D5.3
Theory, methodology, system, and evaluation criteria

Authors	Joseph Corneli, Alison Pease, Tarek R. Besold, Roberto Confalonieri, Danny de Jesús Gómez Ramírez, Maximos Kaliakatsos-Papakostas, Marco Schorlemmer, Asterios Zacharakis
Reviewers	

Grant agreement no.	611553
Project acronym	COINVENT - Concept Invention Theory
Date	November 7, 2016
Distribution	PU/RE/CO

Disclaimer

The information in this document is subject to change without notice. Company or product names mentioned in this document may be trademarks or registered trademarks of their respective companies.

The project COINVENT acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant number 611553.

Abstract

This deliverable evaluates the contributions of the COINVENT project to the theory and practice of computational concept invention. We apply several sets of criteria for the quality of a theory to review our contribution. We then assess the system itself, again using multiple sets of criteria for evaluating work in computational creativity. We include a meta-evaluation of these criteria, based on interviews with the researchers who made the most significant contributions to the evaluation activities detailed in earlier sections.

Keyword list: **evaluation**, **theory**, **methodology**, **criteria**

Changes[1]

Version	Date	Author	Changes
0.1	07.11.16	Joseph Corneli	Move text into this template

[1] The list of changes is optional and can be removed in the final version.

Executive Summary

This deliverable evaluates the contributions of the COINVENT project to the theory and practice of computational concept invention. We apply several sets of criteria drawn from the philosophy literature that can be used to assess the quality of a theory. We make the case that the theory of blending developed in this project is a “good theory” in that it matches these desiderata:

- It is as general as possible
- It explains more than it set out to explain
- It does not assume what it sets out to explain

We then assess the system itself, using multiple sets of criteria for evaluating work in computational creativity. Specifically, we evaluate the realised system’s ability to be “creative,” using an assessment framework of 14 criteria devised by Anna Jordanous. Among these criteria, the system’s strengths were Generation of Results, Originality, and Value. Several other criteria, like “General Intellect”, could not be projected onto the system. This was in line with the project’s theoretical and development goals, since COINVENT focused on developing “a computationally feasible, cognitively-inspired, formal model of concept invention” and not a general-purpose AI.

Validation of the realised goals is found in both mathematical and musical case studies. The system assisted in creating at least one novel and valuable concept in mathematics; and a user study showed that the system is able to create perceptually meaningful blends based on self-evaluation of its outcome. Thus, within the specific domains we worked with, and given its limitations as a concept-invention tool, the system was seen to perform well.

We describe the successive prototypes developed in the project, showing their increasing autonomy and sophistication with tasks like those mentioned above.

We describe the relationship between the developed system, and point out some limitations of the system that could be addressed in future work – thereby continuing the trajectory outlined in earlier sections. In addition to technical work, we highlights directions for research that would:

- model and incorporate additional familiar aspects of the mathematician’s toolkit into computational systems;
- develop additional iterative methods for generation-and-testing of system outputs in the various branches of the system; and
- connect creativity evaluation research that have been applied and reviewed above to the study, use, and development of social machines

Contents

1 Introduction

The COINVENT project has the aim of developing "a computationally feasible, cognitively-inspired, formal model of concept invention, drawing on Fauconnier and Turner's theory of conceptual blending" [51]. Theoretical aspects of the project were to be grounded in a realised computational implementation, which would be piloted in the domains of mathematics and music. Specifically, the COINVENT system would be able to invent new mathematical and musical concepts, or "rationally reconstruct" old ones in a way that allows us to more completely theorise their discovery. In these several regards, COINVENT certainly seems to have been successful.

In the present document, we develop a detailed evaluation of the project as a whole, considering both these explicit high-level goals, and the project's broader scientific contributions. In Section 2, we first evaluate our formal model as a theory of combinatorial concept invention. Then, in Section 3, we evaluate the realised system's ability to be "creative," using an assessment framework devised by Jordanous [29]. Here, we conclude that although the system has some of the features Jordanous outlined, it there are several creative qualities (like "General Intellect") that it is not possible to project on the realised system. In Section 4, we describe several experiments that examined user perceptions of blends generated by the music sub-system. One of these studies showed that the system is able to "create perceptually meaningful blends based on self-evaluation of its outcome" [58], which is a useful result towards the rational reconstruction aims of the project. In Section 5, we describe the successive prototypes developed in the project, showing their increasing autonomy and sophistication. Thus, while the system does not yet possess a "General Intellect," we can understand the past and to some extent the future development as part of a developmental trajectory. In Section 6, briefly describe the relationship between the developed system, pointing out some limitations of the system that could be addressed in future development work – thereby continuing the trajectory outlined in earlier sections. Section 7 comprises a meta-evaluation of the various evaluation criteria used in the earlier sections, drawing on interview data. Finally, Section 8 summarises all of the foregoing material, pointing to several new directions for research into modelling creativity.

2 Theory

In order to evaluate our formal model, we will use criteria suggested by [54], [53], [45] and [36] (see Appendix A for a more detailed description of these criteria). The first three consist of Thagard's notions of consilience, simplicity and analogy; Sloman's criteria that a theory *(i)* explain a range of possibilities, and a good theory is *(ii)* definite, *(iii)* general, *(iv)* able to explain fine structure, *(v)* non-circular, *(vi)* rigorous, *(vii)* plausible, *(viii)* economical, *(ix)* rich in heuristic power, and *(x)* extendable; and Popper's criteria of being independently testable, and rich in content. These three sets of criteria overlap in the criteria shown below. A good theory should:

- be as general as possible (Thagard's notion of consilience, and Sloman's third "generality" criterion);
- explain more than it set out to explain (Thagard's notion of simplicity, and Popper's richness in content criterion);

- not assume what it sets out to explain (Sloman's fifth non-circularity criterion, and Popper's independently testable criterion).

Since there is no ready made integrated, coherent, well-motivated and widely accepted set of criteria relevant to evaluating our work, we will focus on these criteria in our evaluation of our formal theory. Furthermore, we shall compare our formal model of concept invention with the most representative computational models proposed in the literature: Sapper [56], Divago [44], HDTP [27], and Alloy [22].

2.1 To be as general as possible

We want our model to be as general as possible. As indicated above, this criterion overlaps with Thagard's notion of consilience and with Sloman's generality criterion. According to Thagard, consilience is a measure of how many observables a theory explains, and the variety and importance of the facts explained. According to Sloman, a theory should explain many significantly different possibilities, preferably some which were not known about before the theory was invented; however it should not explain too many possibilities which have not been shown to exist, i.e., it should not be too general. Both Thagard's consilience and Sloman's generality are related to the *expressivity* of the representation formalism used for modelling. We review the expressivity of several computational models of concept invention below.

Sapper [56] uses a simple formalism to represent complex concepts – namely semantic networks: directed graphs whose nodes represent concepts and whose edges represent binary relations between concepts. Input spaces are subgraphs determined by spreading activation, and a cross-space mapping between activation paths of each input space is established on the basis that they are structurally isomorphic (i.e., node labels may vary, but edge labels have to be identical — this isomorphism requirement can be relaxed a bit by means of so called *slippage* rules). The blended space is obtained by fusing the corresponding nodes of this mapping.

The model of conceptual blending put forward in Divago [44] is more expressive than that of Sapper because it allows for the additional specification, by means of Prolog clauses, of rules, frames and integrity constraints for each input space, which play an important role in the blending process. For instance, in the conceptual blend of a bird with a horse to create the concept of Pegasus [44, p. 110] , an additional *transforming frame* identifies a transformation that occurs during blending, such as the transfer of wings from bird to horse giving it the ability to fly.

Blending as modelled with HDTP [48] goes beyond Divago's expressiveness as it allows for specifying input and blend spaces by means of (many-sorted) first-order theories. The cross-space mapping is also more general than the mere pair-wise identification of correspondences between concept entities, because mappings between theories are established via a restricted form of higher-order anti-unification, providing a more generic way to relate entities between input spaces (if required also allowing for many-to-one or one-to-many mappings between theories). The blends that underlie the grounding metaphors of arithmetic as put forward by [37], for instance, exceed the modelling capacity of Sapper and Divago, but can be recreated with HDTP [27].

The most general formalisms proposed so far for modelling conceptual blending has been the one proposed by Goguen in the context of his work on algebraic semiotics [20]. He proposes to model conceptual spaces by means of sign systems of a particular kind (based on constants representing concepts and relations between them — much as a semantic network), and a conceptual blend as particular kind of pushout in the ordered category of sign systems and semiotic morphisms (structure-preserving partial maps between sign systems). This framework based on category theory allows for capturing many aspect of the phenomenology of blending, such as for instance the partiality of the projections occurring between input spaces and blend space. The conceptual blend underlying the solution of the Buddhist Monk Riddle, for instance, can be nicely modelled in this formalism [21].

Unfortunately, unlike Sapper, Divago and HDTP, Goguen's model has no implemented realisation besides the one based on BOBJ [24], which is not in consonance with the theory of algebraic semiotics. BOBJ cannot model sign systems as defined by [20], only hidden algebras, and morphisms do not represent partial mappings, making the category handled by BOBJ not an ordered category. Utimately, ordinary pushouts are computed, and not the 3/2-pushouts proposed to model for blending. Hence, the expressiveness of the actual computational systems for blending proposed by [22] lies in between Divago and HDTP (as equational logic covers first-oder logic with universal quantification).

In [7], we have suggested a model of conceptual blending that generalises the notion of amalgam as it has been used in case-based reasoning [43]. Following Goguen's intuition, our objective was to be as general as possible by being representation-independent and expressing the core notions of blending in a category-theoretic framework. For this reason our model would cover input spaces as expressed by semantic networks, Prolog clauses, equational theories or first-order theories. We have used CASL [2] to provide concrete descriptions of blending [8, 15, 25] and this specification formalism subsumes the ones just mentioned. But unlike the work of Harrel and Goguen, our model also corresponds to a computational realisation that is in tune with the formal model. Consequently we have been capable of modelling conceptual blends tha lay outside the scope of Sapper, Divago, Alloy or HDTP, as for instance the conceptual blends underlying mathematical concepts such as Prime Ideals or Dedekind Domains [25].

In this sense the theory explains many significantly different cases of concept invention, some which were not explicitly identified as cases of conceptual blending before, but which are coherent with Fauconnier and Turner's model. One paradigmatic example is the identification that a jazz cadence such as the Tritone Substitution can actually be considered as a blend of the historically previously used perfect and Phrygian cadences [15]. Still, the theory is not too general as it retains the constitutive elements of blending put forward by [18]:

- input spaces — modelled as objects in a category. From the fact that we focus on a category, it is implicit that these objects have some structure that is preserved by the morphisms of the category. Our model is neutral about the concrete structure, allowing for many modelling possibilities for input spaces; but since morphisms are based on monospans (see [7]), we need a notion of generalisation that is captured categorically by the notion of subobject. This is in tune that mental spaces are "small conceptual packets constructed as we think and talk".

