackle the question 'Who's being creative?' (Maher, 2012). She points out that creative responsibility can be assigned either to individuals collaborating in the process, or the set of entities as a collective whole. She proposes categorising and ordering co-creative systems via levels of ideation and interaction. Ideation is measured in two dimensions, ranging from modelling to generating on one axis, and on the other axis, from abilities to suggest ideas, through to enhancement and generation of ideas. Interaction is similarly categorised by the number of entities involved, human and computational.

Links can be seen between Maher's proposals and Bown's later commentary on the need to recognise the contributions of different agents in the creative process (Bown, 2015). Bown proposes breaking down the creative process into dynamic activities over the course of the creative process, measuring for each activity the involvement of or influence over different creative entities.

Both Maher's and Bown's proposals are well thought through and useful as summative attributions of creativity to participants in the creative process, but are affected by two main problems. Firstly, neither approach acknowledges that we may be biased in attributing or describing creative abilities to computational agents, compared to how we do this for people, nor how to circumvent ethical issues. Secondly, both approaches are relatively lightweight and only give us certain information on ideation and interaction, rather than a broader picture

A more complex approach has been proposed that uses mixed methods for evaluation of computational collaborative creativity (Kantosalo et al., 2015), with human opinion a key part of analysis. Following from inspiration from prior work considering co-creativity as strongly related to Interaction Design (Bown, 2014), Kantosalo et al. have investigated (favourably) the suitability of cross-applying Interaction Design research methods and frameworks to computational creativity research methods (Bown, 2014). Following this inspiration, Kantosalo et al. proposed and evaluated the DECIDE framework (Rogers et al., 2011) as a suitable framework for approaching contextual evaluation of several systems collaborating and being co-creative with humans; this is the scenario we have with the creative systems used for *Beyond the Fence*.

The DECIDE framework employs the following steps (not necessarily in a linear or one-directional manner):

(1) 'Determine the goals
(2) Explore the questions
(3) Choose the evaluation methods
(4) Identify the practical issues
(5) Decide how to deal with the ethical issues
(6) Evaluate, analyze, interpret, and present the data' (Rogers et al., 2011)

DECIDE offers a framework within which evaluation can be planned and carried out in a systematic way. Using DECIDE, Kantosalo et al. (2015) evaluated several co-creativity research projects (creativity research projects that involve both human and computational participants in the creative task), including one system of their own. The feedback they retrieved from this analysis was useful in comparing systems' success, 'providing results that can be put to practical development use' (Kantosalo et al., 2015, p. 282).

Thus DECIDE is an evaluation approach for collaborative creativity scenarios which enables considerable detail, breadth and focus in evaluating computational creativity collaborators. Currently it is under-explored in computational creativity outside of Kantosalo et al's work, but given the reasoning above, it is worth further exploration. Hence it is adopted for evaluation in this present paper, and the paper takes on a second purpose in addition to evaluating the *Beyond the Fence* project: that of a case study to evaluate DECIDE as a suitable framework for evaluation of computational collaborative creativity.

3. Evaluation using the DECIDE framework

To what extent is *Beyond the Fence* successful and how has it engaged with computational creativity research? To evaluate these questions, we consider *Beyond the Fence* as a case of multiple creative systems and people generating a musical. Evaluation models typically focus on evaluating a single system, as we saw in Section 2.2, however recently Kantosalo et al. (2015) has proposed and investigated the DECIDE framework (Rogers et al., 2011) as an appropriate framework for evaluating multiple co-creative systems.

Following Kantosalo et al. (2015), therefore, here the DECIDE framework is now applied to *Beyond the Fence* to evaluate as a whole the various creative entities involved in this project. Each subsection in Section 3 reports one step of the DECIDE framework being carried out to evaluate *Beyond the Fence*.

3.1. *DECIDE-1: Determine the goals*

This step of DECIDE requires the evaluator to *determine the goals* of evaluation, which are to question the extent to which the project is successful. As part of this step, it is useful to consider the goals of the overall project, and see if those have been achieved. Such considerations give us more details on what it means for the project to have been successful, to inform the next steps of the DECIDE analysis.

The Wingspan Productions team conducted this project to explore if a 'computer-generated musical' was possible. Specifically, their goal was to create and stage a musical generated in collaboration between creative software and human musical theatre experts. As Archie Baron, Creative Director of Wingspan Productions, put it:

> 'We pretty much pitched the experiment and its end product the musical from the very first contact with the broadcaster as an attempt to provide a precinct for a fascinating and important debate about man, machine, creativity etc. ... We wanted to see whether it was possible and to create and stage it precisely in order that we could provoke the wider discussion. ... the main criterion for success for us as producers and Sky as broadcasters ... broadly is to make a mark with a project that is original, innovative, thought-provoking, entertaining, genre-busting etc. It falls into the category TV folk talk about as 'reputational' (i.e. it's not about viewer numbers but about television's impact and cultural agency). '
> (Baron, 2016, personal communications)

The Wingspan Productions team also maintained throughout the project an interest in what 'being creative' actually means and how creativity might emerge from rules and be assessed, given the difficulties we have in assessing creativity in humans. One particular debate that Dr Catherine Gale (from Wingspan Productions) probed during her conversations with computational creativity researchers is one that relates these thoughts directly back to *Beyond the Fence*: does a generative process have to result in a good quality product (i.e. the musical) to qualify that process as having been creative? This distinction between the process and the generated artefact when assessing/recognising creativity has often arisen in computational creativity research (Gervas, 2009; Jordanous, 2016, for example). For *Beyond the Fence*, then, can the project be considered successful even if the generated show (Till et al., 2016) is not well received as a musical in its own right? To some extent, it can: another subgoal of the project was to engage with research in creative software as much as possible. This is a two-way process; as well as learning from computational creativity research and using software in the musical's production, the Wingspan team hoped the project would contribute back to research knowledge.

It should be noted that different members of the team involved in *Beyond the Fence* would have had differing goals. For the musical theatre experts in the team, those directly involved in making the musical happen, they have a vested interest in the musical being a success commercially, a hit with audiences, and culturally, in terms of awards and other recognition. As key figures in musical theatre, they are investing their reputations in the project (as much as the teams behind the creative software are investing their software's evaluation in a successful end product). However, all (human) members of the *Beyond the Fence* team agreed to abide by 12 ground rules (see Appendix A) determining how compromises between the quality of the end product and the 'purity' of the computer-generated aspect of the project should be handled.[6] Typically, the process should involve the use of as much computer-generated material as possible, within constraints of available time, with minimal intervention from people.

In this first step of DECIDE, however, we are searching for the goal of the DECIDE evaluation, so will now return to consider this. The evaluation process itself also introduces goals. One criterion of what makes a good methodology for evaluating creative systems (Jordanous, 2014) is the usefulness of feedback. As we have seen, the *Beyond the Fence* project comprises multiple software, each tackling different types of creative task. This illustrates the vast and varied scope of creativity in musical theatre. For this current work, this evaluation is intended to uncover formative feedback for future development and recognition of this work's contribution to research.

[6] Section 3 raises in step DECIDE-5 the associated ethical issues with such compromises.

In short, the two-fold goal of this current evaluation is to **determine to what extent the project was successful (or not) in meeting its aims, and to seek feedback on how the project could increase its successfulness**.

3.2. *DECIDE-2: Explore the questions*

This second step of the DECIDE framework asks us to *explore the questions* that could be asked to achieve the goal asserted, and decide upon what specific questions our evaluation will answer.

For this project, what will constitute a success? The question of whether there was an underlying hypothesis or research question for the project was raised in a question-and-answer session post-performance I took part in with Catherine Gale (Wingspan Productions), Bob Sturm and Benjamin Till (respectively representing computational and human parts of the music generation team) on 2nd March 2016. This discussion revealed some difficulty in pinning down an exact single scientific hypothesis or research question for the project (one by which the project success could be tested against and/or measured).

Various questions have driven the project, though, some as a higher priority than others. It is worth considering what questions came up during the process of doing the project. As part of this project, the Wingspan Productions team sought out and engaged with several leading computational creativity researchers, as described above. During this process, an extensive number of questions arose around various debates and issues around computational creativity, and how the Beyond the Fence project sits in the wider context of computational creativity. For example (Gale, 2016, personal communications):

(1) 'How has our attitude towards how we use computers changed in recent decades?
(2) Why do people develop machines that are creative?
(3) So is it right to paint a portrait of machines as a young artist? One that is maturing?
(4) What kind of systems have been/are being developed?
(5) What different approaches do people take? 'Heroic' methods where computer is [an] artist in its own right, or more collaborative approaches?'[7]

The *Computer Says Show* documentary (Gale et al., 2016) explores many of these questions in its reporting of the *Beyond the Fence* project.

This current work, though, is to evaluate whether the overall project has been successful in achieving its goal, and to what degree. Returning briefly to the discussions in step DECIDE-1, we see that the key aspects of success of this project revolve around two questions: (1) whether (and to what extent) the project was successful in achieving its goal of creating and staging a computer-generated musical; and (2) how well the project has engaged with and contributed to computational creativity. So we can settle on two evaluation questions, which will give us useful detailed feedback towards learning from this project:

(1) **To what extent was *Beyond the Fence* successful as a computer-generated musical?**
(2) **How well has the project engaged with and contributed to computational creativity research?**

[7] Probably inspired by d'Inverno and McCormack (2015).

3.3. *DECIDE-3: Choose the evaluation methods*

In the previous step, DECIDE-2, two evaluation questions were determined. The first [the project producing a computer-generated musical] we can investigate with the aid of traditional review metrics and methods for musical theatre in step DECIDE-3. The second [engagement with research into creative software] we can investigate by reflecting on our data for the first question alongside study of interactions between Wingspan Productions and researchers. This next step, DECIDE-3, asks us to determine what evaluation methods we should use to seek data and answers for our evaluative questions.

Of these, methods and metrics for the first evaluation question need a little more investigation, to identify existing metrics for evaluation of success in musical theatre. Within the DECIDE framework, Rogers et al. advocate that a combinatory approach, using multiple methods, allows for a broader and richer evaluation. Such a mixed-methods approach was applied to good effect by Kantosalo et al. (2015), generating various complementary types of feedback for their main case study.

