Empirical Evidence of the Limits of Automatic Assessment of Fictional Ideation

A. Tapscott[1⋆], J. Gómez[1], C. León[1], J. Smailović[2], M. Žnidaršič[2], P. Gervás[1]
[1]Facultad de Informática, Universidad Complutense de Madrid
[2]Department of Knowledge Technologies, Jožef Stefan Institute

Abstract. Automatic evaluation of fictional ideation systems and their output is a topic relevant to Computational Creativity. Models and techniques have been proposed for this task, but their applicability is limited to the field of fictional ideation. In this paper we describe an evaluation procedure for fictional ideation, which compares human validation of the ideas with a number of automatically generated metrics obtained from them. We report on the observed limits of this procedure. The results suggest that, besides technical limitations, providing a stable evaluation method is fundamentally incomplete unless the full creative phenomenon is modelled, including aspects that are beyond current technical capabilities.

Keywords: Automatic evaluation, ideation, empirical study, narrative, computational creativity

1 Introduction

Evaluation of creative processes and artefacts is key to computational creativity. Explicitly reflecting on the relative value and novelty is crucial if machines are to produce content that would be *deemed creative* [6]. As such, addressing evaluation is fundamental for computational creativity that can successfully fulfill human needs.

This crucial aspect contrasts with the relative scarcity of systems explicitly generating rich evaluation of their own generated material or inner processes. Some systems arguably control the quality of their artifacts by carrying out a process that ensures a minimum relative quality, but an explicit evaluation arguably represents a qualitative advantage, both theoretical (as studied by computational creativity frameworks [29]) and practical ([4]).

Although the semantics of creativity are elusive and usually problematic, the vision that quality and novelty influence the perception of the creativity of an artifact (at least from the point of view of observation) is commonly accepted. Still, quality and novelty vary depending on the domain and context. Theoretical

⋆ Supported by the project WHIM (611560) and PROSECCO (600653), funded by the European Commission, Framework Program 7, the ICT theme, and the Future Emerging Technologies FET program.

discussion on this exists and it is seminal in the field [1, 2], while other works attempt to offer either formal or procedural techniques for evaluating creativity [18, 25, 30]. These efforts address the evaluation of creativity in generic terms, and they are of limited applicability for the evaluation of the quality of specific artifacts generated automatically. It might be the case that the assumption that there is a global definition of creativity applicable to every creative domain is not possible, but we still need more empirical evidence supporting whether this is so.

Moreover, even when working within a domain in which there is an agreed definition of characteristics assumed to play a role in creativity (let us say *quality*), addressing explicit automatic evaluation can be a costly task, even more costly than creating the generative system that is being evaluated. It is not uncommon that being able to generate appropriate artefacts is doable, while yielding an explicit, measurable evaluation is not (for instance, in images generated by evolutionary computing [15]).

This paper reports on an empirical study in which the output of an automatic ideation system is assessed by computational means. When compared to human evaluation, the conceptual and practical limits of the approach were evidenced. This led to an in-depth analysis of the challenges, which is provided in Section 5.

2 Previous Work

While all scientific exploration requires thorough evaluation of the steps taken, doing so in creativity represents a challenge. How to assess creativity itself is a commonly discussed aspect of the whole phenomena of creative generation. While most authors agree on the correlation between a number of features and the perception of creativity, there is no consensus either on what these features are or how they really correlate. Moreover, adding computers to the problem makes it even more difficult to know whether a system has been successful or not. There is still a debate on what parts should be evaluated, the influence of the programmer on the output, the very definition of creative behavior, the decision of whether to focus on the process or the artifacts (or both), and many others.

The few examples present in the literature describing actual evaluation of automatic creative systems usually focus on less ambitious, more measurable aspects. This makes these systems less useful from a general perspective, but they nonetheless provide insight on the current capabilities of computer systems to assess their own production.