- cross-space mapping — modelled as a span of two morphisms. Actually this allows for modelling a cross-space *relation*, rather than a mapping. Certain cases of blending, such as the one underlying the Buddhist Monk Riddle, are more naturally modelled using a relation rather than a mapping [50].

- generic space — modelled as the object acting as common source of the two morphisms of the span. It necessary will contain the sub-structure that is common to the initial input spaces.

- selective projection and blend — modelled as an amalgam. Each amalgam amounts to a pair of subobjects of the initial input spaces together with an optimal way to combine them (the pushout). Consequently it represents a selective projection of certain structure originally in the input spaces, and its combination according to the cross-space relation and generic space. No other combination is considered a blend, in accordance to Fauconnier and Turner's model.

- emergent structure — modelled, for instance, as a Yoneda-based creative process. The generality provided by our model makes it possible to relate it with the mathematical model of the creative process proposed by Mazzola et al. [40, 1]. Hence the processes of composition, completion and elaboration can be nicely captured by process of search, colimit computation and reasoning at a distance, as exemplified in the Buddhist Monk Riddle in [50].

Uniformity. As a particular feature aligned with this criterion, our model has a great degree of *uniformity*, in the sense that it is parametrised by a category, who can be conveniently chosen for each case. This gives our model a representation-independent character, so that we were able to explore cases of blending with a varied number of representation formalisms such as OWL (see [35, 33]; [42]) – or a particular sublanguage of thereof, such as the one based on description logic, $\mathscr{E}\mathscr{L}$++ (see [11]) – the specification language CASL (see [15, 8]), or feature structures (see [14]).

2.2 To explain more than it set out to explain

We want our model to explain more than it set out to explain. This criterion overlaps with Thagard's notion of simplicity and with Popper's richness in content criterion.

According to Thagard, simplicity is a way of constraining consilience by ensuring that the theory is not ad hoc. This means that the theory explains more then just the data which it was introduced to explain, i.e., it is not fine tuned. Similarly, according to Popper, a theory that explains phenomena other than the specific phenomena it was designed to explain has richer content, and is therefore of greater value, than one which is less general.

We show that our formal model is able to encompass, besides Fauconnier and Turner's, other notions previously considered in the literature. In particular we will focus on how, with our formal model,

- we gain a a deeper understanding of the relationship between analogy-making and conceptual blending;

- we can show how conceptual blending relates to ontology alignment and ontology engineering;

- we can situate conceptual blending in the larger picture of a model of the creative process in general.

Analogy-Making. The amalgam-based model of conceptual blending put forward by [7] sheds light into the relationship of conceptual blending and generalisation-based models of analogy, such as the Heuristic-Driven Theory Projection (HDTP) framework of [47]. Amalgams were originally proposed in the field of Case-Based Reasoning (CBR) as a technique for the transfer and combination of knowledge from the retrieved case to the current problem. An amalgam can be conceived of as a generalisation of the notion of unification: if joining two descriptions leads to inconsistency, then unification is performed on a generalisation of the initial descriptions. Here, while in analogy-making generalisation is mostly modelled on the syntax level (two terms are generalised by identifying syntactically identical subterms and antiunifying these), amalgamation often also applies semantic generalisation returning a joint ontological super-concept for two input concepts (for instance generalising “convertible” and “station wagon” into “car”). A special case of particular interest is called an asymmetric amalgam, in which one input (called the source) is allowed to be generalised, while the other one (called the target) is not. Asymmetric amalgams can be seen as models of a form of analogical inference, transferring information form source to the target by creating a new amalgam that enriches the latter with the content of the source. It should be noted that analogy has long been seen as one of the thought-processes that underly mathematical creativity [57].

An *analogy-inspired view* of concept invention by concept blending will only need the mapping mechanism — based on the computation of the generalisation G (generic space) of input domains S and T that reflects common aspects of both spaces — and will replace the transfer phase by a new blending algorithm. While in analogy-making the analogical relations are used in the transfer phase to translate additional uncovered knowledge from the source S to the target space T, blending combines additional facts from one or both spaces. Therefore the process of blending can build on the generalisation and specialisations provided by the analogy engine, but has to include a new mechanism for transfer and concept combination. Here, amalgams naturally come into play: The set of specialisations can be inverted and applied to generalise the original source theory S, and the more general version can then be combined into an asymmetric amalgam with the target theory T, forming a (possibly underspecified) proto-blend of both. In a final step, the proto-blend is then completed into the blended theory B by applying corresponding specialisation steps stored from the generalisation process between S and T [3, 4].

Ontology Engineering. A representation-independent formalism such as the one put forward in our formal model of conceptual blending [7] allows for exploring the similarities apparent between the goals of heterogeneous ontology ‘alignment’ and the creative combination of thematically distinct information spaces. Ontological blending has been proposed as a new method for combining existing ontologies to create new ontologies on the basis of input ontologies whose domains are thematically distinct but whose specifications share structural or logical properties [28, 34].

We see the almost unlimited space of possibilities supported by ‘ontological blending’, offering substantial benefits not only for ontological engineering, but also for conceptual blending and related framework themselves. Re-considering some of the classic problems in conceptual blending in terms of ontological modelling and ontological blending opens up an exciting direction for future research, and we have already done some significant steps in this direction by exploring the representation of blends using the Distributed Ontology Language (DOL) [35] and its computational support via the Ontohub.org platform [33]. The relationship between conceptual blending, upward refinement operators and lightweight ontology representation languages have become evident in the proposal by [11] of a particular formal approach to resolve inconsistencies arising in the conceptual blending of $\mathscr{E}\mathscr{L}$++ concepts.

Creative Process. Conceptual blending, according to [18], underlies much of everyday thought and language, and is a fundamental cognitive principle in the creative process of concept invention. It is therefore reasonable to see how our formal model of conceptual blending relates to models of creativity in general. Since the model described in [7] is a very abstract model based on category theory, it is particularly suitable to relate to the model of the creative process as proposed by [40].

Mazzola et al. propose to take the insights offered by the Yoneda lemma of category theory as a metaphor for the process by which an open question may be solved in a creative way. [1] illustrate this for the particular case of Beethoven's piano sonata op. 109. [50] show by means of the Buddhist monk riddle [32] that Mazzola et al.'s metaphor for the creative process can be useful to make explicit the external structure of the concept or idea we want to creatively explore. This metaphor likens the creative process to the task of finding a canonical diagram that externalises the structure of a categorical object. In particular we have focussed on the image-schematic structure — in such a way that the solution to the riddle can be found by conceptual blending, using an amalgam-based process as the one put forward in our model.

2.3 To not assume what it sets out to explain

We want our model not to assume what it sets out to explain. This covers Sloman's fifth "non-circularity" criterion, and Popper's independently testable criterion.

According to Sloman, a good theory is non-circular when it does not assume that which it purports to explain. According to Popper, a theory that is independently testable is one that is not *ad hoc*, i.e. the theory cannot itself be evidence for the phenomena to be explained, or vice versa.

Conceptual blending has been thoroughly used as an analytic tool for understanding the origin of ideas and concepts *a posteriori* [55]. Consequently, the constitutive elements that conform the theory and that have guided our formal model are an abstraction covering the cases of blending studied in the literature. We have provided detailed formalisations of these cases in our model. Some examples are:

- the houseboat-boathouse blends [33]
- the Buddhist monk riddle [50]
- complex numbers [19]

Conceptual blending, however, has also been proposed as a basis for computational models of creativity, using the theory to guide the design and implementation of algorithms for generating novel ideas and concepts [56, 44, 23, 27, 17]. This is also the aim of our model: **to set out to explain how concepts are invented**. So, in order to *not* assume what we set out to explain, we have shown how our models has been capable to invent new concepts or conceptual structures that were not analysed as blends before. Most notably we have shown the conceptual blending structure underlying

- tritone substitutions in jazz music [15]
- prime ideals and Dedekind domains [25]
- novel harmonic spaces [9, 31]

3 The system's creativity: Jordanous's 14 components applied to COBBLE

3.1 Conjectures as intersections of models

Let $C_1,\dots,C_n$, and T_1 be mathematical concepts: for example, abelian groups, topological spaces, prime numbers, smooth functions and affine varieties. Let us assume that we need to solve a mathematical conjecture of the form: any mathematical object x belonging to the intersection of the models of the concepts $C_1,\dots,C_n$, is a model for the concept T_1.[2]

For instance, one can re-write Goldbach's Conjecture in the following way:

- Let $C_1(R)$ be the concept defining the 'even-ness' of an element a of a particular commutative ring R with unity, i.e., there exists an element $b \in R$ such that $a = b + b$.
- Let $T_1(S)$ be the concept defining elements c which can be expressed as the addition of two 'prime numbers' in a commutative ring S, where the notion of primeness in S is defined accordingly.[3]
- Then Goldbach's conjecture states that any natural number x belonging to the models of $C_1(\mathbb{Z})$, is also a model of $T_1(\mathbb{Z})$.

3.2 The mathematical research problem and how the COBBLE system could help

Suppose X is a researcher interested in solving a conjecture A that is expressed using the concepts $C_1,\dots,C_n$, and T_1, as before. Our system COBBLE should be considered as a creative support for X regarding the solution of A if it can give to X, in an interactive way, new relevant concepts $B_1,\dots,B_k$ which allow X to solve, at least some particular cases of A (or, respectively, $\neg A$, in which case COBBLE has helped to find a counterexample for A).

The concrete way for getting this kind of help from COBBLE is the following: X gives as input spaces a concept related with the hypothesis concepts $C_1,\dots,C_n$ and another concept related with the thesis concept T_1. So, COBBLE can compute a blend B that X should analyse in order to

[2] Many conjectures dealt with by a working mathematician have this 'model form'. There are also other kinds of 'meta conjectures' such as those involving the re-interpretation of mathematical phenomena from one area into another one that require additional formal tools.

[3] $p \in S$ is prime if for all $a, b \in S$, p divides $a * b$ implies that p divides a or p divides b.

see if it gives a new hint related with the solution of A. Theoretically, X can use B again as new input concept for obtaining another blending helping in the solution of A. So, this process will have a quite clear interactive nature.