How do we identify a successful musical? As reported in Section 2.1, the big data analysis by the Cambridge team distinguished between characteristics of 'hits' (culturally and commercially successful), 'flops' (culturally and commercially unsuccessful), 'critically acclaimed' musicals (successful culturally but not commercially) and 'crowd pleasers' (successful commercially but not culturally). Therefore, we should aim to evaluate both commercial and cultural success. Commercial success is relatively easy to gauge after a show has been performed for a length of time: for example by looking at the amount of money a musical makes, the length of its run, the number of tickets sold and any touring the show does. Typically, cultural value is more subjective and so trickier to measure objectively. While people involved in the project have reported that they felt the show was a success (Gale et al., 2016) (and Taylor, 2016, personal communications), empirical metrics are possible (Jordanous, Allington, & Dueck, 2015). For musical theatre, cultural value can be measured via awards, press attention, reviews, audience reaction questionnaires, location of show venue, funding, pick-up by amateur companies, influence on other shows,[8] and other non-commercially measurable aspects.

Another critical part of this project's success is the extent to which it can be considered a 'computer-generated' musical. As Wingspan's ground rules for this project dictate (see Appendix A), the aim was to maximise the amount of computer-contributed material used in the musical (within the time and practical constraints of the project). Another ground rule for this project was (again within practical limitations) to document and keep track of sources used, actions taken and decisions made. The intention, according to the ground rules, was to report these as faithfully and as fully as possible in the resulting documentary (Gale et al., 2016) and in other news outlets/publications such as media articles and academic publications such as Collins (2016); Colton et al. (2016). We can use these sources to trace the extent to which the musical was computer-generated.

It should also be remembered that we are evaluating the overall project, not just the show itself. Some objective data that might be useful in evaluating the success of the project is to study the success of the *Computer Says Show* documentary (Gale et al., 2016), which was Wingspan Productions' main target output during the project's lifetime. Again, there are commercial and cultural metrics for success of this documentary. The main commercial metric available is the number of people who saw the documentary; cultural metrics include awards and award nominations for the documentary, and discussion of the documentary in other sources (such as media articles and social media).

[8] The latter two aspects, pick-up by amateur companies and influence on other shows, are difficult to measure at this point in time, as discussed in step DECIDE-4.

As will be discussed in step DECIDE-4, viewing figures are unfortunately not available for the documentary, but we can trace how the project was discussed in media articles and on some social media.

As mentioned above, for our second evaluation question investigating how the project has engaged with and contributed to computational creativity research, this is less straightforward to answer using data, as we are understanding more subjective interactions and lessons learned. Hence, addressing the second evaluation question will require us to take a reflective approach, examining in what ways the project has interacted with computational creativity research: how it has influenced research and how it has been influenced by research. We can use this reflection to consider what activities and knowledge exchanges have been useful, and in what ways the project could better engage with computational creativity research.

We have now identified several ways in which we can collect data, evidence and reflections towards our two evaluation questions. The next step is to consider any practical issues which affect evaluation.

3.4. *DECIDE-4: Identify the practical issues*

Before proceeding with evaluation, the DECIDE framework requires us to *identify the practical issues* that may affect our evaluation, and consider how to deal with them.

Data availability is a major practical issue, manifesting itself in various ways. Practically, one issue is that the show is no longer being performed; its run ended in early March 2016. Any evaluative data needed must therefore already be collected in a timely fashion, or must not depend on being collected during a show's performance. Some data will not be available due to webpages being out-of-date (such as the `http://www.beyondthefence.com` site, which is now no longer available), though some of this content can still be accessed using web archiving solutions such as the Internet Archive (`http://www.archive.org`). Another timeliness issue is the difficulty of measuring cultural value indications such as whether the show is adopted for performance by amateur companies, and whether the show has influence on other shows. These are difficult to measure at this point in time, as the musical has probably not been in existence for long enough for these two aspects to have had time to develop and take place. Realistically, these indicators of cultural value would not be fully apparent for a matter of years after the show's original dates, so are not taken into account for this present evaluation.

Other data are inaccessible (at least in part) due to privacy restrictions. For example in terms of social media, much content on sites such as Facebook will not be accessible due to privacy restrictions, but sources such as Twitter are typically open data and hence better sources to mine data from.

Other data are not made available by the appropriate sources; a particular example of this are the data relating to viewing figures for the Sky Arts documentary (Gale et al., 2016). Archie Baron (Wingspan Productions) confirmed that:

> '[Viewing figures] aren't available (to us or to you). Sky don't release them. Its commercial model is based on subscription of customers and it chooses not to disseminate overnights. Sky Arts content is also frequently repeated and available on catch-up (and in curated seasons on demand) so it argues that figures are meaningless.'
> (Baron 2016, personal communications)

Another issue is that the show's context differed from a commercial musical in a number of respects, which affect how it can be evaluated as a piece of musical theatre. We can still use the same metrics as for traditional musicals - for example, the 'ground rules' governing the show's processes (Appendix A) stressed that 'we have to sell tickets' - but should do this in the context that they may be less helpful/less available in terms of evaluating the musical as a show in its own right. As Archie Baron said:

> 'though it was modelled on commercial musicals in terms of content and form- it certainly wasn't a commercial musical in most other respects. There was almost no significant advertising or paid marketing - unlike for most musicals of this scale. Sky and our investment only sustained to a two-week run with little or no word-of-mouth effect possible once we'd opened because of the very short run (so we knew from the outset it wasn't a strictly commercial proposition given that e.g. actors were engaged for 7 weeks in total). Above all, the fact that its experimental/computer-derived genesis was its main PR angle rather skews how and why audiences came to it.[9]'
> (Baron 2016, personal communications)

One thing to stay mindful of, in evaluation, is that it is difficult to remove effects of bias when evaluating computational creativity; typically people evaluate output of creative software differently to the output of creative people (Jordanous, 2012a; Lamb, Brown, & Clarke, 2015; Moffat & Kelly, 2006; Pasquier, Burnett, & Maxwell, 2016). This is an issue that should be accommodated when interpreting data during analysis, where possible.

3.5. *DECIDE-5: Decide how to deal with the ethical issues*

In this step we must identify and acknowledge ethical issues relevant to our evaluation, and *make decisions on how to deal with these ethical issues.*

Ethical issues exist around the project itself and also around computational creativity more generally. Taking the latter type of issues first, in the context of the project, Gale (2016, personal communications) has reflected on whether computers could (and should) be considered creative entities, at a level which is comparable or equal to humans. She discussed with various computational creativity researchers what might make people working in the creative industries more receptive or better suited to collaborate with computers than others, and questioned how to talk to creatives about their attitudes, perceptions and potential biases towards working with computers. Certainly this latter question deserves greater attention if computational creativity research is to reach a wider audience and broader range of collaborative partners. Some interesting points have already been raised which we can build upon such as how people may perceive the creativity of a system by looking for key aspects that a system should generate if it might be described as 'creative' (Colton, 2008); and the role of people's reactions on interacting with creative systems, as a key contribution to that system's creativity (Gervas, 2009; Gervás & León, 2014; Jordanous, 2016). Computational creativity researchers could also consider to what extent it is reasonable (or productive?) to attribute creative agency to a computer when featuring computational creativity software in public engagement activities, following discussions on creative agency and creative responsibilities (Bown, 2015; Johnson, 2014; Maher, 2012).

Ethically, there are also challenges for people in experiencing the *Beyond the Fence* musical (Till et al., 2016) as a computer-generated artefact - as we saw, many of the reviews mentioned a feeling of disconnect at times - something missing from the experi-

[9] It certainly cannot be denied that the main angle for advertising the musical was the experiment behind it. Upon leaving one showing of the musical that had included a post-show panel Q&A discussion, however, I overheard one audience member remarking that he had not realised the computer-generated aspect of the show until staying behind to hear the panel discussion. This is of course anecdotal evidence only, but interesting to footnote.

ence. This was discussed in the previous section. Another ethical issue relates not to the human participants in the creative team, but the computational participants. Is it fair to test computational systems at a professional level, where they are being required to generate material at a standard which it takes humans years to reach (founded upon decades and more of human experience more generally?) Typically, evaluation of computational creativity systems has been undertaken on a more controlled and less professional level, away from the public eye - although there are notable exceptions to this as exemplified by the Painting Fool (Colton, 2012) or the Unnatural Selection (Eigenfeldt, 2015), both of which have recently 'participated' in public professional displays of their creativity (exhibitions and concerts respectively).

One other issue relating to the computational participants is the level at which we are evaluating them. Each software takes on a creative task assigned to it, but here we judge the overall success of the project rather than the success of each task. What we do not consider is the success to which individual tasks have been identified, and to what extent the software fulfils any original requirements. If poor decisions have been made and a vital part of the task overlooked, the computational participants may be judged more harshly as a result, even though they were not asked to perform this missing part of the task. The global focus of this current analysis can only touch on individual system issues; we leave more detailed analyses of success at individual tasks up to the researchers behind the systems involved. Here we focus on what we can learn from the project as a whole.

Archie Baron from Wingspan Productions highlighted in personal communications another set of significant ethical issues, around how the human participants in the project dealt with issues working with computational participants:

> 'it was very thorny ... [we had] ground rules by which we attempted to define and hold on track the project. It was contentious and difficult - but everyone signed up to them ... but in ways you saw in the docs and could no doubt detect from the Q&A etc at the theatre, the process was fairly fraught and traumatic.'
> (Baron 2016, personal communications)

Several of the human participants in the project were musical theatre professionals; although the show was billed as being computer-generated, in reality the success of the show would have influence over many of the human participants' professional reputations. Particularly if the show was not successful, this lack of success could easily be associated with the people involved. As mentioned above in step DECIDE-1, therefore, different people within the project had different personal goals for the project, but had to reconcile their personal goals with the overall 'ground rules' (see Appendix A) they were asked to abide by. Baron (2016, personal communications) expressed an interest in evaluating these psychological aspects of the show in future work, as discussed in Section 5.2 below. The project ground rules openly acknowledged such potential conflicts and advocated a 'respectful' and 'honest' approach to dealing with these conflicts, making these experiences part of the project's research process and documenting/reporting them (particularly in Gale et al. (2016)):

> 'Where this process is liberating / painful / humorous / surprising / impossible / scary / cruel / disturbing we record that honestly and respectfully. We want to experience what it is for artists to be confronted with these new tools and honestly engage with them.'
> *Beyond the Fence* ground rules, see Appendix A.

3.6. *DECIDE-6: Evaluate, analyze, interpret, and present the data*

We are now in a position, in step DECIDE-6, to implement the evaluation that has been designed in steps DECIDE-1 to DECIDE-5. We analyse the evaluative data, interpret the results and present our findings in this subsection. This is the final step in the DECIDE framework, concluding our evaluation.

Previously we identified the goal of evaluating the success of the musical, and identified these two evaluation questions to address:

(1) **To what extent was *Beyond the Fence* successful as a computer-generated musical?**
(2) **How well has the project engaged with and contributed to computational creativity research?**

During this application of the DECIDE framework for evaluation *Beyond the Fence*, various metrics have been identified for how to address these evaluation questions, as well as practical and ethical issues to keep in mind. Now, therefore, we perform the actual evaluation activities.