There is, however, a number of proposals that try to provide guidelines to evaluate creative systems. For instance, Ritchie [24, 25] addresses the issue of evaluating when a program can be considered creative by outlining a set of empirical criteria to measure the creativity of the program in terms of its output. He makes it very clear that he is restricting his analysis to the questions of what factors are to be observed, and how these might relate to creativity, specifically stating that he does not intend to build a model of creativity. Ritchie's criteria

are defined in terms of two observable properties of the results produced by the program: *novelty* (to what extent is the produced item dissimilar to existing examples of that genre) and *quality* (to what extent is the produced item a high-quality example of that genre). To measure these aspects, two rating schemes are introduced, which rate the typicality of a given item (item is typical) and its quality (item is good). Another important issue that affects the assessment of creativity in creative programs is the concept of *inspiring set*, the set of (usually highly valued) artifacts that the programmer is guided by when designing a creative program. Ritchie's criteria are phrased in terms of: what proportion of the results rates well according to each rating scheme, ratios between various subsets of the result (defined in terms of their ratings), and whether the elements in these sets were already present or not in the inspiring set. Ritchie's criteria have been used in subsequent evaluations of creative systems output [7, 21, 8].

Pease et al. [19] discuss relevant factors to evaluating systems in terms of creativity. The proposed framework mainly takes into account input provided, output produced and process employed. Each of these categories are detailed in depth, detailing their required measures. Before detailing the measurement methods, Pease et al. provide assumptions regarding creativity, also admitting their 'somewhat arbitrary' nature. The evaluation tests proposed deal with two main aspects: how close does the test predict human evaluation of creativity and how possible and practical it is to apply the test to a system. Overall, this work suggests that the very definition of creativity is subjective and that evaluating systems in a general way is problematic.

Colton et al. [5] propose an extension of Ritchie's criteria [24] that attempts to determine the impact of the input data on the creative artifact produced by a system. This more agnostic approach attempts to obtain an objective measure by comparing the output of the system to the inspirational material used as input. This investigation attempts to discriminate systems that overfit or shuffle input data (fine-tuning) instead of producing genuine novel artifacts. Among other conclusions, the authors state that comparing creative systems might not be viable, suggesting their criteria to be used as guidelines for program construction rather than post-hoc evaluation.

The creative tripod framework, proposed by Colton [3], is built around the premise that a creative system must demonstrate skill, imagination and appreciation. These qualities are not required to be possessed by the system, but rather to be perceived as possessed by the system. This is an important remark by Colton to avoid debates around the definition of creativity. The framework also includes the programmer, the system and the consumer, however Colton is only interested in the program's behavior.

Pease and Colton [18] propose an alternative to the Turing Test to assess computational systems' creativity, the FACE (Frame, Aesthetic, Concept, Expression of concept) and IDEA (Iterative Development Execution Appreciation) model. The model includes creative acts and audiences, with relevant measures such as popularity, appeal, provocation, opinion, subversion and shock. Putting the focus on the reaction produced by the creative artifact, this model attempts

to avoid the shortcomings of the Turing Test by going further than merely assessing the capacity of a creative system to imitate human behavior. By including the audience into the model, this approach acknowledges the highly subjective nature of creativity evaluation.

SPECS [9], introduced by Jordanous as "a standardised and systematic methodology for evaluating computational creativity", represents a substantial effort to provide a standard for evaluating the creativity of a system in the field of computational creativity and address the multi-faceted and subjective nature of creativity. Its flexible nature allows SPECS to adapt to the demands of the researchers' field, applying the required demands and standards. The methodology informs researchers of their system's strength and weaknesses, providing useful feedback for achieving creative results.

2.1 Evaluation of Automatically Generated Narrative

Automatic generation of narratives has been a long-standing goal of Artificial Intelligence since its very beginning. There are a number of systems described in the literature, but the evaluation of these systems – be it its output, its creative process or whatever other aspect – is seldom found. This is most likely due to the fact that the average quality or variety of the generated stories is not really comparable to those written by most humans, not necessarily professional writers.

The Mexica system [23] includes procedures for the dynamic assessment of the novelty of a story in progress with respect to previously known stories. Novelty is considered in terms of how the stories differ in terms of the actions they include and their frequency of appearance.

In Pérez et al [22] three different characteristics are considered as relevant for measuring story novelty: sequence of actions, structure of the story, and use of characters and actions.

Peinado & Gervás [20] carried out an empirical study of how generated stories were perceived by a set of human volunteer evaluators. Human judges blindly compared one of the generated basic stories to two alternatives: one rendered directly from a stored fabula of the knowledge base and another randomly generated. Values were collected for: *linguistic quality* (how well is the text written), *coherence* (how well is the sequence of events linked), *interest* (how interesting is the topic of the story for the reader) and *originality* (how different is the story from others).