3.3 Evaluating the current version of COBBLE

Now, let us assume that we want to grade our system from 1 (minimum) to 10 points (maximum) regarding each of 14 criteria of Anna Jordanous's Evaluation. Then, focusing on to the mathematical part of our system, we may form the following qualitative estimations:

- **Active Involvement and Persistence**: $\boxed{3}$ Our system has no a heuristic routines for producing suitable iterative blends directed towards an specific global goal.
- **Dealing with Uncertainly**: $\boxed{1}$ Our system is quite sensitive to small syntactic typos.
- **Domain Competence**: $\boxed{1}$ We have made no implementation concerning a domain-specific mathematical 'skill' or 'expertise'.
- **General Intellect**: $\boxed{1}$ The same as before for any kind of "Flexible and adaptable capacity" [29] of our system.
- **Generation of Results**: $\boxed{7}$ Partial versions of our system have produced some valuable and innovative mathematical notions – such as the notion of Containment-Division rings [6].
- **Independence and Freedom**: $\boxed{1}$ Our system is completely dependent of the input spaces.
- **Intention and Emotional Involvement**: $\boxed{2}$ Our system has just one concrete task to do: compute a colimit in HETS using a generic space coming from HDTP.
- **Originality**: $\boxed{6}$ The blended output spaces could have a mathematical value – as a novel contribution – in particular, in the case of input spaces with hundreds of axioms defining them.
- **Progression and Development**: $\boxed{1}$ COBBLE is a quite local oriented system, at least the mathematical part. Besides, the user is the only one responsible for the global development of the entire process.
- **Social Interaction and Communication**: $\boxed{5}$ As indicated before, one can conceive COBBLE as a system with a coherent interactive component, at least if the topic of the corresponding 'conversation' is the solution of a mathematical conjecture and/or the discovery of new mathematical notions.
- **Spontaneity/Subconscious Processing**: $\boxed{1}$ Every part of our implementations is 'quasi-conscious' in the sense that each step of the computation is quite well defined and there is no parallel processing at least from the mathematical point of view.
- **Thinking and Evaluation**: $\boxed{1}$ Our system does not choose on its own among any kind of options.

- **Value**: $\boxed{6}$ In the case of the Containment-division rings, our system has made valuable contributions in an specific domain.

- **Variety, Divergence and Experimentation**: $\boxed{3}$ COBBLE could increase this special component of creativity if we integrate to it sub-routines for finding maximal consistent and minimal inconsistent blends as in similar creative systems [39].

In conclusion, according to Jordanous's Criteria our system is just at its very beginning.

3.4 Potential improvements

Whether or not these are ultimately realised as blending operations, a key point for increasing COBBLE's supporting capacity during human creative mathematical research is the integration of computationally-feasible formalisations for familiar cognitive capacities, such as formal conceptual generalisation and particularisation, and metaphoric, analogical and inductive reasoning, among others.

It is worth mentioning that the creative support of our software would be bigger if we could allow X to put constraints and relevant information in COBBLE within the generic space in order to obtain a blending that satisfied more precisely the conditions required by X. This could be the topic of future research for improved versions of our system.

Another situation where COBBLE can be used is when X is studying how to transfer methods of an area F_1 of mathematical research (for example, number theory, algebra, homology theory) into a second area F_2 (e.g. geometry, topology). Here, COBBLE can be quite useful, since X can translate concepts coming from F_1 and F_2 into a many sorted first-order logic language and then form conceptual blendings with our system which could, in principle, be part of the desired new mathematical setting (e.g., some notions of homological algebra, algebraic geometry, algebraic topology, algebraic and analytic number theory).

4 The system's output: Jordanous/Ritchie applied to the harmonic blending system

The evaluation of the harmonic blending system needs to deal with multiple issues. One important objective is to assess the creativity of the system per se. However, since this tool is designed as a harmonisation assistant it is also important to complement the first aim with an assessment of the system's potential to enhance human creativity. Finally, rating the perceptual and cognitive effects of the blended harmonisations (for instance, testing whether they are perceived as blends or not) is of particular interest. To meet the above goals we have performed a number of empirical experiments. Since the melodic harmoniser is a computational tool that requires an external user, the evaluation mostly concerns the products of the system [30]. This is similar to the situation considered by [46], but for reasons described in the Section 7, we have not used his methods directly.

As a first step, we approached harmonic blending through one of the best-defined concepts in western harmony: the harmonic cadence. The cadence is defined as a harmonic progression (consisting of at least two chords) that serves as a phrase conclusion. Using two distinct cadences as input and through computational modelling of their characteristics, we produced a number of blended cadences. These cadences were subject to empirical evaluation in order to gain insight into their perceptual relationships. This was first addressed with a pairwise dissimilarity rating task, which is a common approach used in psychoacoustics in order to create perceptual spaces of the objects under study. Additionally, we acquired ratings concerning the qualitative characteristics of results, partially inspired by [30] and [46] (i.e., we focused on the research subject's perceptions of originality/expectancy and value) but also including some music-related characteristics (i.e, closure effect, degree of fit and tension). The results of this study are currently under review at the journal *Music Perception*.

We then proceeded to evaluate larger and more complex harmonic structures. For this purpose, we conducted an experiment whereby listeners were asked to perform an idiom classification of various melodies harmonised by blending between harmonic idioms. Thus, we wanted to test whether the blended idioms are perceived as such and to also assess the value that listeners attribute to the products of the system. One final point of interest was to examine whether the use of the harmonisation assistant can enhance the creativity of human users. Towards this end, we devised an experiment whereby music students were instructed to harmonise a traditional Scottish melody in the way they felt was most appropriate. After they had provided us with their initial harmonisations we presented them with a small number of alternative harmonisations produced by the blending system (although the participants were naive regarding their source) and gave them the opportunity to reconsider their initial approach once they had carefully listened to this input. The hypothesis here was that the revised approaches would be more diverse harmonically as a result of the influence by the computer-generated products. The last two experiments are currently in the phase of data analysis.

5 Progress towards a creative system: the FACE diagrammatic formalism applied to the trajectory of system development

Here we focus in one of the development trajectories taken within the COINVENT project. Referring to Figure 1, via the interface components Υ_1 and Υ_2, the user has access to specific *concepts* to blend, as well as a *generic space* that imposes constraints on the generated blend. These inputs are then processed by system in three phases: Φ_1, which generates potential blends, and Φ_2, which checks them for consistency and Φ_3, which applies some evaluation criteria.

The development of the system is illustrated in Figure 2, using the diagrammatic extension to the FACE model described in [10]. Here, capital letters F, A, C, or E are creative acts that generate a framing, aesthetic, concept, or example, respectively. Administrative acts S and T denote selection and translation. Lower-case letters denote the generated artefact in each case (e.g., the concept c corresponding to the concept-creation act C, or the aesthetic a corresponding to the aesthetic-creation act A). Subscripts p, g, or m indicate whether the act takes place at the process, ground, or meta level. Inside each box, stacks show the dependence in development epochs, and arrows show run-time message passing. Acts taken by the programmer or user are decorated with a bar, whereas acts taken by the system itself receive no extra decoration.

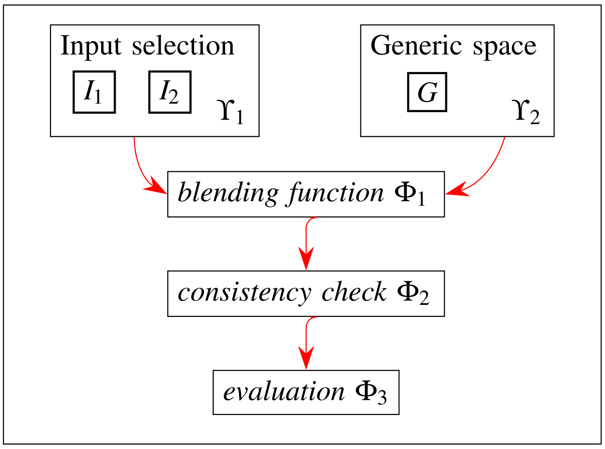

Figure 1: Overview of the blending system.

P1 In the initial prototype, $\overline{C_g^{I_1}}, \overline{C_g^{I_2}}$ and $\overline{C_g^G}$ denote the concepts supplied by the user. These are processed by the system with a previously programmed blending routine, $\overline{c_p^{\Phi_1}}$, to generate candidate blends, $\langle E_g^{\Phi_1} \rangle^*$. These are subsequently filtered according to the pre-specified aesthetic of consistency, denoted here by $S[\overline{a_g^{\mathfrak{C}}}]$; note that the computer carries out this check. The result is a list of consistent blends $\langle e_g^{\Phi_2} \rangle^*$, with further evaluation.

P2 In the second prototype, **P2**, the programmer enhanced the system with the capability of computing the generic space of the input concepts supplied by the user. This is done by a generic space routine, $\overline{c_p^{\Upsilon_2}}$, that generates a set of candidate generic spaces, $\langle E_g^{\Upsilon_2} \rangle^*$, for the input concepts $\overline{C_g^{I_1}}$ and $\overline{C_g^{I_2}}$. The generic space is computed by means of anti-unification in *amalgams* or *HDTP*.[4] The best generic space, $e_g^{\Upsilon_2}$, is selected according to some given heuristics, $S[\overline{a_g^{\mathscr{H}}}]$, and then is translated to the concept C_g^G in a format that module Φ_1 can process. The computation and elementary evaluation of conceptual blends is done as described in the prototype **P1**. (Note that in practice the computation and evaluation of conceptual blends is carried out for each generic space found in Υ_2 as part of its heuristic method.)

P3 In the third version of the prototype **P3**, the input concepts are not supplied by the user anymore, instead they are retrieved from a Rich Background, $\langle \overline{e^{\Upsilon_1}} \rangle^*$ according to some user requirements $\overline{a_g^{\mathscr{R}}}$.[5] This is implemented by a selection procedure in module Υ_1 that retrieves the best pairs of input concepts that meet the requirements of the user, $S[\overline{a_g^{\mathscr{R}}}](\langle \overline{e^{\Upsilon_1}} \rangle^*)$. The retrieved concepts are then passed to Υ_2. The computation then proceeds as in prototype **P2**.

[4] [52, 12] and [16] describe how the generic space is computed.

[5] [13] explains how discovery in the Rich Background can be implemented.

P4 In the fourth prototype, **P4**, after the consistency check, further evaluation is carried out to pick ‘good’ blends. This is done by means of a pre-defined aesthetic, denoted here by $S[\overline{a_g^{\mathscr{E}}}]$, that uses either value argumentation and audiences or notions of conceptual coherence to rank the blends.[6]

6 Relationship between System and Theory

The COBBLE system exploits HDTP, Amalgams and HETS in order to produce blends. In this section we give some examples of what is possible and go on to discuss limitations which render some examples impossible to compute with the current system. We discuss some of the techniques we employ to overcome these limitations.

6.1 Simple Example from Automated Theorem Proving

In Automated Theorem Proving, one of the most challenging problems to overcome is the use of creative lemmas to unblock a proof. The COBBLE system can be used to exploit known creative lemmas in a blended theory. For example let us consider a simple theory of the natural numbers as show in Figure 3. In this theory, a lemma (annotated by *(sl)* has been suggested by a human in order to prove a theorem (annotated by *st*). Figure 4 now shows a theory of lists which contains a theorem to prove (annotated by *st*), but no corresponding lemma.