As outlined above, to address Q1, we consider the cultural and commercial value of the *Beyond the Fence musical*, considering both the success of the show itself as a piece of musical theatre and by the contributions made to computational creativity. We then reflect on the interaction between the project and computational creativity research, in addressing Q2.

3.6.1. Evaluating Q1: Commercial value

In terms of attendance, a total of 3047 audience members[10] saw BTF across 15 shows (12 regular evening performances, 3 of which were preview performances and occurred before the show had officially opened, 1 of which was the gala and 2 of which were matinees). The total capacity of the venue over the 15 shows was 5173, meaning that the musical achieved 59% of capacity across the show's run.[11]

Financial data on commercial benchmarks for success of *Beyond the Fence* (Till et al., 2016) is not yet available. The lack of such data for commercial metrics of success, however, is not an issue that Archie Baron (Creative Director of Wingspan Productions) sees as a problem:

> 'though it [Beyond the Fence] was modelled on commercial musicals in terms of content and form- it certainly wasn't a commercial musical in most other respects. ... There was almost no significant advertising or paid marketing - unlike for most musicals of this scale ... the fact that its experimental / computer-derived genesis was its main PR angle rather skews how and why audiences came to it.'
> (Baron, 2016, personal communications)

[10] This figure refers to the sum of audience numbers attending across each show. So if a person attended the show twice on two different shows, they would be counted as 2 audience members for the purpose of these figures. As this figure is generated from ticket-related metrics, though, it is not possible to detect how many distinct individuals saw the musical.

[11] From informal feedback, one of the creative team reported that he felt the show was getting good audiences every night, which he had been able to observe since he had attended every show to date (due to having different guests coming to see every show - a mark of cultural interest in its own right). The creative team member reported they were happy about this observation, particularly as the show was being performed in a London 'West End' venue.

Table 1. The results of a poll of audience members on the last three performances of the *Beyond the Fence* musical, in answer to the poll question: 'Please rate your enjoyment of the show from 1 to 5 where 1 is low and 5 high'. Poll carried out by Wingspan Productions; data provided by Archie Baron, Wingspan Productions (Baron 2016, personal communications).

Score	# of respondents giving that score	% of respondents giving that score
1	1	1.7%
2	1	1.7%
3	6	10.3%
4	10	17.3%
5	40	69.0%

3.6.2. Evaluating Q1: Cultural value

In terms of cultural value, the timing of this paper also means it is not yet possible to fully reflect on awards, influence on other shows or whether the show is picked up by amateur companies. However the project has thus far attracted some attention in terms of awards. As a piece of musical theatre it is nominated for the 'Most Underrated West End Show' category in the *WestEnd Wilma* awards, with the show's female lead C.J. Johnson being nominated in the 'Rising Star' category of the same awards as a result of her performances in Beyond the Fence.[12] The *Computer Says Show* documentary around the project have been selected for screening and awards nomination in the *51st International Festival of Science Documentary Films Academia Film Olomouc*,[13] a leading European Science Film Festival. It was also nominated for a *Banff Rockie Award* in their Science and Technology category,[14] though it was unsuccessful in achieving this award.

Audience appreciation of the (final version of the) show was measured via audience polls during the last three performances. From an audience comprising 490 people in total for these three performances, 58 audience members (11.8%) filled out and returned a ratings form with their answer to the question 'Please rate your enjoyment of the show from 1 to 5 where 1 is low and 5 high'. Table 1 shows the results. Those audience members that participated in the poll typically enjoyed the show to a large extent: 50 out of 58 people gave the show 4 or 5, with only two of the 58 giving the show a score below 3 for enjoyment.

As a project it has also received media interest beyond the typical review interest, for example: its inclusion in the 2016 Edinburgh Digital Festival, being screened and being included in a discussion looking at the application of AI and its implications; a discussion of the project in the BBC Breakfast (the main morning programme of the British national broadcaster, the BBC); and the inclusion of speakers from the Beyond the Fence team in a 'Digital Conversations' evening panel event entitled 'Portrait of the Machine as a Young Artist', organised and hosted by the British Library. Also, at the 2016 International Conference on Computational Creativity (the key international conference in computational creativity), there was a special session of papers and a panel discussion on *Beyond the Fence*, including panellists from musical theatre, Wingspan Productions, computational creativity and psychology research.[15] There has also been some discussion

[12] `http://www.westendwilma.com/nominees-announced-for-the-2016-west-end-wilma-awards/`, last accessed August 2016.

[13] `http://www.afo.cz/index.php?seo_url=official-selection`, last accessed August 2016.

[14] `http://banffmediafestival.com/rockie-awards/program-competition/`, last accessed August 2016.

[15] See `http://www.computationalcreativity.net/iccc2016/program/`, last accessed August 2016.

on a mailing list for computational creativity research.[16]

The show was performed in the Arts Theatre, London, a reasonably well-known independent theatre in a part of London strongly connected to theatre (colloquially known as 'Theatreland'). The show and the wider project received support from a number of avenues. The overall project was supported by the Sky Arts television channel as well as Wellcome Trust funding. The viewing figures for the documentary have not been released by Sky but Archie Baron from Wingspan Productions remarked in personal communications that in his opinion Sky Arts' aim in this project was reputational-based rather than commercially-based: 'signalling loudly that they're involved in original, genre-busting, innovative arts commissioning' (Baron, 2016, personal communications). From early in the project, Phil Edgar-Jones, Director of the Sky Arts Channel, was very positive:

> 'This is a fascinating project that we're extremely proud to be a part of. At Sky Arts, we're always excited by innovation and this venture offers an intriguing glimpse into how technology is changing music evolution. Can an algorithm create music with all the humanity, emotion and drama that a person can bring? This question captivates us. We cannot wait to see the result.'
> (`http://www.wingspanproductions.co.uk/news-and-awards/read/48/Beyond-the-Fence-the-world-s-first-computer-generated-musical`, last accessed August 2016).

An exhaustive search of Google results towards the end of the musical's run reveals reasonably extensive national/local press attention in the form of 20 reviews, including national newspapers, specialist musical theatre sources, general events listings websites and one technical blog. Most reviews were published directly after the gala performance on February 26th 2016, this (in keeping with traditional musical theatre practice) being the performance that reviewers are invited to attend. Data was also sourced from comments on social media, specifically Twitter. Tweets were mined and analysed using the Twitter R package and manual inspection. After inspecting several accounts related to people involved in the project, various search strings were identified for tweets relevant to evaluating the project: 'BeyondTheFence' (mostly referencing the account @BeyondTheFence_), 'ComputerSaysShow', 'ComputerMusical' and 'AndroidLloydWebber' (referencing hashtags that emerged during the project, the latter referring to the original name for one of the pieces of musical software). Tweets were limited to those that were posted while or after the show was being performed (i.e. from February 22nd onwards), to attempt to restrict tweets only to those evaluating the actual show as opposed to those giving opinions pre-empting the show or marketing tweets. Each set of tweets was manually inspected to filter the set to only those tweets giving some opinion about the project, rather than, for example, personal comments from performers in the show not relevant for an evaluative purpose, or marketing tweets. Tweets were mined using case-insensitive search terms with the @ or # character removed from the start of the search string (for example, to catch situations where tweets mentioned #BeyondTheFence instead of @BeyondTheFence_). The set of tweets remaining was 287 tweets; small enough to be analysed via manual inspection as well as automated methods, as described next.[17] The remainder of this current sub section focuses on an analysis of key issues mentioned in these reviews and Twitter comments.

The textual data from these sources were analysed using both quantitative and qualitative methods. Firstly, they were analysed in terms of what words are used most

[16] `https://groups.google.com/forum/#!topic/computational-creativity-forum/WUxNbLxFXg0`, last accessed August 2016.

[17] The Twitter data set is available at `https://osf.io/exk38/`, under DOI 10.17605/OSF.IO/EXK38.

frequently in reviews, and in tweets about the show. Then sentiment analysis was performed on the shorter texts of the tweets, to see whether the tweets were showing mostly positive sentiment, or being negative about the project, or being neutral in opinion. This analysis was done using the NLTK sentiment analysis API hosted at `http://text-processing.com/api/sentiment/` that implements Pang and Lee (2005), as trained on twitter data and movie reviews. Finally, a qualitative coding analysis inspired by grounded theory was used to identify key reoccurring themes in the tweets, via an iterative process of reading the tweet data, coding them, then classifying the codes into broader categories. Data from these mixed methods is presented in respective order below.

Firstly, word frequency data can be examined to give an impression of common language used in the tweets and reviews. Given that these two types of data are quite different in terms of character length restriction and resulting influences on general vocabulary used, word frequencies are analysed separately for the two sets of data.

Figure 2 shows the most frequently-occurring words seen in the review corpus. A large proportion of these words relate to the content of the musical, as seen in reviews of more typical musicals. Reviews contain many comments relating to work by humans in this musical, such as the strong cast. However, words such as 'computer' and 'experiment' in this word cloud illustrate that these reviewers are well aware of the computational origins of *Beyond the Fence*; several interesting points are examined:[18]

Figure 1. FIG 2 to be placed near here. CAPTION: Content words most frequently occurring in 20 reviews of *Beyond the Fence* (using `http://www.wordle.com`). Common English stopwords and 'beyond', 'fence' and 'musical' are removed.

Figure 3 shows the most frequently-occurring words seen in the relevant tweets about the musical. In this word cloud overview, the comments seem very positive, with words such as 'brilliant', 'great', 'interesting' and 'fantastic' appearing prominently. As observed from manual inspection, a large number of these tweets refer to people involved in the musical; a large number of the tweets seem to be directed positively at individuals or groups of people and their work, suggesting that the tweets come from these people's friends or acquaintances. However, again the prominence of the word 'computer' shows that tweeters, like reviewers, are typically also mindful of *Beyond the Fence* being a computer-generated musical.

Figure 2. FIG 3 to be placed near here. CAPTION: Content words most frequently occurring in relevant tweets about *Beyond the Fence* (using `http://www.wordle.com`). Common English stopwords are removed, as are the phrases relevant to accounts/hashtags typically tagged as identifiers to refer to the project, or to direct tweets towards people involved in the project: 'beyondthefence', 'computersaysshow' and 'computermusical'.

Figure 4 shows the same word cloud data but with Twitter handles removed, so that we see the text that is said about the musical, less dominated by the twitter account names of people who many of the tweets are directed at.