León & Gervás [11] propose a model, intended as a tool to drive automatic story generation, of how quality is evaluated in stories. This paper proposes a computational model for story evaluation in which an evaluation function receives stories and outputs a value as the rating for that story. The value for this function is computed from values assigned to: accumulation of contributions from individual events depending on the meaning of the event – aspects such as whether the reader wants to continue reading the story, or how much danger or love the reader perceives in the story –, appearance of patterns or relationships between the events of a story – aspects such as causality, humour or relative

chronology – and inference – which captures the ability to interpret stories by adding material to explain what they are told even if it is not explicitly present in the story. The evaluation function has been implemented as a rule based system.

Ware, Young et. al. [27] propose a formal model for narrative conflict with seven dimensions from various narratological sources meant to aid in distinguishing one conflict from another: participant, subject, duration, balance, directness, intensity and resolution. Their experimental results [28] suggest the model predicts these seven dimensions of narrative conflict similarly to human criteria. Their good results predicting human-perceived narrative conflict suggest a similar approach may be viable for measures related to creativity.

3 Evaluating Automatic Ideation

Original ideation is central to any creative process. Coming up with innovative ideas that potentially trigger the creation of new material is fundamental to human creativity. It is not uncommon to focus creative processes on the identification of a single, valuable idea that unlocks new paths leading to finished artifacts. Although human creative teams usually rely on pure ideation to foster creativity, there have only been a few small, ad-hoc studies of how to automate ideation until recent times. Section 3.1 describes an effort to provide a system able to produce novel ideas.

3.1 The What-If Machine

Llano et al. have recently proposed an automatic ideation system [13, 14, 12]. This computational system is designed to produce relatively valuable and novel ideas autonomously. This system, the *What-If Machine*[1], includes a module for analysing the ideas and generating narrative metrics, and a module for computing a predictive machine learning model. This model is trained against collected human evaluations of what-ifs, and is intended to learn a robust function from narrative metrics to perceived overall quality. Two main hypotheses guide the design of the What-if Machine and the presented research:

1. There is a strong correlation between the perceived *overall quality* and the perceived *narrative potential*, in the sense that if the audience perceives high narrative potential, it will also perceive a high overall quality. The overall quality is defined in terms of the analyzed response from humans (i.e. no specific model beyond what humans say about quality is assumed), and the narrative potential is assumed to be directly proportional to the amount and quality of the stories a certain what-if can trigger or inspire.
2. There is a set of computable metrics whose values correlate (directly or indirectly) with the overall quality and the narrative potential.

[1] The What-if Machine: `http://www.whim-project.eu/`.

The What-If Machine is, to the best of our knowledge, the only attempt to implement a computer system able to produce novel what-if ideas. The What-If Machine is a distributed computer system in which several modules collaborate in order to output rendered what-ifs. Five modules compose the system:

1. The **ideation module** produces, using a knowledge base, what-if ideas formalized as *mini-narratives*.
2. The mini-narratives are fed into the **narrative-based metric generation**, which generates values for a set of metrics which hypothetically have a correlation with human perception of quality. These metrics are based on narrative properties of the what-ifs.
3. The mini-narratives, now enriched with its corresponding metrics, are sent to a **crowd-sourcing evaluation module**, which applies machine learning to create and refine models for predicting overall quality against human ratings.
4. The **world view** creation, providing knowledge for what-if generation, story creation and metric computation.
5. The finished, filtered what-ifs are finally passed to a **rendering module**, which creates artifacts from the final what-ifs (stories, texts or images, for instance).

A subset of the What-If Machine (modules 1, 2 and 3) was used to generate the material for the study, which is described in detail in Section 4.

4 Study

A pilot study was performed to determine the feasibility of predicting the perceived *quality* and *narrative potential* in the artifacts created by a computable creative system. Both magnitudes have been introduced in the previous section, and in order to avoid influencing our subjects, no definition for them is provided in the questionnaires (as seen in Fig. 1). This naive approach is a result of our focus on the model and its capability to predict human assessment instead of introducing our own views or definitions. The study was conducted to obtain the human rating of perceived *quality* and *narrative potential*.

Using both measures, a machine learning process will search for correlations between some metrics (detailed in the next section) and the perceived *quality* and perceived *narrative potential*. This should allow us to determine what measures are relevant to predict human-perceived *quality* and *narrative potential* to produce what-ifs that present both qualities to human observers.