In principle, one can define a *derived signature morphism* which identifies the element *s* from the natural theory, and the element *cons* from the list theory. When we run COBBLE on this example we get the generic space and morphisms defined by GENERALISATION0, MAPPING0_1 and MAPPING0_2 shown in Figure 5. As can be seen from the Generalisation, there is no element found which maps to both *s* and *cons*. In order for this to work we would need to be able to support lambda abstractions on morphisms, but this is not currently possible in COBBLE.

In order to overcome this problem in the current setting, we add “dummy” arguments to the function *s* so that the arities of *s* and *cons* are the same. This pre-processing generalisation step can be performed automatically by amalgams to the theory. One performed, the generic space contains an element to which both *s* and *cons* can map and the blend thus contains a candidate lemma in the blend theory.

6.2 Theoretical restrictions of COBBLE

COBBLE has been able to work through blending in mathematics whose mechanisation has helped to identify some novel and unexpected results Section 3 for some qualitative reflections on the user experience. The blending itself was essentially a one-step interaction, where human input was required to identify the input concepts. COBBLE itself can search the space of generated blends for optimal candidates, in the manner of **P4**, described in Section 5.

However, more ambitious aims would apply blending more extensively to the problem of computational discovery in mathematics. This might be done iteratively, by allowing search over

[6] See [13] and [49].

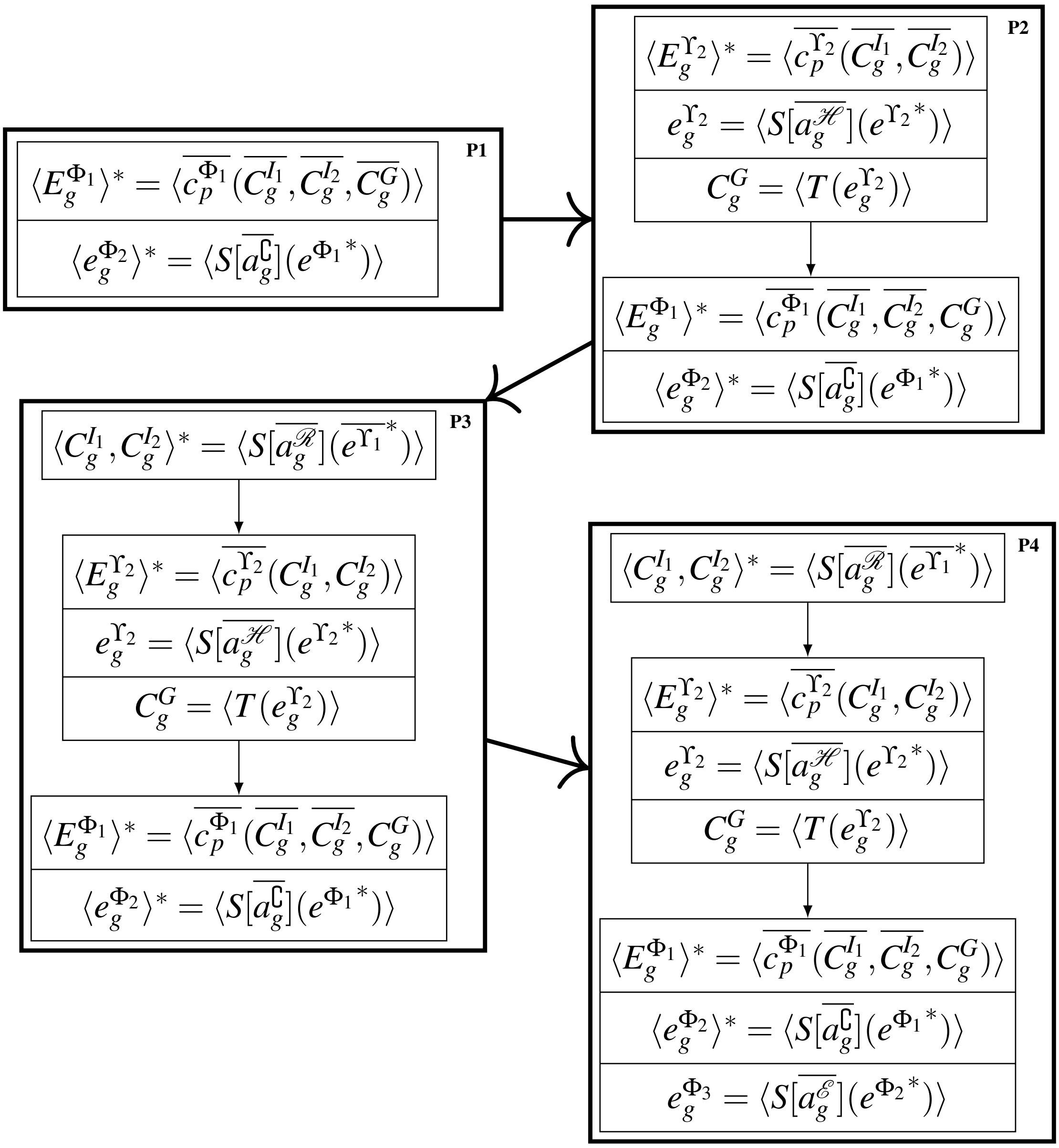
P1
$\langle E_g^{\Phi_1}\rangle^* = \langle\overline{c_p^{\Phi_1}}(\overline{C_g^{I_1}},\overline{C_g^{I_2}},\overline{C_g^{G}})\rangle$
$\langle e_g^{\Phi_2}\rangle^* = \langle S[\overline{a_g^{\complement}}](e^{\Phi_1 *})\rangle$
P2
$\langle E_g^{\Upsilon_2}\rangle^* = \langle\overline{c_p^{\Upsilon_2}}(\overline{C_g^{I_1}},\overline{C_g^{I_2}})\rangle$
$e_g^{\Upsilon_2} = \langle S[\overline{a_g^{\mathscr{H}}}](e^{\Upsilon_2 *})\rangle$
$C_g^G = \langle T(e_g^{\Upsilon_2})\rangle$
$\langle E_g^{\Phi_1}\rangle^* = \langle\overline{c_p^{\Phi_1}}(\overline{C_g^{I_1}},\overline{C_g^{I_2}},C_g^{G})\rangle$
$\langle e_g^{\Phi_2}\rangle^* = \langle S[\overline{a_g^{\complement}}](e^{\Phi_1 *})\rangle$
P3
$\langle C_g^{I_1},C_g^{I_2}\rangle^* = \langle S[\overline{a_g^{\mathscr{R}}}](\overline{e^{\Upsilon_1}}^*)\rangle$
$\langle E_g^{\Upsilon_2}\rangle^* = \langle\overline{c_p^{\Upsilon_2}}(C_g^{I_1},C_g^{I_2})\rangle$
$e_g^{\Upsilon_2} = \langle S[\overline{a_g^{\mathscr{H}}}](e^{\Upsilon_2 *})\rangle$
$C_g^G = \langle T(e_g^{\Upsilon_2})\rangle$
$\langle E_g^{\Phi_1}\rangle^* = \langle\overline{c_p^{\Phi_1}}(\overline{C_g^{I_1}},\overline{C_g^{I_2}},C_g^{G})\rangle$
$\langle e_g^{\Phi_2}\rangle^* = \langle S[\overline{a_g^{\complement}}](e^{\Phi_1 *})\rangle$
P4
$\langle C_g^{I_1},C_g^{I_2}\rangle^* = \langle S[\overline{a_g^{\mathscr{R}}}](\overline{e^{\Upsilon_1}}^*)\rangle$
$\langle E_g^{\Upsilon_2}\rangle^* = \langle\overline{c_p^{\Upsilon_2}}(C_g^{I_1},C_g^{I_2})\rangle$
$e_g^{\Upsilon_2} = \langle S[\overline{a_g^{\mathscr{H}}}](e^{\Upsilon_2 *})\rangle$
$C_g^G = \langle T(e_g^{\Upsilon_2})\rangle$
$\langle E_g^{\Phi_1}\rangle^* = \langle\overline{c_p^{\Phi_1}}(\overline{C_g^{I_1}},\overline{C_g^{I_2}},C_g^{G})\rangle$
$\langle e_g^{\Phi_2}\rangle^* = \langle S[\overline{a_g^{\complement}}](e^{\Phi_1 *})\rangle$
$e_g^{\Phi_3} = \langle S[\overline{a_g^{\mathscr{E}}}](e^{\Phi_2 *})\rangle$

Figure 2: The development of the blending system, showing the different prototype refinements from **P1** to **P4**.

multiple inputs, with optimality criteria determined mechanically.

Indeed, this notion of a process-with-memory suggests that it should be possible to use of blending to work out steps in a proof. The primary problem is: if blending is the realisation of "combinatorial creativity" [5], how can the process avoid getting swamped in the combinatorial explosion of possible things to combine? Here the local problem is one of fitting different mathematical components, or other structures, together in a sensible manner. The global problem is to selectively combine objects either to work towards some goal, or to build in a playful way and learn from the results. As indicated in the *s/cons* example described above, COBBLE has some limited ways to address the local problem. It does not, yet, tackle the global problem.

Nevertheless, some related work does move in this direction. [38] describes the use of CASL (via Hets) and TPTP a "proof checker," using the "auto-DG prover" that can discharge some proof obligations associated with the development graphs (DG) of Hets. Again, this focuses on goals that arise in that context, and is suitable for addressing simpler goals, but in general not strong, nor intended to be used to deal with proofs in general.

While COBBLE can support novel ideation – "*running the blend*" to elaborate or complete a mathematical theory formed by combinatorial means – its ability to reason about a broader context is limited, in part, by the lack of interfaces to the languages and repositories that express formal mathematical concepts (e.g., Mizar, HOL).

There are other practical limitations, which may be best described with an example. In [6], we describe a thought experiment that would realise the core notion in Galois theory, the "splitting field extension" as a blend. Recall that Galois theory develops a relationship between a polynomial $p(x)$ with coefficients in some field F and the extension K of F (written "K/F") containing all of the roots of $p(x)$ in the algebraic closure of F. The *Galois group* $\mathbf{Gal}(K)$ is the group of automorphisms of K/F that fix the elements of F. The fundamental theorem of Galois theory states that there is a bijection between the subfields of K/F and the subgroups of $\mathbf{Gal}(K)$; namely, subgroups correspond to fixed fields. A challenging but creative step is to discover the theorem that extending F *only with the roots of* $f(x)$ is a field.

In building field extensions, mathematicians need to consider "roots of the polynomial $p(x)$ in the field which is the algebraic closure of F." There is no obvious way to deal with the notion of "algebraic closure" in our framework. The blending machinery removes the distinction between polynomials with coefficients in F, K, F's algebraic closure, and K/F. It may be possible to address this limitation using parameterised types, but this remains to be seen.