Next, sentiment analysis was performed using an established NLTK toolkit implementation of Pang and Lee's sentiment analysis tool (Pang & Lee, 2005). This analysis was restricted only to the twitter data. Reducing a lengthy review to a single indicator of sentiment is less useful than the same summative action for a (maximum) 140-character

[18] For sources for each review quote here, see Table 3 in Appendix B.

Figure 3. FIG 4 to be placed near here. CAPTION: Content words most frequently occurring in relevant tweets about *Beyond the Fence* (using `http://www.wordle.com`). This figure is built from the same data as Figure 3, with the difference being that twitter account names (handles) for people mentioned in the tweets are removed, to allow us to see more easily the words that are being used to talk about the musical.

Table 2. Results of sentiment analysis of the tweet data. Sentiment was analysed as either Negative, Neutral or Positive, and each tweet was assigned one of these labels with a confidence of between 0 (lowest confidence) and 1 (maximum confidence).

Sentiment	Num tweets	Average confidence (to 3sf)
Negative	63	0.608
Neutral	66	0.708
Positive	158	0.691

tweet; the sentiment analysis tool used is excellent at highlighting the key emotion (positivity or negativity rating) of a short text, but it disguises subtleties in changes in sentiment throughout a text by reducing these different sentiments within a text to a single average rating of positivity or negativity. However, given the 287 tweets, it is interesting to summarise their positivity or negativity towards the project via sentiment analysis.

Table 2 shows us that 158 of the 287 tweets (approximately 55% or just over 1 in every 2 tweets) expressed positive opinions of the project, with an average confidence of 0.691 recorded by the sentiment analysis toolkit for the correctness of this analysis (out of a maximum of 1.00). 66 tweets (approximately 23% or just under 1 in every 4 tweets) were neutral about the project, with an average confidence of 0.708. 63 tweets (approximately 22% or just under 1 in every 4 tweets) expressed negative comments about the project, with an average confidence of 0.608.

So sentiment analysis shows that roughly half the tweets expressed positive opinions, and half expressed negative or neutral opinions. To interpret the data another way, three quarters of the tweets were positive or neutral about the project.

As examples of positive tweets, below we see the five tweets with the highest confidence rating for assignment of a positive sentiment to the tweet (confidence ratings given to 3 significant figures):

(1) @JackReitman @BeyondTheFence hope you enjoyed the show Jack. Nice to see you. I absolutely loved it! [0.916 confidence]
(2) At the wonderful @ArtsTheatreLDN for @BeyondTheFence_ #intervaltweet [0.886 confidence]
(3) #BeyondTheFence was brilliant @ArtsTheatreLDN tonight. Many thanks @cjjohnsonsinger and all the rest of the hugely talented cast. I loved it [0.883 confidence]
(4) Beyond The Fence was beyond amazing. Cast were phenomenal, the music was beautiful. Laughter, tears (many), fantastic show! #beyondthefence [0.875 confidence]
(5) Loved #BeyondTheFence_ tonight @ArtsTheatreLDN great cast, great songs, funny and sweet #ComputerSaysShow #theatre #london #musical [0.875 confidence]

Positive sentiment is occasionally being influenced by positive sentiment on topics other than the show, for example in tweet 2 above, but the five tweets above are good representatives of positive tweets about the show. For examples of negative tweets, similarly we look at the five tweets with the highest confidence rating for assignment of negative sentiment to the tweet:

(1) Sorry to miss you tonight @roblevybass @Bobolobolobolob @BeyondTheFence_ as I stayed for q&a, but well played chaps! Hope to catch up soon [0.852 confidence]
(2) I'm sorry for nearly falling over during your curtain call. I was just clapping too enthusiastically...) @BeyondTheFence_ [0.841 confidence]
(3) @rlizcooper @BeyondTheFence_ Wish I could have. Went to the initial read through so thought I might get a few secs of airtime! [0.833 confidence]
(4) Seriously miffed on closer inspection of @BeyondTheFence_ website that Sky have already nabbed 'Computer Says No' @lsrushton, that's my pun! [0.801 confidence]
(5) #computersaysshow on @SkyArts is really, really thought provoking... I do think the fundimentals are being nudged by humans tho! [0.761 confidence]

There are more examples here of sentiment analysis being influenced by comments unrelated to opinions on the show, such as example number 2, 3 and 4. Though this analysis suggests a positive bias, we can find more pertinent examples of negative tweets tagged as negative by looking further into the data.

(1) Really interesting, but a computer's no match for #Sondheim @composerstephen or #Hamlisch. #beyondthefence #stageyAvril Thornton
(2) Anybody seen that Sky Arts #BeyondTheFence ? Yes n No for me. Some of it was a bit formulaic but then, of course it was? Worthy
(3) So #ComputerSaysShow is telling me that robots are taking the job of #actors. We have enough competition already! #actorslife
(4) Just watched #ComputerSaysShow about #BeyondTheFence: 'almost 25% of the lyrics were written by computer....' NOT first computer musical!

Analysis via automated techniques has given some indication of the nature of the twitter data; now qualitative analysis can help us to dig deeper into comments made about the show. A qualitative coding analysis was performed on the textual data to identify key reoccurring themes in the data. As remarked above, the Twitter data potentially contains some bias; there seemed to be in the data a number of tweets directed at congratulating individuals within the show, though other tweets took a more generalised view. Tweets often came from people who were reporting their opinions having just seen the show, so give us an interesting perspective of examples of reactions immediately on leaving the theatre. The reviews contain typically more balanced, reflective and (due to the lack of character limits) more lengthy published considerations on the show itself. Taking these data sources together, then, but prioritising the reviews as a potentially less biased (overall) source of evaluative data and treating the Twitter data more as examples than as indicative of general trends, several interesting points emerge for analysis:[19]

3.6.2.1. Reviews of success. As one reviewer comments, the Beyond the Fence project is indeed '[s]eemingly wanting to be judged as the output of an experiment rather than a "proper show" ' (BroadwayBaby, see Table 3, Appendix B). While this seems like a criticism, it is actually not too far from the truth as an appraisal of the project's aims (Gale, 2016, personal communications). Typically, reviews focused on judging the show on its quality and fit as a 'proper show': 'Computers can help write a musical, it seems, but they can't yet write a good one' (Engadget UK). 'This show is as bland, inoffensive, and pleasant as a warm milky drink' (Guardian). The Londonist was a little more encouraging: '[w]hat's our measure for success? Well, the audience we saw Beyond The Fence with was applauding just about every number and was brought to tears by

[19] For sources for each review quote here, see Table 3 in Appendix B.

at least two of them.'

Several reviewers reflected at the experiment as a whole as part of their reviews. What's On Stage asked if computers can create more challenging material and concludes: '[o]ne day, maybe, but not yet. Not yet.' The Financial Times reviewer judges that 'this bold experiment doesn't solve the many contradictions it throws up.' Similarly, the Reviews Hub says that 'Beyond the Fence is an interesting experiment but it shows that computers are a long way off from creating a hit musical.' 'Here, the puppet masters' digital strings are still a little too visible' (Musical Theatre Review). The West End Frame concludes that 'as a theatrical event the show is remarkable. However, as a piece of theatre in its own right Beyond The Fence doesn't stand strong; it feels contrived and clunky.' But perhaps the computational creativity community can feel more heartened for future work here: 'it is an interesting development in the intersection between theatre and technology that I suspect we haven't heard the last of' ('There ought to be clowns' blog).

Opinions on Twitter were divided as to the success of the musical, though typically tweets were mostly positive overall in tone, particularly from those who seemed to be tweeting just after having been to the musical. Tweets ranged in a spectrum from 'Finally watching #BeyondTheFence .. It's actually really good! Just How???' (JenniferRaddit, March 6 2016) to 'Can a computer create a hit musical? .... Nope' (Will_So, March 5 2016).

3.6.2.2. The validity of computers as creative entities within the creative team. 'This experiment ... has plainly benefited from a lot of human intervention ... To call it "computer-generated" is misleading. "Computer-initiated" and "computer-assisted", though less grabby, are more accurate - and in their own way provide a thought-provoking novelty' (Telegraph). A number of reviewers commented on how human members of the creative team 'are, essentially, curating and correcting the computers' output' (What's on Stage). Rarely if ever do the reviewers allow the software participants any creative agency or responsibilities for their output. Instead of being treated as co-creators in an interactive creative situation, computers are often portrayed in reviews as learners rather than creatives, whose work the human participants are being asked to bring up to professional level for the final product. In this project the computer software is not able to engage in revision, respond to feedback (particularly as musicals can change from night to night in response to feedback) - they provide material for people to curate. '[h]ere is where the computer generated claim starts to unravel. There's no software that can put all of these elements together and turn them into a musical. That requires a human'[20] (Engadget UK). Some similar comments were made in tweets about the show, for example: 'I do think the fundimentals [sic] are being nudged by humans' (kevlawdrums, March 3 2016) and 'it's absolutely brilliant - but had a lot of human creative input too' (SemarkJ, February 27 2016).

A notable and fascinating exception to this generalisation is by Musical Theatre Review: ' "What if a wounded soldier had to learn how to understand a child in order to find true love?"' was generated by WHIM, the "What If Machine". And in tribute, Ako Mitchell's US soldier Jim rubs his thigh in pain occasionally - less wounded soldier, more bloke who should have done a few more warm-ups before exercising. There's an emotional need for him that is implied in WHIM's question that is not addressed here, leaving the show's central love story to feel a little anaemic.' In other words, software has supplied a creative idea to its human collaborators that is not used well, to the detriment of the overall effect of the show. Several Twitter commentators also expressed their approval

[20] A provocative request for future research?

of the computational aspect of the show, e.g. 'outstanding!That "infographic" worked a treat. Ups and downs in all the right places' (TVStix, March 2 2016), and 'Better than a fair few human hacks' (StuBlackEsquire, March 1 2016). Other tweeters had less complimentary opinions of the success of the computer-generated material but typically the tone on Twitter was more accommodating of computational attempts than in reviews, for example: 'Not sure computers are as good as human composers as didn't go home humming tunes but fab idea!#Bravo' (emilysalmon, February 24 2016).

In the Twitter data, a lot of interest was expressed in the computational creativity aspects of the project, often sparked in response to the *Computer Says Show* documentary (Gale et al., 2016), e.g.: 'Really enjoyed the show. Need to catch up on the @SkyArts show - v interested in the process involved' (vikkistone, March 6 2016) and 'En route to @BeyondTheFence_, a computer-generated musical. There is no way this won't be interesting..' (charlotte_dann, February 25 2016). Equally interesting, though, were the tweets that disregarded the computational aspects of the project and reacted to the musical as a show in its own right: 'Computer music aside, thoughts on @BeyondTheFence_ - Women are brilliant - We have SO many stories to tell - Telling them is "at our feet" ' (KateMalyon, February 27 2016).