4.1 Metrics

Since we have no certainty about what metrics extracted from each what-if's mini-narrative may impact over the perceived *quality* and *narrative potential*, we focused on generating the maximum amount of computable features. The impact of these features on the perceived *quality* and *narrative potential* may be

obtained with machine learning techniques (we refer to these features as *metrics*). This approach is similar to the one used by Nowak for image classification [17] that generates a high number of arbitrary features from each image.

A mini-narrative is a structure that contains a set of **narrative points** linked to schemas like *setting* or *resolution*. Each **narrative point** is a set of **narrative statements** that provide information about characters or events through predicates (e.g. *dog is old* or *dog learns to play a piano*). **Narrative statements** may be related to one another (*caused by* or *inferred by* another statement).

The next list includes the set of implemented features along with their description:

- **Length**: mini-narrative **narrative points** amount.
- **SettingQuality**: Amount of schemas divided by 3.
- **ExplicitFact**: the amount of **narrative statements** in the mini-narrative.
- **RatioCharacters**: the character/statement ratio.
- **Originality**: hits returned by the full text of the mini-narrative in the *Bing* search engine.
- **OriginalityAccurate**: hits returned by the **exact** full text of the mini-narrative in the *Bing* search engine.
- **Divergence**: average hits returned by the mini-narrative statements in the *Bing* search engine.
- **DivergenceMinimum**: minimum hits returned by the mini-narrative statements in the *Bing* search engine.
- **Evolution**: amount of `learnTo` predicates found in the mini-narrative.
- **Handicap**: amount of negated `capableOf` predicates found in the mini-narrative.
- **InterestingLife**: amount of negated `doesFor` predicates found in the mini-narrative.
- **TotalStoriesGenerated**: amount of stories generated by the story generator from the current mini-narrative.
- **StoryCharacters**: average number of characters in the generated stories.
- **Names**: StanfordNLP [16] queries for the what-if's names.
- **NamesRatio**: *Names*/*ExplicitFact* ratio.
- **Valence**: Sum per statement, each statement codified as +1 if a fact is positive, -1 if negative and 0 otherwise).
- **ValenceAverage**: *Valence*/*ExplicitFact* ratio.
- **JointWordsProbability**: joint probability average for each set of words using *ngrams*. For this metric we use the Project Oxford[2] services.
- **JointWordsProbabilityMinimum**: the minimum joint probability for the set of words using *ngrams* from Project Oxford.
- **RealityDistortionRatio**: events in the mini-narrative that negate a fact from the *knowledge base* are considered a *reality distortion*. This metric provides the *reality distortion* amount/*ExplicitFact* ratio.

[2] `https://www.projectoxford.ai/`

- **FictionalAdditionsRatio**: any event in the mini-narrative that is missing from the *knowledge base* is considered a *fictional addition*. This metric provides the *fictional addition* amount/*ExplicitFact* ratio.
- **FictionalRatio**: *reality distortion* amount plus *fictional addition* amount/*ExplicitFact*.
- **ResolutionTriggerRatio**: *resolution events* solve *conflicts* from the mini-narrative. Provides the *resolution event* amount/*ExplicitFact* ratio.
- **MainCharacterEventsRatio**: *protagonist statements* are statements in which this actor plays any role. This metric provides the *protagonist statement* amount/*ExplicitFact* ratio.

4.2 Methodology

A set of 890 what-ifs were generated by the What-If Machine. All of their source mini-narratives were processed by the metric generation system. A total of 15 different questionnaires were created, each including 10 what-ifs rendered as text from the original set of 890. 150 what-ifs were included in the evaluation set. 101 volunteers received a link that randomly redirects to one of the 15 possible questionnaires through email. Given the simplicity of the questions, Google Forms was our platform of choice. The platform was robust and stable and all of the answers were successfully stored in a Google Sheet document automatically. There was no active supervision for each subject given the remote nature and limitations of the Google Forms platform.

4.3 Questionnaire

The questionnaire informed subjects about their participation in a study related to computer-generated content (Figure 1). Some demographic information was queried (age, gender and English level) and then they were asked to evaluate the overall quality (on a 0-5 Likert scale) of each what-ifs plus its narrative potential (yes/no binary answer). A text box accepting any comment was also provided in order to gather additional qualitative information.

> You are about to evaluate some of the preliminary results of the "WHIM: The What-If Machine" research project from the European Union. The overall objective of the What-If Machine is to automatically generate fictional ideas with cultural value. You will be presented a number of what-if style ideas and we kindly ask you to rate them according to the following features:
> – Overall quality: from 0 (no quality) to 5 (superb quality). – Narrative potential (yes/no). – Any observation you can provide.
> Completing the questionnaire should not take more than 10 minutes. We really appreciate your contribution to the project.