Perhaps most problematically for the development of Galois theory within COBBLE, currently there is no way of computing colimits if automorphisms are characterised using higher-order logic – as would be quite natural. Although higher-order logic (with Henkin semantics) is available in HETS (and indeed in CASL) and the colimits are well-known to exist (because higher-order in this form is reducible to many-sorted first-order logic), the calculus of such colimits is not currently available in HETS.

spec NatSuc =

sort Nat

ops $0 : Nat;$

$s : Nat \to Nat$

op $sum : Nat \to Nat$

op $qsum : Nat \times Nat \to Nat$

op $plus : Nat \times Nat \to Nat$

$\bullet\ sum(0) = 0$

$\bullet\ \forall\, x : Nat \bullet sum(s(x)) = plus(s(x), sum(x))$ (sl)

$\bullet\ \forall\, x, y : Nat \bullet qsum(s(x), y) = qsum(x, plus(s(x), y))$

$\bullet\ \forall\, x : Nat \bullet qsum(0, x) = x$

$\bullet\ \forall\, x : Nat \bullet plus(0, x) = x$

$\bullet\ \forall\, x, y : Nat \bullet plus(s(x), y) = s(plus(x, y))$

$\bullet\ \forall\, x : Nat \bullet sum(x) = qsum(x, 0)$ (st)

$\bullet\ \forall\, x, y : Nat \bullet plus(sum(x), y) = qsum(x, y)$

end

Figure 3: A simple theory of the natural numbers

spec List =

sort El

sort L

op $nil : L$

op $cons : El \times L \to L$

op $app : L \times L \to L$

op $rev : L \to L$

op $qrev : L \times L \to L$

$\bullet\ \forall\, x : L \bullet app(nil, x) = x$

$\bullet\ \forall\, x, y : L; h : El \bullet app(cons(h, x), y) = cons(h, app(x, y))$

$\bullet\ rev(nil) = nil$

$\bullet\ \forall\, x : L; h : El \bullet rev(cons(h, x)) = app(rev(x), cons(h, nil))$

$\bullet\ \forall\, x : L \bullet qrev(nil, x) = x$

$\bullet\ \forall\, x, y : L; h : El \bullet qrev(cons(h, x), y) = qrev(x, cons(h, y))$

$\bullet\ \forall\, x : L \bullet rev(x) = qrev(x, nil)$ (tt)

end

Figure 4: A simple theory of lists

```
spec Generalisation0 =
      sort  Nat_L
      op    G_G394866 : Nat_L
      op    G_G394161 : Nat_L × Nat_L → Nat_L
      op    G_G394533 : Nat_L × Nat_L → Nat_L
      op    G_G394315 : Nat_L → Nat_L
      • G_G394315(G_G394866) = G_G394866
      ∀ G_G395416 : Nat_L
      • G_G394533(G_G394866, G_G395416) = G_G395416
      ∀ G_G395263 : Nat_L
      • G_G394161(G_G394866, G_G395263) = G_G395263
      ∀ G_G394729 : Nat_L
      • G_G394315(G_G394729)
        = G_G394533(G_G394729, G_G394866)
end


%Generalisation0 mapping NatSuc to List Cost = 82
view mapping0_1 :
      Generalisation0 to NatSuc =
      G_G394866 ↦ 0, G_G394161 ↦ plus, G_G394533 ↦ qsum,
      G_G394315 ↦ sum
end

view mapping0_2 :
      Generalisation0 to List =
      G_G394866 ↦ nil, G_G394161 ↦ app, G_G394533 ↦ qrev,
      G_G394315 ↦ rev
end
```

Figure 5: Generic space and morphisms calculated for list and nat by COBBLE

7 Metaevaluation

Our examination and use of the Jordanous and Ritchie models, and our new diagrammatic version of the FACE model, have afforded the opportunity to evaluate the practical applicability of these models. We assess how applicable each of the models are by logging the time spent and expertise required in using them. We further evaluate all of the creativity measures based on criteria from the philosophy of science on what constitutes a good theory. We have interviewed the members of the COINVENT consortium who made the most significant contributions to evaluation, in order to discern which methods: (i) are the easiest to apply; (ii) are the most useful in guiding progress; (iii) best capture intuitive notions of creativity. See Appendix B for our interview questions. We are using data from these initial interviews to devise a further round of questions for experts in CC evaluation.

7.1 Study design

We conducted four interviews, of approximately one hour in duration. These interviews were carried out face to face in a room with two interviewers and one interviewee, and were audio-recorded and then transcribed. Interviewers also took notes during the interview. Participants were recruited via personal connections and via an open call at a recent Computational Creativity (CC) conference. The main inclusion criterion were that participants had applied some evaluation metrics to software in CC. At the beginning of the interview participants were given an information sheet about the study, a consent form to sign, and were shown a copy of the interview guide so they could see the sorts of questions that they would be asked. All participants were offered a break partway through the interview.

The people who volunteered for the study were all male. They had advanced degrees, in computer science, artificial intelligence, electrical engineering and psychoacoustics, and mathematics, respectively. They had all worked in the field of CC for between 1.5 to 3 years.

7.2 Key findings

All participants were motivated by the idea of building systems that enhance human creativity and cognition, rather than building fully autonomous creative systems. In line with their level of expertise, participants suggested several domain-specific evaluation criteria such as logical consistency, consequence requirements, preference, mathematical validity, originality, closure effect, and tension. Participants varied as to whether these domain-specific criteria would be seen as measuring creativity, or related to but not specifically evaluating creativity.

Some of the participants felt that Jordanous and Ritchie required more understanding of computational creativity in general, and that the FACE model would be a better fit for their background in computer science. Others chose to apply the FACE model because they were system developers and the FACE model specified the steps that need to be implemented in the system. All who applied FACE had been involved in the system development. This evaluation was carried out at the end of the system development, with retrospective reflection on previous development cycles.

The FACE model was seen as rather complicated and, by some, as unintuitive. This applied in particular to the notation in the diagrammatic schemes: the concepts they represent were seen as necessary, but the language could be clarified. One participant suggested that a graphical notation might be more intuitive than the current algebraic notation. Participants took from two days to a week to apply this model. It was observed that it could have been useful to apply it earlier on, to early stages of the development. One further comment concerned large team projects and how the evaluation rubric should be designed to be useful to everyone on the team – including managers, domain experts, etc. – and not just the developers: "it might have been more interesting to have something that was a shared object that everyone in the project was contributing to, not just the software folks."

One participant liked the component-based approach developed by Jordanous because of its focus on the often intuitive themes related to creativity. Relatedly, in a separate user study carried out by one of our study participants, it was seen to be more productive to ask participants questions concerning their preferences, rather than whether they thought a particular artefact "was creative." In part, this was because of complex sociological factors in how the latter question is answered – for instance, in this study participants were students and there was reason to doubt that they would be comfortable judging the creativity of a professor or even one another, when a negative judgement could be taken as an insult. Less value-laden questions about personal preference, closure, etc., were seen as simpler. Additionally, questions concerning preferences were seen as being less varied and more meaningful and measurable. Problems related to ambiguity in translation of study texts were noted as a concern.

In this study, originality and preference were selected as evaluation dimensions from Jordanous's 14 components. While these were closely connected to Ritchie's criteria, Ritchie's approach was not applied. The main reason for this was that the underlying system only generated one artefact per run, and thus the notion of estimating proportions could not be applied in a straightforward manner. The fact that Ritchie's criteria only apply to a given set of artefacts was also seen as problematic for theoretical reasons. This participant commented that we don't use the idea of proportions to evaluate human creativity: for instance, we don't know how many draft versions of a given sonata Beethoven wrote before the final one, nor is this relevant for our evaluation of his creativity.

7.3 Practical applicability of the evaluation methods

In order to evaluate the practical applicability of the several evaluation methods, we carried out interviews with the people who had applied them (or, in one case, who chose to apply a nearby variant). Quotes in this section come from this round of interviews.

Jordanous's criteria were viewed as quite straightforward. Section 3 above took our respondent "one whole day" to complete. The respondent described the criteria as "complete," "sound," and even "in a sense also creative." The criteria were seen to be "centr[ed] on the human perspective" and the evaluator's "emotional experience." Unlike in certain of Jordanous's own applications of these criteria, the evaluator had a long experience with the system in several stages of development.

The diagrammatic FACE model was, by contrast, seen as relatively difficult to use, even with guidance. The developers were led through two in-person training workshops by one of the persons involved in developing the model. Drawing on feedback from other contributors to these workshops, completing Section 5 took one member of the developer team "a couple of days." About half of this time was spent working on paper and the other half of the time revising the handwritten draft in LaTeX. Another member of the developer team attended the second preparatory workshop and contributed to the discussion, but found the model “extremely complex” and felt it would not be “realistic for me to try to formalise everything that way.” This person, a computer scientist, further remarked: “I can imagine a software engineer, like a professional software engineer, having real troubles understanding the notations.”

Ritchie’s criteria were by and large not a particularly good fit for the kind of evaluation we aimed to carry out; accordingly, Section 4 does not apply the Ritchie criteria directly. The music domain was seen as the most likely match for Ritchie’s criteria within the COINVENT project, since this was the domain that generated the widest range of novel outputs. The reasons are as follow.

When applying the Ritchie model to a particular run or a particular session, the evaluator must define some specific system parameters, and then given all of its output, measure how much of it is novel, how much of it is valuable according to the judgement of experts. In our current harmonisation system, one set of parameters correspond to exactly one generated harmonisation.

The team behind the music system did, nevertheless, apply something similar to the Ritchie value/typicality judgements internally: after generating numerous examples, they would “present our [study] participants with some things that we think are valuable.” The reason for filtering the system’s output is that “you don’t [want to] waste people’s time.” In one case, the researchers did present participants with “a ‘wrong’ harmonisation that was completely bad” to test whether it would be perceived that way, which it was. Repeating this particular experiment was not seen as worthwhile. Nevertheless, some of the data gathered from study participants were quite close to those asked in Ritchie’s experiments, since participants were asked to rate the “originality” and “preference” of generated musical artefacts. Later, due to perceived confusion among respondents about how to interpret the term, “originality” dropped as a criterion and the study was repeated using the criterion of “expectancy” instead.

These criteria, selected from Jordanous’s 14, are quite close to Ritchie’s “typicality” and “value.” Respondents in the listening study were also asked about certain additional musical properties of the prompts, including “pairwise (dis)similarity,” “musical tense,” and “closure effect.”

Here it should be noted that the system itself does a round of internal evaluation (comparable to **P4**, discussed earlier), in which it considers several possible blends, from which “it will select one, trying to maximise the probability, the overall probability of this path.” Future work might “introduce some uncertainty” into this process and thereby generate a family of different outputs from one input, making Ritchie more directly applicable, but this has yet to been done.