3.6.2.3. Perception of the musical as too formulaic. '[A]ny show built according to a formula runs the risk of sounding, well, formulaic. Beyond the Fence doesn't avoid this risk: despite the cutting-edge technology involved in its creation, the show itself is middle-of-the-road and predictable.' (Financial Times). 'Follow a formula and - who would have thunk it - you get something formulaic' (What's on Stage). Most of the reviewers criticised the musical for feeling too pattern-driven and predictable, rather than including content to challenge rather than replicate musical theatre. Several reviewers criticised the musical for its lack of avant-garde, challenging or ground-breaking content. 'Picking through old shows for tricks evidently means the plotting is predictable and at times a bit shallow' (Londonist). The issue with this type of criticism is that the software involved was developed on the task of 'replicating the past, not challenging it' (What's On Stage). 'Nothing moved the needle. Nothing felt fresh.' (Engadget UK). One or two tweeters also made similar (though less critical) comments specific about the formulaic nature of the show, such as 'Anybody seen that Sky Arts #BeyondTheFence ? Yes n No for me. Some of it was a bit formulaic but then, of course it was? Worthy [thumbsup emoji]' (joglasg, June 5 2016). Not all tweeters felt that the show was too mechanically derived though, for example 'Could you tell? I'm not sure I could' (BBCTech, March 2 2016).

Where computer-generated aspects moved away from typical output, this also attracted criticism. For example several reviewers criticise the lyrics for being atypical - an ironic example of this is where during such criticism, a reviewer highlights the *We are Greenham* song as one of the songs 'that work' (Musical Theatre Review), even though that song consists entirely of computer-generated lyrics.[21] Sometimes criticism of machine-generated content permeated into parts of the production humans had responsibility for: one reviewer suggested that the live band (of human performers, performing music that had been orchestrated by humans), 'sounded extremely robotic and monotonous - it sounded as if backing tracks were being used' (West End Frame).[22] The Financial Times said, though: 'writers have been using formulae for years to make commercially-minded culture and so what difference does it make if it's a formula developed by a machine?

[21] It should be noted that this song's lyrical content was generated differently to many other songs, using corpus analysis of protest songs from Greenham Common rather than the Cloud Lyricist.

[22] For balance: this was the only reviewer who criticised the band's performance; another review praised how the tunes were 'well played' (Theatreworld Internet Magazine).

This is the main talking point of the show ... a debate that will run and run.'

3.6.2.4. Lack of context awareness?. As the Guardian reviewer observed, 'The software appears to have ... zero grasp of 1980s feminism and the Greenham Common women's peace camp. A pity, because that's where it's set.' (Guardian). It is interesting to discuss the extent to which this is a fair criticism. The Greenham Common setting and feminist themes were chosen by the humans in the creative team rather than via software, so one could argue that the software should not be expected to know about this contextual setting. Certainly many of a younger generation of musical theatre professionals could also be criticised for not knowing about this particular event in UK history, and the software is given no chance to research these themes post-decision to use them for *Beyond the Fence*. Criticising the computer participants for not being more knowledgeable seems harsh. But we learn from this project the perceived importance of computational systems demonstrating knowledge of context when being creative.[23] should (and could) the computers have used more contextual awareness, to develop ideas based on contextual information available e.g. via web resources? It seems from a number of review and twitter comments that this wider contextual knowledge was expected, for example where reviewers criticised the show for not engaging more with feminism issues, or debates about nuclear weapons that concurrently happened at the same time as the musical was being performed. More than one reviewer commented that they would have liked to see a plot centred around scenarios a computer might have some knowledge of (e.g. the Financial Times suggest 'circuit boards in revolt'), though no follow-on comment considers how human audiences might perceive the results.

One area where the computational creativity software was roundly criticised was in the ability to understand content unfolding over time, in longer-term structures. 'Even if they give you a stroke of genius, they can never follow that up... every thought is a new thought' says Benjamin Till, in (Gale et al., 2016). Nick Collins also reflects in (Gale et al., 2016) that this issue raised by Till is an area for further research.

Occasionally tweets bucked this trend of judging computers to be out of touch with a wider context: 'To think a computer wrote @BeyondTheFence_ is insane, just how?! Such a diverse piece of theatre that is so current in today's society' (@charlottewood15, February 29 2016). Opinions on twitter were divided, though, for example: 'It's #Hair meets #StarlightExpress but without the wit' (MrJonathanBaz, March 3 2016).

3.6.2.5. Gimmicks by boffins? Preconceptions and biases about computational creativity. It was interesting to see several reviewers make negative comments to the effect that poor human-produced musicals appear as if they were written by a machine, e.g. 'Plenty of musicals written by humans sound as if they have been composed by a machine' (Guardian). The Telegraph reviewer reported how 'the evening somehow over-rode my default scepticism'; others made more negative comments about 'gimmicks' introduced by 'boffins'. Interestingly, during development, the work-in-progress musical was performed to a test audience of regular theatre-goers who were unaware of the origins of much of the material being computer-generated. The audience reacted positively to the preview performance, but what is more telling is their reaction once they were informed about the computational work and its contribution to the musical (Gale et al., 2016). They reacted with stunned silence, followed by nervous laughter. Preconceptions about what computers can (and cannot) do are there to be dealt with, in computational creativity -

[23] This project reinforces the perceived importance of contextual awareness for computational systems that are beginning to be discussed more and more in computational creativity.

the reviews here show that this issue should not be ignored when engaging with the public in computational creativity research. Perhaps one solution is to be guided by proposals requiring transparency and clarity in computational creativity processes; Colton et al. (2011) argue for the need to provide a 'frame' or description that details the processes used in a computational creativity system, for computational agents to be more easily perceived as creative in their own right.

Interestingly, the Twitter data - largely coming from tweets by people who had seen the musical or show, or engaged with the project to some extent - are largely positive towards the overall project, with reduced or no negativity or resistance towards the computational aspects of the show. Perhaps the increased engagement of the tweeters with the show allowed them to override negative preconceptions to some extent?

3.6.2.6. Response to critics' reviews from Wingspan Productions. In personal communications, Archie Baron from Wingspan Productions responded to the various points raised by the critics in reviews:

> 'We decided - rightly - to deploy two different methodologies. First we attempted to model a commercial hit based on the stats for a 'recipe for success'. This gave us various constraints which we had to task everyone - humans & software - with delivering. Downsides: the formulaic criticism and an end-product which was never going to be as original in end-product as it was in methodology - we couldn't let rip creatively because everyone was working within the constraints of a commercial-facing model. Upsides: the project needed constraints - with multiple systems and commercial investment these were unavoidable. Ours at least had an impeccable logic behind them. Secondly we needed something which the commercial musical fan might buy tickets for - they want a familiar product not an experiment in the avant garde.'
> (Baron, 2016, personal communications)

3.6.2.7. Summary of evaluating Q1. As a reminder, the evaluation Q1 is:

To what extent was *Beyond the Fence* successful as a computer-generated musical?

> 'Beyond the Fence is conceived by computer and substantially crafted by computer.'[24]

The *Beyond the Fence* project has achieved the goal of staging a musical which has been generated through collaboration between creative computer software and human musical theatre experts. The *Beyond the Fence* (Till et al., 2016) musical was performed for two weeks in London's 'Theatreland' (the West End district of London, UK). The premise, plot elements, storyline, music and approximately a quarter of its lyrics were computer-generated, using various creative systems in conjunction with human experts. The project has been recorded in documentary form (Gale et al., 2016) and its technical details reported in Colton et al. (2016).

The human creatives involved have reported that they feel the project has been a success, and the show has been performed to good-sized and receptive audiences each evening of its run. Audiences seem to have enjoyed the musical show itself, though reviewers and online commentators remain divided in their opinions on the computationally created content. While 'this bold experiment doesn't solve the many contradictions it throws up' (Financial Times, see Table 3 in Appendix B), it has nonetheless contributed towards making these 'contradictions' and debates open and relevant for discussion among a much

[24] `http://beyondthefencemusical.com/about-the-show` (last accessed March 2016, now retrievable via `http://web.archive.org/web/20160418004951/http://beyondthefencemusical.com/about-the-show`).

broader audience.

3.6.3. Evaluating Q2: Computational creativity and Beyond the Fence

Evaluation Q2 is:

How well has the project engaged with and contributed to computational creativity research?

How has the *Beyond the Fence* project engaged with computational research? What can the field of computational creativity learn from this project? Where has computational creativity successfully contributed to this project? What has not worked so well in terms of computational creativity's application to this problem of creating a new musical? And where could future work in this direction be directed?

In the short term, this project has played an important part in raising the profile of computational creativity research. This project has taken on an ambitious task, and has tasked computational creativity researchers with applying the software we create both as individual pieces of software and (importantly for computational creativity) in combination with other systems.

Several leading computational creativity researchers and projects engaged with the project in terms of providing software and being involved in the process of the project. The 'ground rules' for the project (Appendix A) dictate the need to 'team up with the best scientists in the field in Europe so our experiment is as good as it can possibly be in 2015' and let 'people judge the very best work that can be computationally derived in 2015' (for preparation and performance in early 2016).

According to these ground rules, Wingspan were keen to 'encourage debate and shine a light on some profound issues we're confronting'. For example, during research for the project documentary (Gale et al., 2016) Gale also investigated questions with various computational creativity researchers (including myself) about the role of computer software in the 'creative conversation': 'Computers can become another voice in the room - speaking "from the data" as it were - and we instinctively question that' (Gale 2016, personal communications). In her research, Gale observed through this experiment (Gale 2016, personal communications) that people are often surprised at the challenges and difficulties in computationally generating creative artefacts - perhaps underestimating the complexity of the tasks involved. She also saw resistance in people's reactions to computers being creative. One example of this resistance is illustrated in the documentary that reports on this project (Gale et al., 2016). Benjamin Till reflects a number of times on his apprehensions about working with computational software, particularly music generation software. He says: 'maybe I was a little bit harsher on it than I should have been'(Gale et al., 2016), going on to explain that because music was composed by computer software and not by humans, he partly felt defensive, that he did not want the results to be good. This is useful to observe, in the context of general trends noted before on the effect of bias in judging computational creativity (Jordanous, 2012a; Lamb et al., 2015; Moffat & Kelly, 2006; Pasquier et al., 2016).