Fig. 1. Information presented to the user in the evaluation questionnaire.

4.4 Results

101 subjects participated in the study. Statistical analysis of the results revealed no significant differences between evaluators in terms of English level, age or gender. For instance, the quality (Q) for gender yielded $\mu(Q)_{male} = 2.66$, $\sigma(Q)_{male} = 0.75$; $\mu(Q)_{female} = 2.69$, $\sigma(Q)_{female} = 0.89$. The corresponding results for English and age are comparable.

Questionnaires provided 1,007 *Quality* and 1,004 *Narrative Potential* rankings for the 150 *What-Ifs* used. *What-Ifs* were ranked between 1 and 27 times. For the *Narrative Potential* (P) measurements, we mapped "Yes" to +1, "Not sure" to 0, and "No" to -1. Overall measures resulted in $\mu(Q) = 2,4$ and $\sigma(Q) = 1,3$ for *Quality* and $\mu(P) = -0,05$ and $\sigma(P) = 0,89$ for *Narrative Potential*. Individual *What-Ifs* aggregated ranking values were used for calculating:

- Pairwise correlations between perceived *Quality* and perceived *Narrative Potential*, perceived *Quality* or perceived *Narrative Potential* and the metrics, and between individual metrics.
- Global measure of attribute importance for these metrics in predictive modeling of the average perceived *Quality* or perceived *Narrative Potential*.

Pairwise correlations Metrics that provided the same values for all *What-Ifs* in the dataset were discarded. Correlation coefficients were calculated with the Pearson Product-Moment. There is a strong positive correlation between *Quality* and *Narrative Potential* averages (0.83) and medians (0.758). As seen in table 1, both measures correlate positively with some metrics, such as *MainCharacterEventsRatio* and *RatioCharacters* and correlate negatively with others, such as *ExplicitFact* and *Length*.

Table 1. The correlation coefficient between average/median *Quality (Q)* or *Narrative Potential (P)* labels and the metrics. The values are sorted by correlation coefficient values of the average *Quality*.

	Avg Q	Mdn Q	Avg P	Mdn P
MainCharEventsRatio	0.371	0.346	0.379	0.329
RatioCharacters	0.354	0.296	0.368	0.307
ResolutionTriggerRatio	0.342	0.303	0.305	0.261
TotalStoriesGenerated	0.312	0.250	0.321	0.264
JointWordsProbMin	0.308	0.289	0.367	0.314
...	...	...	...	...
ValenceAverage	-0.219	-0.188	-0.296	-0.249
ValenceSum	-0.258	-0.234	-0.323	-0.276
StoryCharacters	-0.283	-0.269	-0.327	-0.285
ExplicitFact	-0.379	-0.336	-0.406	-0.345
Length	-0.379	-0.336	-0.406	-0.345

Importance for Predictive Modeling In order to determine the importance of each metric in predicting perceived *Quality* and *Narrative Potential* we used the Relief measure [10, 26], which is a method commonly used for feature selection in machine learning. This measure does not assume independence among the metrics, but takes their possible interdependence into account. The more the Relief scores are positive, the more a metric contributes to prediction of a target value (in our case, the value of average Quality or the average Potential). The ones that scored close to zero or negative are irrelevant and those with negative values have even a negative impact.

According to the results in Table 2 it seems that most of the metrics have no use in predictive models of average Quality. For the average Narrative Potential, however, most of the metrics seem to be slightly informative . According to Relief ranks for the metrics results, usefulness of the metrics for average Quality is to some extent inversely proportional to their usefulness for the average Narrative Potential. The absolute values of the Relief scores depend on the characteristics of data and the parameters of the assessment, which makes it difficult to use absolute thresholds for judgements on the relevance of features. However, a strong correlation among the Quality and Narrative Potential values and a mismatch of the Relief scores of metrics for these two targets provide an indication that also the contributions of the positively scored metrics are likely to be too low to be considered relevant.

Table 2. Relief measure results for average Quality (Relief Avg Q) and average Narrative Potential (Relief Avg P). Rows sorted by Relief Avg Q. The best three results are in bold and the worst three are in italics.