Although the *originality/expectancy* and *value* data that has been gathered could be mapped to Ritchie’s framework to compute the various derived metrics he suggested (such as average quality, or proportion of atypical results that are good, etc.), this has not been carried out either. For one thing, it is not entirely clear that either “originality” or “expectancy” in the context of musical cadences would map cleanly to “typicality.” More fundamentally, the listening studies focused on understanding the small-scale correlates of different musical qualities, rather than one overall measure of “quality” or “creativity.” For instance, if a future system user “wanted to get an ending that had more tension than another” it would be important “to know how tension was perceived.”

7.4 Philosophical validity of the evaluation methods

Here we will walk through the major criteria discussed in Section 2 for each of the evaluation criteria discussed above, considered as a theory of creativity. Thus, for each of FACE, Jordanous, and Ritchie we ask to what extent it is "as general as possible," whether it "explains more than it set out to explain" and whether or not it "assumes what it sets out to explain."

To be as general as possible

Jordanous Jordanous's criteria appear to describe perceptions of creativity quite generally – certainly, the criteria could describe a much wider range of systems than COBBLE, as evidenced by the number of dimensions for which COBBLE received a rating of "1". In a sense it is too general for this system, since only some of the criteria seem to apply. From an epistemic point of view, we have seen that the methods can be applied both by someone who is well-acquainted with the system and (in Jordanous's own work) someone who is only familiar with a description of the system. It is not entirely clear how gracefully the ratings degrade when users have less experience with the system.

Ritchie Ritchie's criteria aim to be domain-general and as such do not intersect domain-specific, value-neutral, issues like the similarity of two chords. Ritchie's methods seem to apply exactly in the case in which the researcher wishes to understand a system's capability of generating items with a high overall quality, within some domain or genre, even though in his formalism the particular genre is abstracted away. As we saw above, there are several places where we have availed ourselves of Ritchie-like ratings, which shows an inherent degree of generality to this way of assessing things. The meta-criterion should perhaps be phrased "To be as general as possible—but no more." Our commentary indicates some of the boundaries of Ritchie's theory, but also possibilities for its extension.

FACE The diagrammatic version of the FACE model is intended to model progress towards a creative system in various development epochs. Again, there are emerging boundaries around the domain of application: for example, in the words of one of our interview subjects, there is a difference between "a good model for software" and "good evaluation criteria for creativity." The diagrammatic version of the FACE model somewhat uncomfortably straddles the two. Leaving aside the difficulties that its notational complexities pose there was a deeper question: namely, to what extent turning over responsibilities to the system really measures progress towards a *creative* system, or whether this instead models progress towards an *autonomous* system? [10] argue that more, more varied, and higher-level creative acts should all be seen as progress. The formalism is able to express all of these, but if it is to be truly general as a way of modelling progress towards creative systems, there may be other issues to capture as well. One which came up in our interviews was the effect of human factors and development arrangements amongst the developer team. [41] divide development of "social machines" into the *development* and *target* social machines. Although adding more notational complexity to FACE would make it harder to use, perhaps more attention could be given to the human factors issues.

Another related issue that came up in the interviews was the question of the way in which any of the evaluation criteria could be used in a more "formative" way, as a tool for communication among participants in the project team.

To explain more than it set out to explain

Jordanous Jordanous's collected criteria are based on her research into the way people talk about creativity. Thus, if some hypothetical system was rated highly on all of the criteria, it could be assumed that most people would agree that this was a "creative" system. However, what about systems – like COBBLE – that only rate highly on some subset of the criteria? In the assessment recorded in Section 3, COBBLE was rated above the "theoretical average" (namely, 5.5) only on three criteria: *Generation of Results*, *Originality*, and *Value*.[7] As a stand-alone "theory" the criteria do not explain what this means. Taken in context with the other theories under consideration, however, we can regard *achieving originality and value through the generation of results* as a kind of low-hanging fruit – precisely since, so to speak, these items form the "generic space" of each of the three models under consideration. Thus, although in outlining these criteria Jordanous did not set out to create a hierarchy of creative systems, it does such a hierarchy might be induced as a knock-on effect of empirical use.

Ritchie As indicated above, somewhat Ritchie-like ratings were used internally by the computational music research team, and also by the system itself. In the first case, items were filtered for quality, and in the second case, filtered (down to the last one standing) for probability of occurring within a corpus. Even though Ritchie assumes that the system is doing work that can be assessed within a genre, it should also be considered that applying Ritchie-like criteria (or indeed any coherent set of filtration criteria) also induces a genre. Again, although Ritchie does not set out to explain where genres come from, it may be that judgements of typicality, for instance, are even more fundamental than he says. Groups of items that are *atypical* within the broader population but mutually *typical* in the sense that they bear a family resemblance to one another may begin to express sensory "qualia."

FACE This model is primarily designed to be *descriptive*, so that the primary challenge in its application being to identify the examples, concepts, aesthetics, and framings that are present within a given system. However, in light of the various notions of progress (such as more, more varied, and higher-level creative acts), the theory also takes on a *prescriptive* flavour. In this way the diagrammatic FACE model can be used to envision further prototypes, such as a **P5** in which a generated "good" blend $e_g^{\Phi_3}$ feeds back into the Rich Background $\langle\overline{e^{\Upsilon_1}}\rangle^*$, expanding the range of examples available to choose from in further iterations, and stopping when some global criterion $\overline{a_g^{\mathscr{R}'}}$ is met. Indeed, each of the 14 criteria from Jordanous can be viewed as aesthetics which apply to the system as a whole (see Section 3, above). By viewing the course of system development as a generative process, FACE could incorporate any of these meta-level aesthetics – although it must be emphasised that at the moment this correspondence between the two theories exists only in principle.

To not assume what it sets out to explain

[7] The *empirical* average of the ratings provided is 2.86.

Jordanous Again, Jordanous's framework is based on research into the way people talk about creativity; it is not assumed that these criteria represent a definitive checklist for creative systems. The criteria may provide an easy way to guide discussions among developers and users, or other stakeholders, but they do not in themselves direct where the discussion should go, nor do they constrain the discussion to lie within the specific dimensions that have been put forward. Given the ratings in Section 3, it is interesting to ask how future enhancements might lead to improved ratings in some of the dimensions – but it is not assumed that these dimensions are specific "targets" for development.[8]

Ritchie Ritchie puts forth several assumptions, which, as indicated earlier, were not a perfect match for our studies: namely that the system produces numerous artefacts that fall within some genre, and that these can be evaluated with respect to their typicality and value. His paper explained how, from gathered ratings, numerical estimates for several derived metrics (including novelty) could be calculated. Ritchie refrains from setting thresholds for the production of valuable and novel artefacts that would qualify a given system as "creative." In our music listening studies, we took a similar tack, at an even further extreme: we did not even attempt to measure factors that would play into estimates of creativity, but rather sought to find parameters that may be unexpected or that lead to the perception of certain sensory qualities, like tension. Although the experiments measured "preference," we did not assume that this was an indicator of creativity. One of the studies specifically asked whether use of the system could enhance the users' creativity, which is rather different from the framework proposed by Ritchie. On a whole there are some overlaps with his method, but by working within a specific domain, our research was able to focus in on modelling specific experiential factors.

FACE The FACE model assumes that the salient features of software development in computational creativity can be described in terms of framings, aesthetics, concepts, and examples. Notably, it does not attempt a domain-general description of what these model-components actually are, leaving this to the interpretation of the researcher carrying the evaluation. As our discussion so far has shown, certain meta-level aesthetics are suggested as indicators of progress, but the evaluator is free to use others. Perhaps most fundamentally, the diagrammatic FACE model assumes that architecture and development work can be used to flesh out any particular notion of generativity or creativity, which we have certainly seen to be the case here. However, FACE does not assume any particular development trajectory or architecture, which instead follows the outlines and requirements indicated by the problem domain.

8 Conclusion

We have described the theoretical and practical aspects of the model of concept invention that was investigated in the COINVENT project, comparing it with predecessor systems. We outlined some novel contributions: to analogy making, ontology engineering, and modelling the creative process.

We then presented an evaluation of the COBBLE system, showing both the strengths and limitations. We described empirical research in listening studies based on the parallel cadence blend-

[8] "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" [26].

ing system, which resulted in new understanding of the ways to parameterise blends to achieve salient perceptual effects. We then described some of the technical advances made in the project – illustrating the make-up of several increasingly sophisticated prototypes in Figure 2. We described the alignment of the theory and the implemented system. Lastly, we presented a detailed meta-evaluation of the evaluation methods we applied, noting the relevance of both domain-specific and domain-general aspects of evaluation – as well as both product-focused and process-focused methods – in forming a full impression of progress in computational creativity.

These evaluations pointed to a number of ways in which the COINVENT prototypes could be improved in future work, notably: (1) more expressivity could be found by incorporating additional familiar aspects of the mathematician's toolkit; and (2) greater system autonomy could be obtained by developing additional iterative methods for generation-and-testing of system outputs in the various branches of the system. As indicated in our discussion of analogy-making in Section 2, blending can help understand and decompose some of these mechanisms.

We noted that our mechanisms for modelling progress towards a creative system could potentially be adapted to model within the research and development effort itself. We look forward to future work that connects the several strands of creativity evaluation research that have been applied and reviewed above to the study, use, and development of social machines.

References

[1] ANDREATTA, M., EHRESMANN, A., GUITART, R., AND MAZZOLA, G. Towards a categorical theory of creativity for music, discourse, and cognition. In *Mathematics and Computation in Music. 4th International Conference, MCM 2013. Montreal, QC, June 2013. Proceedings* (2013), J. Yust, J. Wild, and J. A. Burgoyne, Eds., vol. 7937 of *Lecture Notes in Artificial Intelligence*.

[2] ASTESIANO, E., BIDOIT, M., KIRCHNER, H., KRIEG-BRÜCKNER, B., MOSSES, P. D., SANNELLA, D., AND TARLECKI, A. CASL: the common algebraic specification language. *Theoretical Computer Science 286*, 2 (2002), 153–196.

[3] BESOLD, T. R., KÜHNBERGER, K.-U., AND PLAZA, E. Analogy, amalgams, and concept blending. In *Proceedings of the Third Annual Conference on Advances in Cognitive Systems (Poster Collection)* (2015), Cogsys.org.

[4] BESOLD, T. R., AND PLAZA, E. Generalize and blend: Concept blending based on generalization, analogy, and amalgams. In *Proceedings of the Sixth International Conference on Computational Creativity, Park City, Utah, USA, June 29 - July 2, 2015.* (2015), pp. 150–157.

[5] BODEN, M. A. Computer models of creativity. *AI Magazine 30*, 3 (2009), 23.