Gale and her team were interested in what computational creativity researchers thought about the project; such discussions receive attention in the documentaries resulting from this project (Gale et al., 2016). For example, Gale was interested in whether the *Beyond the Fence* project was doing work that was in some way different to existing current work in computational creativity, or work that was exciting for the field. Relevant aspects that emerged in such discussions included the collaborative aspects of the creation of the musical, and the scale of the overall project (especially as the project resulted in public performances presented in a venue in a high profile London location.).

The team also became interested in the nature of creativity itself, and how it can be modelled, replicated or simulated using computational means; computational creativity researchers share these same research questions and many stimulating discussions arose as a result of people within Wingspan raising these questions. Many of these points are documented in Gale et al. (2016).

During the panel discussion on the *Beyond the Fence* project at the 2016 International Conference on Computational Creativity, interesting points were raised by one of the panellists about the perceived and actual complexity of the task of creating a musical. Melly Still, an experienced opera and theatre director, argued that the project team underestimated the huge creative capabilities needed for creating and staging a musical. In her opinion, the parts of this complex process that were performed by a computer represented only a small and relatively simple part of the overall process; the claim of the musical being *computer-generated* was too ambitious. This echoes similar objections discussed in the analysis of reviews and comments about the musical above, making the same questions about this claim.

In this current project, one important lesson to learn is illustrated for computational creativity research. The evaluation in this paper has showed us that the computer software are for the most part considered to be passive participants in a process curated by humans. Essentially, as discussed above, the computational participants typically contribute material that is shaped by the human participants in the creative team, and the human participants have the final say in what makes it to the 'final cut'. Perhaps, to paraphrase the title of this musical, the 'fence' in musical theatre represents the recognition of computers as genuine creative participants contributing to the creative process. In this interpretation, the *Beyond the Fence* project does not fully see creative software moving 'beyond' this 'fence'. But certainly the debate on computers being creative has been opened up to wider public scrutiny, a debate to which the project makes a significant contribution.

In post-conference discussion via a computational-creativity-specific mailing list, Mark d'Inverno (a commentator on computational creativity research for Gale et al. (2016)) notes this issue, giving an interesting possible interpretation and suggestion for future work to start to address it:

> ' it still seems we are some way of the stage where content generated autonomously by machine could ever sustain our interest for any period of time. As a computer scientist this is an interesting challenge, but as a musician myself, and as a potential audience member, I find it hard to imagine a scenario where we could sustain interested in solely generated artificial content for very long. The times when something has sustained interest in me is in music performance situations because the human is put under new challenges to work with an autonomous system because it can take them out of their comfort zones and they have to work harder to make things work musically. And for the musical to work in this way they need to embue the system with its own creative agency. They need to give it equal billing to get the best out of themselves and of the unfolding creative collaboration.
>
> So I think that where the future lies is exploring artificial creative agency. This is the idea that machines enable new kinds of creative partnerships for humans. That they stimulate, challenge, provoke us to work in new ways and to produce content that would not have been possible without the system. And, come to that, would not - or could not - have occurred working with any other human collaborator.
>
> So we need to start with the human creative, and build systems that demonstrate this creative agency to creative.' (Mark d'Inverno, `https://groups.google.com/d/msg/computational-creativity-forum/WUxNbLxFXg0/w2F8nlQwAwAJ`, posted 31st August 2016)

This point is taken up again in Section 5.2, when future work is discussed.

4. Evaluation of the DECIDE framework for computational creativity

Given that the DECIDE framework has only very recently been posited as a research tool in computational creativity, with little application so far beyond the work reported in Kantosalo et al. (2015), what have we learned from this application of DECIDE for evaluation? Has it been useful? What has it enabled us to find out - and what is missing or inadequate from DECIDE as an evaluation framework? Can the DECIDE framework fill a methodological gap in the evaluation of co-creative computational systems?

DECIDE has enabled evaluation to be carried out in a planned and constructive manner. Various data has been collected from multiple perspectives for an informed evaluation, even given that a number of issues have hindered data collection (as discussed in Section 3 for step DECIDE-4).

What DECIDE has not provided (perhaps because it is not reasonable to expect this) is a formal set of steps to objectively carry out a full and complete evaluation. We cannot be certain that the evaluation takes into account all necessary data, nor that the subset of data we have is sufficient for drawing conclusions from. In fact, especially given timeliness issues, we can be reasonably sure that there is missing data, such as long-term indicators of the musical's cultural value in terms of take-up and influence in other musical shows.

The evaluation is quite subjective, in a scientific research domain which continually wrestles with the computational, objective treatment of the subjective, intangible topic that is creativity. DECIDE does, however, help to systematise the approach to evaluation in a more formal manner. Intended as a checklist for planning and carrying out evaluation Rogers et al. (2011), it fulfils this role, rather than that of a more prescriptive and rigorous methodology for evaluation. Thus it is most helpful to treat DECIDE as such: a framework that helps us remember and deal with several important aspects when planning and undertaking evaluation of computational creativity scenarios with multiple creative contributors (computational or human). In this role, it has been a useful framework to guide the current evaluation reported here. Further application of the DECIDE framework in more evaluation case studies would help to identify and refine more operational, relevant methods for use in evaluation, particularly for step DECIDE-6.

5. Conclusions

5.1. *Summary*

To tackle the task of evaluating *Beyond the Fence* as a computational creativity project, we noted in Section 2.2 that existing evaluation tools in computational creativity research were inadequate for the task of evaluating a project with multiple distinct creative contributors to an end product. Existing methods tend to assume that only one creative system is being evaluated. Following recent work by Kantosalo et al. (2015) for evaluating computational creativity projects with more than one creative participant, in this current paper the DECIDE framework (Rogers et al., 2011) has been employed to plan and undertake the evaluation process.

This paper has investigated and evaluated the *Beyond the Fence* musical theatre computational creativity project from two interlinked but distinct perspectives. The project is evaluated to learn more about in what ways it is successful as a computer-generated musical, and as a computational creativity project. Given the challenges in evaluating computational projects with this collaborative complexity, though, this evaluation also affords observations about the suitability of the DECIDE framework for evaluation of collaborative computational creativity projects.

From the first of these perspectives: the project was evaluated above using the DECIDE framework. The goal of evaluation was identified as ascertaining the extent to which the *Beyond the Fence* project has been successful, and this was further refined into two evaluation questions:

(1) **To what extent was *Beyond the Fence* successful as a computer-generated musical?**
(2) **How well has the project engaged with and contributed to computational creativity research?**

Various aspects were identified as metrics and data sources for evaluating the first question, including indicators both of cultural and commercial value. The data for the first question and relevant other observations were subsequently considered in the context of the second question. Methods were identified for evaluating these aspects, with relevant ethical and practical issues identified, then evaluation was carried out: results were summarised in the previous Section 5.1.

The *Beyond the Fence* project has been successful at showing how existing creative software can indeed be used within the scenario of creating a plausibly acceptable musical. The resulting musical has been moderately well-received by most critics, though it has exemplified how expectations can be set very high for computational creativity in the public eye in order for the outputs to be judged successful. Questions have been raised over the extent to which the computer software was a genuine creative agent within the processes of creating the musical, with criticism attracted over whether the end result was too formulaic, gimmicky, or lacking in a suitable grounding of the chosen plot context. The performed musical, however, did reasonably well in terms of audience attendance and audience satisfaction, despite (or perhaps because of) a lack of traditional marketing for the show in comparison to other West End shows in London; as remarked above, the show was marketed as the ‘world’s first computer-generated musical’. The project overall has received various expressions of cultural value including award nominations both for the project documentary and the show as a musical in its own right.

The project has raised the profile of computational creativity research by attracting media attention and theatre audiences, and some some useful lessons have also been highlighted for computational creativity. In particular: *computational creativity should encompass more than a. replicating norms and b. completing independent tasks within the creative process (with little feedback or collaboration between tasks).*

From our second perspective: considering this evaluative work as a case study to investigate the proposal of the DECIDE framework as an evaluative tool for computational co-creativity systems, we have gained some useful experience via this case study. Application of a methodology is a useful way of observing how a methodology works in practice, and we have learned from this application.

We have been able to design and carry out a useful, detailed, broad yet targeted evaluation of the success *Beyond the Fence* project, generating useful formative feedback as outlined above that we can investigate in future work (see Section 5.2). However, a lot of interpretative load is placed on the evaluator when using DECIDE. DECIDE is not a definitive and static set of steps for a one-size-fits-all co-creativity evaluation tool, but a set of guidelines that can be implemented for assisting us in a transparent and systematic evaluation of a subjective set of results. We conclude that the DECIDE framework should be treated as a framework to guide us in designing and implementing a pragmatic and informed evaluation. Further application and development of the DECIDE framework could be done to make the framework more operational as an evaluative tool, as discussed during the next section. However we are still a considerable distance from being able to realise the idea suggested by this tweet:

> Has anyone invented an algorithm that will automatically review #BeyondTheFence @ArtsTheatreLDN yet, put us critics out of a job? :):):):):)

Not yet...

5.2. *Future work*

As remarked above, further case studies and investigation into the DECIDE framework would be helpful in making it more operationally useful for computational creativity research. There are a number of areas of future work specifically related to the *Beyond the Fence* project.

The project led the Wingspan team to became interested in questions to do with creativity in different domains. Perhaps because of her biochemistry research background, Gale particularly focused on people's perceptions of creativity outside of artistic domains, and cultural issues that may affect how we distinguish between creativity in scientific domains compared to artistic domains (Gale, 2016, personal communications). Although this current project concentrated on creativity in musical theatre, it will be intriguing to see the directions of any future projects by Wingspan Productions concerning computational creativity.

Wingspan Productions have identified other ways in which they would like to develop academic parts of this work. One of those relates to sociological and psychological issues encountered during the project. In step DECIDE-5 of Section 3, one ethical issue identified was the conflict between the project's overall overall 'ground rules' (see Appendix A) and the professional goals and standards of several of the human musical theatre professionals participating in the project. Baron (2016, personal communications) has expressed interest in investigating these more psychological aspects of the show with the human participants, seeing how they dealt with ethical and professional conflicts. Another avenue Baron has raised for potential future work is in how successfully the show's structure matched the emotional arc it was intended to follow. Part of the data analysis revealed points at which the show should arouse different emotions and reactions from the audience. In analysis of this, Baron suggests in personal communications (2016) an experiment with a group of people watching the DVD[25], surveyed at regular intervals on what emotions they are experiencing. Such an experiment could, if handled carefully, reveal more about the success of transferring the computational data analysis to the final musical, helping us to learn from the successes and failures in this process for future such work.