Metric	Relief Avg Q	Relief Avg P
Handicap	**0.027**	*-0.009*
MainCharacterEventsRatio	**0.007**	0.004
NamesRatio	**0.001**	0.006
DivergenceMinimum	0.000	0.000
JointWordsProbabilityMinimum	0.000	*0.000*
Divergence	0.000	*0.000*
Originality	-0.006	0.013
...	...	...
FictionalAdditionsRatio	-0.075	0.028
InterestingLife	-0.116	**0.045**
TotalStoriesGenerated	-0.116	**0.045**
OriginalityAccurate	-0.126	0.024
FictionalRatio	-0.142	**0.039**
RatioCharacters	-0.142	0.039
SettingQuality	*-0.147*	0.024
Names	*-0.147*	0.024
ValenceSum	*-0.174*	0.033

5 Relative Limits of Evaluating Quality

The results previously presented evidence that there is a strong correlation between narrative potential and perceived overall quality of a what-if, which indicates that focusing on narrative plausibility as one of the main factors of quality can lead to better results. Moreover, some of the metrics are weakly correlated to narrative potential. However, these results are still inconclusive, and there is a number of aspects worth mentioning for their influence on the results.

Automatically generating stories and computing useful values for metrics is heavily dependent on the available knowledge. The outcome of the system is constrained by the use of ConceptNet. The amount of relations that can be safely used in ConceptNet is small and the richness and depth of the chains of properties is limited regarding to its use as a source for narrative processing. This makes it necessary to address knowledge management from a different perspective. The WHIM project currently includes a whole module for providing robust knowledge to the rest of the modules, and the impact of the application of this subsystem on the creation and evaluation of what-if ideas will be reported once the results are ready.

The generation process (for the what-ifs, the stories and the metrics) strongly influences the overall outcome. Many design decisions have been taken in order to provide a working, implemented prototype able to generate actual what-ifs, and these decisions set the kind of what-ifs generated, the complexity of the stories and many other aspects. The provided results are then the outcome of a specific implementation which does not claim any generality. However, the approach itself (namely the generation-metric computation-evaluation process) is presented as a generally applicable method for producing novel what-if ideas.

The used metrics for labeling narrative properties do not cover all computable features. There is a large number of aspects that can be extracted from a what-if, and the narrative-based feature extraction module of the What-If Machine does not currently provide coverage for all of them. This is considered to be not strictly relevant with regard to the methodology and scope of the study. To test the second hypothesis (the existence of a correlation between a certain set of metrics and the overall quality and plausibility), the metrics must be improved. For that purpose, the presented study gives valuable insight on which direction to go next.

The weak correlation between our metrics and the quality perceived by humans suggested that considering more sophisticated metrics was necessary. Some of them were considered:

1. **Humanization**: An approximation of how much human-like the main character is, assuming that fictional scenarios use characters that, while behaving like humans, can be non-human.
2. **Empathy**: How much empathy will a reader feel about the characters.
3. **Tragedy**: The amount of tragedy in the story.
4. **Reality**: How real and current the context is. An approximation of fictionally in terms of context.

5. **TimeSpan**: The time span the story covers. It could be minutes, days or years.

Modelling and implementing these metrics proved to be beyond technical capabilities because it required complex, rich knowledge bases (1, 4), reliable text understanding systems (5), sophisticated emotional models (2) or formal versions of narratological models (3). All of these resources are currently not available.

6 Conclusions

The current paper has presented a pilot study trying to gain insight on two hypotheses, namely that (1) human evaluation on overall quality of what-if ideas correlates to the perception of narrative potential and that (2) there is a set of computable metrics that also correlate to this perception. The study has evidenced that there is a strong correlation between quality and narrative potential for humans (1), but failed to prove such a strong correlation between the current metrics and the human ratings. These results have been analysed and discussed in terms of the limited potential of the current implementation of both the fictional ideation procedure and the method employed to evaluate it. Actual implementations lack the required complexity to approximate evaluations with a relatively acceptable level of accuracy, mainly due to the limited technical capabilities of current computational solutions.