[6] BOU, F., CORNELI, J., GÓMEZ-RAMÍREZ, D., MACLEAN, E., SMAILL, A., AND PEASE, A. The role of blending in mathematical invention. In *Proceedings of the Sixth International Conference on Computational Creativity, ICCC 2015*, S. Colton, H. Toivonen, M. Cook, and D. Ventura, Eds. Association for Computational Creativity, 2015.

[7] BOU, F., EPPE, M., PLAZA, E., AND SCHORLEMMER, M. Reasoning with amalgams. Deliverable D2.1, FP7-ICT Collaborative Project COINVENT, October 2014.

[8] Bou, F., Schorlemmer, M., Corneli, J., Gómez-Ramírez, D., Maclean, E., Smaill, A., and Pease, A. The role of blending in mathematical invention. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)* (2015), S. Colton, H. Toivonen, M. Cook, and D. Ventura, Eds., pp. 55–62.

[9] Cambouropoulos, E., Kaliakatsos-Papakostas, M., and Tsourgas, C. Structural blending of harmonic spaces: a computational approach. In *Proceedings of the 9th Triennial Conference of the European Society for the Cognitive Science of Music (ESCOM)* (2015).

[10] Colton, S., Pease, A., Corneli, J., Cook, M., and Llano, T. Assessing progress in building autonomously creative systems. In *Proceedings of the Fifth International Conference on Computational Creativity* (2014), D. Ventura, S. Colton, N. Lavrac, and M. Cook, Eds.

[11] Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., naloza, R. P., and Plaza, E. Upward refinement operators for conceptual blending in the description logic el++. *Annals of Mathematics and Artificial Intelligence* (2016).

[12] Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., Peñaloza, R., and Plaza, E. Upward Refinement Operators for Conceptual Blending in $\mathscr{E}\mathscr{L}^{++}$. *Annals of Mathematics and Artificial Intelligence* (2016). To appear.

[13] Confalonieri, R., Plaza, E., and Schorlemmer, M. A Process Model for Concept Invention. In *Proceedings of the 7th International Conference on Computational Creativity, ICCC 16* (2016).

[14] Confalonieri, R., Plaza, E., and Schorlemmer, M. A process model for concept invention. In *The Seventh International Conference on Computational Creativity (ICCC 2016)* (2016).

[15] Eppe, M., Confalonieri, R., Maclean, E., Kaliakatsos, M., Cambouropoulos, E., Schorlemmer, M., and Kühnberger, K.-U. Computational invention of cadences and chord progressions by conceptual chord-blending. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015* (2015), Q. Yang and M. Wooldridge, Eds., AAAI Press, pp. 2445–2451.

[16] Eppe, M., Maclean, E., Confalonieri, R., Kutz, O., Schorlemmer, W. M., and Plaza, E. ASP, Amalgamation, and the Conceptual Blending Workflow. In *Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings* (2015), F. Calimeri, G. Ianni, and M. Truszczynski, Eds., pp. 309–316.

[17] Eppe, M., Maclean, E., Confalonieri, R., Schorlemmer, M., Kutz, O., and Plaza, E. Asp, amalgamation, and the conceptual blending workflow. In *Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings* (2015), F. Calimeri, G. Ianni, and M. Truszczynski, Eds., vol. 9345 of *Lecture Notes in Artificial Intelligence*, Springer, pp. 309–316.

[18] FAUCONNIER, G., AND TURNER, M. *The Way We Think*. Basic Books, 2002.

[19] FLEURIOT, J., MACLEAN, E., SMAILL, A., AND WINTERSTEIN, D. Reinventing the complex numbers. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence" 2014* (Osnabrück, 2014), T. R. Besold, K.-U. Kühnberger, M. Schorlemmer, and A. Smaill, Eds., vol. 01-2014 of *Publications of the Institute of Cognitive Science*, Institute of Cognitive Science.

[20] GOGUEN, J. An introduction to algebraic semiotics, with applications to user interface design. In *Computation for Metaphors, Analogy, and Agents*, C. L. Nehaniv, Ed., vol. 1562 of *Lecture Notes in Computer Science*. Springer, 1999, pp. 242–291.

[21] GOGUEN, J. Mathematical Models of Cognitive Space and Time. In *Reasoning and Cognition*, D. Andler, Y. Ogawa, M. Okada, and S. Watanabe, Eds., vol. 2 of *Interdisciplinary Conference Series on Reasoning Studies*. Keio University Press, 2006.

[22] GOGUEN, J. A., AND HARRELL, D. F. Foundations for active multimedia narrative: Semiotic spaces and structural blending. Unpublished manuscript available at `https://cseweb.ucsd.edu/~goguen/pps/narr.pdf`, 2005.

[23] GOGUEN, J. A., AND HARRELL, D. F. Style: A computational and conceptual blending-based approach. In *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, S. Argamon, K. Burns, and S. Dubnov, Eds. Springer, 2010, pp. 291–316.

[24] GOGUEN, J. A., LIN, K., AND ROŞU, G. Conditional circular coinductive rewriting with case analysis. In *Recent Trends in Algebraic Development Techniques, 16th International Workshop, WADT 2002, Frauenchiemsee, Germany, September 24-27, 2002, Revised Selected Papers* (2003), vol. 2755 of *Lecture Notes in Computer Science*, Springer, pp. 216–232.

[25] GOMEZ-RAMIREZ, D. Conceptual blending as a creative meta-generator of mathematical concepts: Prime ideals and dedekind domains as a blend. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence" 2015* (2015), T. R. Besold, K.-U. Kühnberger, M. Schorlemmer, and A. Smaill, Eds., vol. 02-2015 of *Publications of the Institute of Cognitive Science*, Institute of Cognitive Science.

[26] GOODHART, C. A. Problems of monetary management: the UK experience. In *Monetary Theory and Practice*. Springer, 1984, pp. 91–121.

[27] GUHE, M., PEASE, A., SMAILL, A., MARTÍNEZ, M., SCHMIDT, M., GUST, H., KÜHNBERGER, K.-U., AND KRUMNACK, U. A computational account of conceptual blending in basic mathematics. *Cognitive Systems Research 12*, 3–4 (2011), 249–265.

[28] HOIS, J., KUTZ, O., MOSSAKOWSKI, T., AND BATEMAN, J. A. Towards ontological blending. In *Artificial Intelligence: Methodology, Systems, and Applications, 14th International Conference, AIMSA 2010, Varna, Bulgaria, September 8-10. 2010. Proceedings* (2010), vol. 6304 of *Lecture Notes in Computer Science*, Springer, pp. 263–264.

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

[30] JORDANOUS, A. K. *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application.* PhD thesis, University of Sussex, 2013.

[31] KALIAKATSOS-PAPAKOSTAS, M., MAKRIS, D., TSOUGRAS, C., AND CAMBOUROPOULOS, E. Learning and creating novel harmonies in diverse musical idioms: An adaptive modular melodic harmonisation system. *Journal of Creative Music Systems 1*, 1 (2016).

[32] KOESTLER, A. *The Act of Creation.* Hutchinson & Co., 1964.

[33] KUTZ, O., BATEMAN, J., NEUHAUS, F., MOSSAKOWSKI, T., AND BHATT, M. E pluribus unum. formalisation, use-cases, and computational support for conceptual blending. In *Computational Creativity Research: Towards Creative Machines*, T. R. Besold, M. Schorlemmer, and A. Smaill, Eds., vol. 7 of *Atlantis Thinking Machines*. Atlantis Press, 2015, ch. 9, pp. 167–196.

[34] KUTZ, O., MOSSAKOWSKI, T., HOIS, J., AND BATEMAN, J. Ontological blending in DOL. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence"* (2012), T. R. Besold, K.-U. Kühnberger, M. Schorlemmer, and A. Smaill, Eds., vol. 1-2012 of *PICS Publications of the Institute of Cognitive Science*, Universität Osnabrück, pp. 33–40.

[35] KUTZ, O., NEUHAUS, F., MOSSAKOWSKI, T., AND CODESCU, M. Blending in the hub. In *Proceedings of the Fifth International Conference on Computational Creativity, Ljubljana, Slovenia, June 10-13, 2014.* (2014), pp. 297–305.

[36] LAKATOS, I. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, A. Musgrave, Ed. Cambridge University Press, 1970, pp. 91–195.

[37] LAKOFF, G., AND NÚÑEZ, R. E. *Where Mathematics Comes From.* Basic Books, 2000.

[38] LANGE, C., CAMINATI, M. B., KERBER, M., MOSSAKOWSKI, T., ROWAT, C., WENZEL, M., AND WINDSTEIGER, W. A qualitative comparison of the suitability of four theorem provers for basic auction theory. In *International Conference on Intelligent Computer Mathematics* (2013), Springer, pp. 200–215.

[39] MARTINEZ, M., ABDEL-FATTAH, A., KRUMNACK, U., GÓMEZ-RAMÍREZ, D., SMAILL, A., BESOLD, T., PEASE, A., SCHMIDT, M., GUHE, M., AND KÜHNBERGER, K.-U. Theory blending: extended algorithmic aspects and examples. *Annals of Mathematics and Artificial Intelligence* (2016), 1–25.

[40] MAZZOLA, G., PARK, J., AND THALMANN, F. *Musical Creativity.* Computational Music Science. Springer, 2011.

[41] MURRAY-RUST, D., AND ROBERTSON, D. Bootstrapping the next generation of social machines. In *Crowdsourcing: Cloud-Based Software Development*, W. Li, M. N. Huhns, W.-T. Tsai, and W. Wu, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 53–71.

[42] NEUHAUS, F., KUTZ, O., CODESCU, M., AND MOSSAKOWSKI, T. Fabricating monsters is hard. towards the automation of conceptual blending. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence" 2014*, T. R. Besold, K.-U. Kühnberger, M. Schorlemmer, and A. Smaill, Eds., vol. 01-2014 of *Publications of the Institute of Cognitive Science*. Institute of Cognitive Science, Osnabrück, 2014.

[43] ONTAÑON, S., AND PLAZA, E. Amalgams: A formal approach for combining multiple case solutions. In *CCBR'10: 18th International Conference on Case-Based Reasoning* (2010), vol. 6176 of *Lecture Notes in Artificial Intelligence*, Springer, pp. 257–271.

[44] PEREIRA, F. C. *Creativity and Artificial Intelligence: A Conceptual Blending Approach*, vol. 4 of *Applications of Cognitive Linguistics*. Mouton de Bruyter, 2007.

[45] POPPER, K. R. *Objective knowledge: An evolutionary approach*. Clarendon Press Oxford, 1972.

[46] RITCHIE, G. D. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines 17* (2007), 67–99.

[47] SCHMIDT, M., KRUMNACK, U., GUST, H., AND KÜHNBERGER, K. Heuristic-driven theory projection: An overview. In *Computational Approaches to Analogical Reasoning: Current Trends*, H. Prade and G. Richard, Eds., vol. 548 of *Studies in Computational Intelligence*. Springer, 2014, pp. 163–194.