Returning to the current project under discussion, what contributions does this *Beyond the Fence* experiment make to the computational creativity field: currently and longer-term? And given directions in computational creativity research, what might this musical be like in a few years?As discussed above in Section 3.6.3, this *Beyond the Fence* project has tasked computational creativity researchers with applying the software we create both as individual pieces of software and (importantly for computational creativity) in

[25] Or possibly another performance of the show

combination with other systems. While some work has been done in combining different creative systems for a broader perspective on creative tasks (Monteith, Francisco, Martinez, Gervás, & Ventura, 2011, e.g.), the idea of different creative systems communicating and/or collaborating with each other (Corneli et al., 2015, e.g.) is an exciting area to look at (especially now many different creative systems have been developed and are potentially at our disposal).

A key point raised in the *Computer Says Show* documentary (Gale et al., 2016) revolves around the generation of the musical: in Gale's words:

> 'right now, [do] humans have to be part of a project like this? [Do we] need some people in the mix to curate? and to make choices? as currently the computers involved can't censor their output very well, and they don't talk to each other yet either!'
> (Gale, 2016, personal communications)

Some recent computational creativity research has focused on how creative computer systems might be able to interact with each other, communicate and give each other feedback towards refining and developing their own creative work (Corneli & Jordanous, 2015; Gervás & León, 2014; Román & Pérez y Pérez, 2014). One exciting potential area for future work is in using computational creativity to carry out this 'curation' step. Can computational creativity software curate parts of a musical (lyrics, plot, characters, emotional timelines) into a single production? To what extent is interaction with a human(s) necessary in this process? Responding to the criticisms raised in reviews that *Beyond the Fence* (Till et al., 2016) is not computer-generated, but 'computer-assisted' or 'computer-initiated' (as discussed above): to what extent can computer software actually generate the full content of a musical? Could software do everything? Could social media comments also be harnessed - perhaps as a way of garnering instant feedback which can inform software in making edits to the show for the next evening's performance? Or does this lead us into a trap where we judge the programs doing tasks set by humans, via subjective opinions of the end result rather than the success at each smaller task (without evaluating decisions taken on how to break complex creative tasks into subtasks)?

As discussed above in Section 3.6.3, Mark d'Inverno has noted that:

> 'In most cases there wasn't a "creative conversation" between human and machine that one would normally expect between humans when devising a show. And that seems to be what the next step for us should be: how do we build systems where these creative conversations can take place in ways which are meaningful for human creative? ... we need to start with the human creative, and build systems that demonstrate this creative agency to creative. Systems which immediately - or at least quickly- open up new opportunities for collaboration where the human creative is happy for the system to take creative control at points in the dialogue. Such systems need agency, and this involves an awareness of the human creative, their goals, their previous content, the way they like to work, the artistic influences of the creative, and also - and this is where it starts to get interesting - influences (algorithmic or human) that could take the human creative into entirely unexplored territories. But I think we need to start with the creative and think about designing systems with the right kinds of agency and flow. Starting with the system and then trying to work out how a human creative might interact with it later seems the wrong way round. '
> (Mark d'Inverno, `https://groups.google.com/d/msg/computational-creativity-forum/WUxNbLxFXg0/w2F8nlQwAwAJ`, posted 31st August 2016)

Frameworks for collaboration between creative participants (software-based and human) are starting to emerge in computational creativity research, such as the Concrete-Flows and FlowR framework system (Charnley, Colton, Rodriguez, & Corneli, 2016; Žnidaršič et al., 2016), which may represent a first start in allowing computer systems and people to engage in dialogue during the creative process. Typically, creative software does not often include the facility to interact with other systems or people, though Kantosalo et al. (2015) reviews several counter examples to this generalisation. We can draw from the *Beyond the Fence* project that this facility is worth including. In the above-cited post, d'Inverno goes as far to suggest that we should, in computational creativity research, start 'not with the question of "I wonder what my system can do" but "how can I help current creative teams in ways that would excite and inspire them?".' This is a controversial statement, and perhaps such a whole-scale change in focus would not be welcome or productive; we should be careful not to throw the proverbial baby out with the bathwater. There is no reason, though, why both foci could not (or should not) co-exist in tandem, and percolate through future computational creativity research.

Overall, the impact for computational creativity research is that for future larger-scale multi-system public-facing projects to be more successful, we are reminded of the need to develop as well as replicate human creative achievements, and to allow our systems to be able to communicate and refine work as well as offer inspirational material. We are also reminded that standards are held high for computational creativity in the public eye.

5.3. *Final remarks*

A central character in *Beyond the Fence* is George, the little girl who is unable to talk for most of the musical - until her trust of the people supporting her allows her to find her voice. The George character can be used as a metaphor for computers involved in this project. The various software play key roles in the unfolding of the musical, but do not necessarily have the ability to join in the conversations around them and talk about what they have done... yet.

Acknowledgments

In collaboration with the computational creativity researchers acknowledged throughout this paper, the Beyond the Fence project was realised through the hard work of Dr Catherine Gale, Archie Baron and their team at Wingspan Productions. Their drive to pursue and share this work opened up much for us to talk about and learn from, both in our research and in the application of this research 'beyond the fence' in real-world creative situations. I hope this paper goes some way to further substantiate a comment made by Gale after the project: 'It's so brilliant to see the project living on in the research space, which was always the hope however it unfolded.' The work of Wingspan Productions was commissioned by Sky Arts and supported by Wellcome Trust. I also acknowledge the cast, crew, band, creative team and management involved in the *Beyond the Fence* performances, as well as the extremely helpful comments made by the anonymous reviewers of this paper during peer review.

References

Bown, O. (2014, June). Empirically Grounding the Evaluation of Creative Systems: Incorporating Interaction Design. In S. Colton, D. Ventura, N. Lavrač, & M. Cook (Eds.), *Proceedings of the fifth international conference on computational creativity* (pp. 112–119). Ljubljana, Slovenia: ACC.

Bown, O. (2015). Attributing creative agency: Are we doing it right? In *Proceedings of the 6th international conference on computational creativity.* Park City, UT: ACC.

Charnley, J., Colton, S., Rodriguez, M. T. L., & Corneli, J. (2016). The flowr online plat-form: Automated programming and computational creativity as a service. In F. Pachet, A. Cardoso, V. Corruble, & F. Ghedini (Eds.), *Proceedings of the Seventh International Conference on Computational Creativity (ICCC'16)* (p. 363-370). Paris, France: Sony CSL Paris.

Collins, N. (2016). A funny thing happened on the way to the formula: Algorithmic composition for musical theatre. *Computer Music Journal*, *40*(3), 41-57.

Colton, S. (2008). Creativity versus the Perception of Creativity in Computational Systems. In *Proceedings of aaai symposium on creative systems* (pp. 14–20). Stanford, CA: AAAI.

Colton, S. (2012). The painting fool: Stories from building an automated painter. In J. McCormack & M. D'Inverno (Eds.), *Computers and Creativity* (p. 3-38). Berlin, Germany: Springer-Verlag.

Colton, S., Charnley, J., & Pease, A. (2011). Computational Creativity Theory: The FACE and IDEA Descriptive Models. In *Proceedings of the 2nd international conference on computational creativity* (pp. 90–95). Mexico City, Mexico: ACC.

Colton, S., Llano, M. T., Hepworth, R., Charnley, J., Gale, C. V., Baron, A., . . . Lloyd, J. (2016, June). The beyond the fence musical and computer says show documentary. In *7th international conference on computational creativity (iccc 2016).* Paris (France): ACC.

Corneli, J., & Jordanous, A. (2015). Implementing feedback in creative systems: A workshop approach. In N. Osman & M. Yee-King (Eds.), *Proceedings of the first international workshop on ai and feedback, international joint conference on artificial intelligence (ijcai).* Buenos Aires, Argentina: AAAI.

Corneli, J., Jordanous, A., Shepperd, R., Llano, M. T., Misztal, J., Colton, S., & Guckelsberger, C. (2015, June). Computational poetry workshop: Making sense of work in progress. In H. Toivonen, S. Colton, M. Cook, & D. Ventura (Eds.), *Proceedings of the sixth international conference on computational creativity (iccc 2015)* (pp. 268–275). Park City, Utah: Brigham Young University.

d'Inverno, M., & McCormack, J. (2015). Heroic versus collaborative AI for the arts. In Q. Yang & M. Wooldridge (Eds.), *Proceedings of the twenty-fourth international joint conference on artificial intelligence (ijcai 2015)* (p. 2438-2444). Palo Alto, California, US: AAAI Press.

Eigenfeldt, A. (2015). Generative music for live musicians: An unnatural selection. In *Proceedings of the sixth international conference on computational creativity june.* Park City, UT: ACC.

Gale, C. V., Baron, A., Lomax, K., & Wingspan Productions. (2016, February/March). *Computer says show (episodes 1 and 2) - sky uk.* www.wingspansproductions.co.uk.

Gervas, P. (2009). Computational Approaches to Storytelling and Creativity. *AI Magazine*, *30*(3), 49–62.

Gervás, P., & León, C. (2014). Reading and Writing as a Creative Cycle: The Need for a Computational Model. *Proceedings of the Fifth International Conference on Computational Creativity*.

Johnson, C. G. (2014). Is it time for computational creativity to grow up and start being irresponsible? In *5th international conference on computational creativity.* Ljubljana, Slovenia: ACC.

Jordanous, A. (2011). Evaluating Evaluation: Assessing Progress in Computational

Creativity Research. In *Proceedings of the second international conference on computational creativity (iccc-11).* Mexico City, Mexico: ACC.

Jordanous, A. (2012a). *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application* (Unpublished doctoral dissertation). University of Sussex, Brighton, UK.

Jordanous, A. (2012b). A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation*, *4*(3), 246–279.

Jordanous, A. (2014). Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of 5th international conference on computational creativity.* Ljubljana, Slovenia: ACC.

Jordanous, A. (2016). Four pppperspectives on computational creativity in theory and in practice. *Connection Science*, *28*(2), 194-216.

Jordanous, A. (2017). Evaluating evaluation: Assessing progress and practices in computational creativity research. In T. Veale & A. Cardoso (Eds.), *Readings in computational creativity (in press)* (p. tbc). Springer.

Jordanous, A., Allington, D., & Dueck, B. (2015). Measuring cultural value using social network analysis: a case study on valuing electronic musicians. In *Proceedings of the sixth international conference on computational creativity* (p. 110). Park City, UT: ACC.

Jordanous, A., & Keller, B. (2012). What makes musical improvisation creative? *Journal of Interdisciplinary Music Studies*, *6*(2), 151–175.

Jordanous, A., & Keller, B. (2016). Modelling creativity: Identifying key components through a corpus-based approach. *PLOS ONE (in press)*, *tbc*.