References

1. Boden, M.: Computational Models of Creativity. Handbook of Creativity pp. 351–373 (1999)
2. Boden, M.: Creative Mind: Myths and Mechanisms. Routledge, New York, NY, 10001 (2003)
3. Colton, S.: Creativity Versus the Perception of Creativity in Computational Systems. Proceedings of the AAAI Spring Symposium on Creative Systems (Colton 2002), 14–20 (2008)
4. Colton, S.: The painting fool: Stories from building an automated painter. Computers and Creativity 9783642317, 3–38 (2012)
5. Colton, S., Pease, A., Ritchie, G.: The effect of input knowledge on creativity. Technical Reports of the Navy Center for (2001), http://www.inf.ed.ac.uk/publications/online/0055.pdf
6. Colton, S., Wiggins, G.: Computational creativity: The final frontier? ECAI (2012)
7. Gervás, P.: Linguistic creativity at different levels of decision in sentence production. In: Proceedings of the AISB 02 Symposium on AI and Creativity in Arts and Science, 3rd-5th April 2002, Imperial College. pp. 79–88 (2002)
8. Haenen, J., Rauchas, S.: Investigating artificial creativity by generating melodies, using connectionist knowledge representation. In: The Third Joint Workshop on Computational Creativity (2006), http://ccg.doc.gold.ac.uk/events/ecai06/proceedings/Haenen.pdf

9. Jordanous, A.: A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. Cognitive Computation 4(3), 246–279 (2012), http://dblp.uni-trier.de/db/journals/cogcom/cogcom4.html#Jordanous12
10. Kira, K., Rendell, L.: A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine learning. pp. 249–256 (1992)
11. León, C., Gervás, P.: The Role of Evaluation-Driven rejection in the Successful Exploration of a Conceptual Space of Stories. Minds and Machines 20(4), 615–634 (2010)
12. Llano, M.T., Colton, S., Hepworth, R., Gow, J.: Automated Fictional Ideation via Knowledge Base Manipulation. Cognitive Computation pp. 1–22 (2016)
13. Llano, M.T., Cook, M., Guckelsberger, C.: Towards the automatic generation of fictional ideas for games. Experimental AI in ... (2014)
14. Llano, M.T., Hepworth, R.: Automating fictional ideation using ConceptNet. Proceedings of the ... (2014)
15. Machado, P., Martins, T., Amaro, H., Abreu, P.: Beyond interactive evolution: Expressing intentions through fitness functions. Leonardo (2015)
16. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations). pp. 55–60 (2014)
17. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. pp. 490–503. Springer Berlin Heidelberg (2006)
18. Pease, A., Colton, S.: On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. AISB 2011: Computing and Philosophy pp. 15–22 (2011)
19. Pease, A., Winterstein, D., Colton, S.: Evaluating machine creativity. In: Workshop on Creative Systems, 4th (2001)
20. Peinado, F., Gervás, P.: Evaluation of Automatic Generation of Basic Stories. New Generation Computing, Computational Paradigms and Computational Intelligence. Special issue: Computational Creativity 24(3), 289–302 (2006)
21. Pereira, F.C., Hervás, R., Gervás, P., Cardoso, A.: A Multiagent Text Generator with Simple Rhetorical Habilities. In: Proc. of the AAAI-06 Workshop on Computational Aesthetics: AI Approaches to Beauty and Happiness, July 2006. AAAI Press (2006)
22. Pérez, R.y., Ortiz, O., Luna, W., Negrete, S.: A system for evaluating novelty in computer generated narratives. Creativity (2011)
23. Pérez y Pérez, R.: MEXICA: A Computer Model of Creativity in Writing. Ph.D. thesis, The University of Sussex (1999)
24. Ritchie, G.: Assessing creativity. In: Proceedings of the AISB Symposium on AI and Creativity in Arts and Science. pp. 3–11. York, UK
25. Ritchie, G.: Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds & Machines 17, 67–99 (2007)
26. Robnik-Šikonja, M., Kononenko, I.: An adaptation of relief for attribute estimation in regression. In: Machine Learning: Proceedings of the Fourteenth International Conference (ICML 1997). pp. 296–304 (1997)
27. Ware, S.G., Young, R.M.: Validating a Plan-Based Model of Narrative Conflict. In: Proceedings of the International Conference on the Foundations of Digital Games. pp. 220–227. ACM Press, New York, New York, USA (2012)
28. Ware, S.G., Young, R.M., Harrison, B., Roberts, D.L.: Four Quantitative Metrics Describing Narrative Conflict.pdf. pp. 18–29. Springer Berlin Heidelberg (2012)

29. Wiggins, G.: A preliminary framework for description, analysis and comparison of creative systems. Knowledge-Based Systems 19(7) (2006)
30. Wiggins, G.: Searching for Computational Creativity. New Generation Computing, Computational Paradigms and Computational Intelligence. Special Issue: Computational Creativity 24(3), 209–222 (2006)