Kantosalo, A., Toivanen, J. M., & Toivonen, H. (2015). Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the 6th international conference on computational creativity.* Park City, UT: ACC.

Lamb, C., Brown, D. G., & Clarke, C. (2015, June). Human competence in creativity evaluation. In H. Toivonen, S. Colton, M. Cook, & D. Ventura (Eds.), *Proceedings of the sixth international conference on computational creativity (iccc 2015)* (pp. 102–109). Park City, Utah: Brigham Young University.

Maher, M. L. (2012). Computational and Collective Creativity: Who's Being Creative? In *Proceedings of the 3rd international conference on computer creativity.* Dublin, Ireland: ACC.

Moffat, D. C. D. C., & Kelly, M. (2006). An investigation into people's bias against computational creativity in music composition. In *The third joint workshop on computational creativity.* Riva del Garda, Italy.

Monteith, K., Francisco, V., Martinez, T., Gervás, P., & Ventura, D. (2011). Automatic Generation of Emotionally-Targeted Soundtracks. In *Proceedings of the 2nd international conference on computational creativity* (pp. 60–62). Mexico City, Mexico: ACC.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115–124).

Pasquier, P., Burnett, A., & Maxwell, J. (2016). Investigating listener bias against musical metacreativity. In F. Pachet, A. Cardoso, V. Corruble, & F. Ghedini (Eds.), *Proceedings of the Seventh International Conference on Computational Creativity (ICCC'16)* (pp. 42–51). Paris, France: Sony CSL Paris.

Plucker, J. A., & Beghetto, R. A. (2004). Why Creativity is Domain General, Why it Looks Domain Specific, and why the Distinction Doesn't Matter. In R. J. Sternberg,

E. L. Grigorenko, & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 153–167). Washington, DC: American Psychological Association.

Ritchie, G. (2007). Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines*, *17*, 67–99. doi:

Ritchie, G. (2017). The evaluation of creative systems. In T. Veale & A. Cardoso (Eds.), *Readings in computational creativity (in press)* (p. tbc). Springer.

Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design: Beyond human computer interaction* (3rd ed.). Chichester, UK: Wiley.

Román, I. G., & Pérez y Pérez, R. (2014). Social Mexica: A computer model for social norms in narratives. In *5th international conference on computational creativity.* Ljubljana, Slovenia: ACC.

Till, B., Taylor, N., The What-If Machine, PropperWryter, Claude-Machine Schonbot, Flow Composer, . . . Wingspan Productions (2016). *Beyond the Fence - Sky UK and Wingspan Theatricals (broadcast of performance at Arts Theatre London, 26th Feburary 2016).* www.wingspanproductions.co.uk.

Žnidaršič, M., Cardoso, A., Gervás, P., Martins, P., Hervás, R., Alves, A. O., . . . Lavrač, N. (2016). Computational creativity infrastructure for online software composition: A conceptual blending use case. In F. Pachet, A. Cardoso, V. Corruble, & F. Ghedini (Eds.), *Proceedings of the Seventh International Conference on Computational Creativity (ICCC'16)* (pp. 371–379). Paris, France: Sony CSL Paris.

Appendix A: Ground Rules

The following 'ground rules' were written and circulated by Wingspan Productions around all people involved in the *Beyond the Fence* project. All participants were required to adhere to these 12 points, to keep the project on track and as close to its original goals and intentions as possible/practicable.

'COMPUTER SAYS SHOW - UNDERLYING PURPOSE AND PRINCIPLES

(1) The underlying purpose of the project is to explore a revolutionary moment in history. Technology, the computer revolution, machine learning and AI are rapidly and fundamentally changing what's done and the way it's done in almost every aspect of human experience. But how does that affect art, humanity and the creative process (those most uniquely human pursuits, almost indeed the defining characteristics of man which are therefore the most precious and hardest for a computer to learn or synthesise)? We involve ourselves in this project not because we're taking sides - wanting to embrace the future or alternately turn back time - but because we want to encourage debate and shine a light on some profound issues we're confronting.

(2) Our chosen method is a deliberately bold, new and provocative experiment - to attempt to create and professionally perform a complex full-length work of musical theatre which is, as far as possible, in as many aspects as possible, 'written' by computer (i.e. computationally generated).

(3) We have chosen musical theatre partly because it has many aspects - premise, story, music, lyrics, design etc. - partly because it is a popular art-form most commonly judged in a commercial space (effectively by the audience), partly because it very frequently has a common form or structure (a supposed 'recipe' which makes modelling it potentially easier than other art forms).

(4) We team up with the best scientists in the field in Europe so our experiment is as

good as it can possibly be in 2015. We accept that it is likely the science will get better later. We want to get the plane off the ground. Not expect it first time to fly the Atlantic. Although we've analysed the data from hundreds of successful musicals, we don't expect to be able to synthesise something better than those musicals just because we've modelled them and derived our data from them to maximise our understanding of what makes a successful musical. So our aim can't be to stage the best musical ever but to stage the best one that is as far as possible computationally generated.

(5) We agree the show must go on - and that to explore the experiment fully we need Benjamin Till, Nathan Taylor and later others in the creative team, to make choices about what is fit to perform and to do those things computers can't. We have to sell tickets - have interested people judge the very best work that can be computationally derived in 2015. So audiences will indeed expect to see a work which is 'by computer' as far as it can be because that's what we're billing. It's for them to judge. But we have to be pragmatic about staging something that is watchable, and - where intervention is required - should give the material the best chance to succeed.

(6) Wherever computers can do something, we pursue that option and try to remain 'pure' for as long as possible. Ben and Nathan's role is as far as possible to curate - to follow the lead of the computers or to make choices thrown up by the computers. But we also need our musical to make sense so where - dialogue for example - we know that computers can't generate language which makes sense or advance computer derived plot sufficiently, Ben and Nathan then become writers not curators.

(7) We are completely transparent (among ourselves, to our audiences, to Sky etc.) about what is and what isn't computationally generated -partly for our own compliance and story-telling purposes and partly because the scientists behind all the computational processes that are being used are doing proper publishable science. Both require that we are scrupulous about documenting and footnoting everything we do so that we can keep track of everything as we go. We will also endeavour to acknowledge, understand and communicate accurately and fairly the motivations, methodologies and processes of the scientific partners whose work has contributed to this project and to maintain a dialogue with the partners where that might either further the experiment or is required by them to augment or complete their research assignments for publication or academic assessment.

(8) We have inserted a development week into the schedule to workshop the show - and then to test parts of it on an audience who won't know that it is computationally derived. This development week - and the filming of it - might represent the stage of maximum 'purity', in terms of pushing the boundaries of the computationally generated and inspired material. We let the company, guided by workshopping discussion, experiment with the material, in order collaboratively to judge its fitness in itself and as part of the whole. The lyric generator, lyric word clouds, back-catalogue of corpus-generated music and the Paris music generator/editor are all tools that will be available during the development week (and beyond), and these should be employed where changes to the material or new material are required.

(9) Manifestly humans could write a 'better' musical than computers who have never been set to that task before. But where computers generate material that should always be the first port of call. Editing, repeat iterations, computer rewrites, cut and paste to impose shape, form, rhyme, sense, mix and match are of course all allowed, as part of a computer collaborative process. Doing that repeatedly is of course as or more time-consuming than writing it from scratch - and almost certainly, currently, less good. But that's the point of the experiment and needs to be honestly and fully

undertaken, within the constraints of the schedule/available time. We can't make our desire to stage the best musical trump the principal objective - which is that the company are staging computationally generated material with as minimal human intervention as possible. If the human intervention is significant then we need to see where, how and why - because that is what the story is all about.

(10) There's an obvious hierarchy of maximum 'purity' which can guide what to do. But we also need to meet deadlines and have a show that opens in late Feb. So many of these choices need to be made by Ben and Nathan and others in the company who, abiding by these principles and recording their choices for transparency purposes, otherwise need to be left to get on with it.

(11) Where this process is liberating / painful / humorous / surprising / impossible / scary / cruel / disturbing we record that honestly and respectfully. We want to experience what it is for artists to be confronted with these new tools and honestly engage with them.

(12) These principles should help to guide choices and act as natural constraints for the project (and avoid having to impose some kind of external referee).'

Appendix B: Reviews of Beyond the Fence

Table 3. The twenty reviews of *Beyond the Fence* sourced via search, for analysis in this paper

Source	Link
	National newspapers
Guardian	http://www.theguardian.com/stage/2016/feb/28/beyond-the-fence-review-computer-created-musical-arts-theatre-london
Telegraph	http://www.telegraph.co.uk/theatre/what-to-see/beyond-the-fence-arts-theatre-review-computer-says-so-so/
Financial Times	http://www.ft.com/cms/s/0/5f993b32-dee2-11e5-b67f-a61732c1d025.html
Independent	http://www.independent.co.uk/arts-entertainment/theatre-dance/reviews/beyond-the-fence-arts-theatre-review-despite-my-reservations-i-was-won-over-a6900836.html
The Times	http://www.thetimes.co.uk/tto/arts/firstnightreviews/article4702133.ece
	Specialist theatre publications
The Stage	https://www.thestage.co.uk/reviews/2016/beyond-the-fence-review-at-arts-theatre-london-futuristic-composition-with-traditional-problems/
What's on Stage	http://www.whatsonstage.com/london-theatre/reviews/beyond-the-fence-arts-theatre_39847.html
The Reviews Hub	http://www.thereviewshub.com/beyond-the-fence-arts-theatre-london/
Musical Theatre Review	http://musicaltheatrereview.com/beyond-the-fence-arts-theatre/
West End Frame	http://www.westendframe.com/2016/02/review-beyond-fence-at-arts-theatre.html
British Theatre Guide	http://www.britishtheatreguide.info/reviews/beyond-the-fenc-the-arts-theatr-12609
Theatreworld	http://www.theatreworldim2.com/#!beyond-the-fence-arts-theatre/fcmry
West End Wilma	http://www.westendwilma.com/beyond-the-fence-review/
BroadwayBaby	http://www.broadwaybaby.com/shows/beyond-the-fence/710587
Carns Theatre Passion	http://carnstheatrepassion.com/2016/02/27/beyond-the-fence-arts-theatre-west-end-until-5th-march/
	Events listings websites
Londonist	https://londonist.com/2016/02/computer-penned-musical-beyond-the-fence-reviewed
There Ought To Be Clowns	http://oughttobeclowns.blogspot.co.uk/2016/02/review-beyond-fence-arts.html
Time Out	http://www.timeout.com/london/theatre/beyond-the-fence
	Technical blogs
Engadget UK	http://www.engadget.com/2016/03/02/beyond-the-fence-computer-generated-musical/