[48] SCHMIDT, M., KRUMNACK, U., GUST, H., AND KÜHNBERGER, K.-U. Heuristic-Driven Theory Projection: An Overview. In *Computational Approaches to Analogical Reasoning: Current Trends*, H. Prade and G. Richard, Eds. Springer, 2014, pp. 163–194.

[49] SCHORLEMMER, M., CONFALONIERI, R., AND PLAZA, E. Coherent Conceptual Blending. In *Computational Creativity, Concept Invention, and General Intelligence: 5th International Workshop, C3GI @ ESSLLI 2016, Bozen-Bolzano, Italy, August 20/21, 2016* (2016), T. R. Besold, O. Kutz, and C. Leon, Eds.

[50] SCHORLEMMER, M., CONFALONIERI, R., AND PLAZA, E. The Yoneda path to the Buddhist monk blend. In *Proceedings of the Joint Ontology Workshops 2016 Episode 2: The French Summer of Ontology co-located with the 9th International Conference on Formal Ontology in Information Systems (FOIS 2016), Annecy, France, July 6-9, 2016.* (2016), vol. 1660, CEUR-WS.org.

[51] SCHORLEMMER, M., SMAILL, A., KÜHNBERGER, K.-U., KUTZ, O., COLTON, S., CAMBOUROPOULOS, E., AND PEASE, A. Coinvent: Towards a computational concept invention theory. In *5th Int. Conf. on Computational Creativity*. Ljubljana, Slovenia, 2014.

[52] SCHWERING, A., KRUMNACK, U., KÜHNBERGER, K.-U., AND GUST, H. Syntactic principles of heuristic-driven theory projection. *Cognitive Systems Research 10*, 3 (2009), 251–269.

[53] SLOMAN, A. *The computer revolution in philosophy: philosophy, science and models of mind*. The Harvester Press, Ltd., 1978.

[54] THAGARD, P. *Computational philosophy of science*. MIT press, 1993.

[55] TURNER, M. *The Origin of Ideas*. Oxford University Press, 2014.

[56] VEALE, T., AND O'DONOGHUE, D. Computation and blending. *Cognitive Linguistics 11*, 3/4 (2000), 253–281.

[57] WEIL, A. De la métaphysique aux mathématiques, 1966. An English language translation by Alan Smaill, entitled "André Weil on analogy in mathematics: translation and comments", is available at `http://dream.inf.ed.ac.uk/projects/coinvent/weil_translated.pdf`.

[58] ZACHARAKIS, A., KALIAKATSOS-PAPAKOSTAS, M., AND CAMBOUROPOULOS, E. Conceptual blending in music cadences: A formal model and subjective evaluation. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2015)* (2015).

A Evaluation Criteria for the Formal Model

A.1 Thagard's criteria

Thagard [54] suggests criteria for evaluating explanatory theories. These are intended for evaluating scientific, rather than philosophical theories, and have been extracted from studying examples of scientific theories. However, Thagard [54, p. 99] also claims that they can be used to determine the best explanation in metaphysical theories. The criteria are consilience, simplicity and analogy. Consilience is a measure of how many observables a theory explains, and the variety and importance of the facts explained. The notion of simplicity is a way of constraining consilience by ensuring that the theory is not *ad hoc*. This means that the theory explains more than just the data which it was introduced to explain, ie, it is not fine-tuned. Hence the first and second criteria need to be taken in conjunction with each other.

A.2 Sloman's criteria

Sloman [53, p. 50-53] sets out criteria which he claims can be used to judge 'theories which purport to explain possibilities', including scientific possibilities as well as theories in human sciences. We set out his criteria below. They are divided into a necessary criterion for something to be considered a theory at all, and a much longer set of criteria for something to be considered a *good* theory.

T is a theory if:

1. it explains a range of possibilities, ie, the possibilities are validly derivable from T , according to criteria for validity generated by the semantics of the language used for T.

T is a good theory if (1) is satisfied, and:

2. it is definite (there is a clear demarcation between what it does and what it does not explain);

3. it is general (it should explain many significantly different possibilities, preferably some which were not known about before the theory was invented. However it should not explain too many possibilities which have not been shown to exist, ie, it should not be *too* general);

4. it accounts for fine structure (the descriptions or representations of possibilities generated by T should be rich and detailed);

5. it is non-circular (it should not assume that which it purports to explain);

6. derivations from it are rigorous (it should be clear how the possibilities which T can generate are generated, or derived, from T);

7. it is plausible (the assumptions made in T should not contradict known *facts* —although they may contradict widely held *beliefs*);

8. it is economical (it does not include assumptions or concepts which are not necessary to explain the possibilities which it explains);

9. it is rich in heuristic power (the components of the theory, ie, the assumptions, concepts, representation language and way in which possibilities are generated, should be such that the detection of errors and gaps, design of problem solving strategies etc is easily manageable); or

10. it is extendable (it should be possible to embed the theory in an improved enlarged theory which explains further possibilities or has a higher degree of fine structure).

These criteria provide us with ways by which we can compare two philosophical theories. Sloman also considers the further criteria that a theory enables us to control or predict phenomena. He argues that these criteria are often over-emphasised, for instance the theory of evolution is arguably one of the most important scientific theories, but its power lies mainly in explaining possibilities, rather than controlling or predicting biological developments.

A.3 Popper's criteria

Following Tarski, Popper [45] suggests that we divide the universal class of all statements into true and false, T and F . He claims that the aim of science is to discover theories (explanations) whose content covers as much of T and as little of F as possible, where the content of a theory is the set of all statements logically entailed by it. This set may also be divided into true and false statements (the theory's truth and falsity content). A good theory should suggest where to look, ie, new observations which we had not thought of making before.

This is comparable to the situation described in [46], in which we divide the universal class of all basic items in a domain into good and bad, V and V'. If we describe the content of a program as its output set O which may be divided into good and bad artefacts, then we can claim that one aim within AI is for a program to generate as much of V (and as little of V') as possible. A good system should suggest new areas of the search space, ie, find artefacts which we had not thought of generating before. If we accept this analogy then Popper's criteria for evaluating theories sheds light on our criteria for evaluating programs.

Popper sets out two criteria for a satisfactory theory (in addition to it logically entailing what it explains). Firstly it must not be *ad hoc*. By this is meant that the theory (explicans) cannot itself be evidence for the phenomena to be explained (explicandum), or vice versa. For example if the explicandum is 'this rat is dead', then it is not enough to suggest that 'this rat ate poison' if the evidence for it having done so is that it is now dead. There must be independent evidence, such as 'the rat's stomach contains rat poison'. The opposite of an *ad hoc* explanation then, is one which is independently testable. Secondly, a theory must be rich in content. For example a theory which explains phenomena other than the specific phenomena it was designed to explain has a much richer content, and is therefore of greater value, than one which is less general (the principle of universality).

Applying these criteria to our programs, if we see a program (K) as the theory and the set of artefacts we wish to generate (I) as the phenomena to be explained, then we are interested in the independent testability of K and the richness of its content. A program which has been carefully tailored in order to produce very specific artefacts cannot be claimed to be a good program on the grounds that it produces those artefacts. There must be independent grounds for its value, such as also generating other valuable artefacts. Within the programming analogy, this is clearly connected to the richness of content criterion; the more valuable artefacts outside of I and fewer worthless artefacts a program generates, the better that program is.

A.4 Lakatos's criteria

Lakatos [36] attempted to salvage some of Popper's falsification methodology, which had suffered in the light of Kuhn's analysis of paradigm change. However, he thought that Popper's focus on the relationship between theory and observation in his falsification methodology was too simplistic, arguing instead that a methodology must take into account the *structure* of a theory. In his account he developed the notion of a scientific research programme, which comprises a *hard core* and a *protective belt*. The hard core consists of the defining characteristics of a programme: these are very general theoretical hypotheses which form the basis of the programme. If a scientist were to reject or modify these hypotheses, then essentially he or she would be abandoning the research programme (this is similar to Kuhn's notion of a paradigm shift). The protective belt consists of explicit auxiliary hypotheses and assumptions which are less central to a research programme; these could be rejected or modified without serious repercussions to the research programme. If a hypothesis from the hard core appears to have been falsified, then in order to remain in the same research programme, appropriate changes or additions would be made to the protective belt, rather than to the hypothesis directly. The *positive heuristic* of a research programme indicates how the protective belt can be altered in order to protect and extend the predictive and explanatory power of the hard core. The negative heuristic states that the hard core must remain unchanged. Another constraint is that modifications made to hypotheses in the protective belt must be independently testable.

Lakatos uses these ideas to show how we can evaluate work done within a research programme, and to evaluate competing research programmes. He also used them as demarcation criteria between science and non-science. Research programmes can be evaluated according to whether they are *progressive* or *degenerative*. A programme is progressive if it satisfies two criteria: firstly if it comprises a coherent hardcore which involves a definite mapping out of predictions and future research, and secondly if it, at least occasionally, leads to the discovery of novel phenomena. These also serve as demarcation criteria. A programme is degenerative if it is gradually coming undone, and there are no recent novel predictions to its name. It is difficult to evaluate a research programme except in retrospect, as we can never be sure that a new discovery is not just around the corner.

One way of evaluating two rival theories is to look at Lakatos's falsification criteria [36, p. 116]. A theory T is falsified if and only if another theory T' has been proposed with the following three criteria:

1. T' predicts novel facts, ie, phenomena which was not predicted by T (this is a sign of a theoretically progressive research programme);
2. T' explains all of the confirming instances the T explained; and
3. some of the excess content of T' is corroborated (this is a sign of an empiricallyprogressive research programme).

Note that T and T' may share the same hard core.

Lakatos differed from Sloman, Popper and Thagard in that the criteria he identified as a good scientific theory were intended to be used for demarcation criteria, ie, they could help to distinguish science from non-science. Conversely, while Sloman, Popper and Thagard set out criteria of a good scientific theory, they also commented that the same criteria could be used to evaluate non-scientific theories.

B Evaluating criteria of progress in Computational Creativity: Interview Guide

Part I: Putting the researcher into context (Warm up)

1. Could you tell us about your academic background and current research?
2. How long have you been working in Computational Creativity (CC)?
3. What are your areas of specialism within CC?
4. Are you familiar with any evaluation criteria (EC) in CC? If so, which ones?
5. Have you applied any of the criteria in the context of software design, development or evaluation? If so, proceed to Part II. If not – Have you ever considered doing so? What were your reasons for not applying them? Now go directly to Part III.

Part II: Application of the criteria (in depth)

Fact gathering

1. Evaluation criteria:
   (a) Which criteria have you applied?
   (b) What were your reasons for selecting those particular criteria?

   (c) Did you apply more than one set of criteria? If so, why?

2. Software:
   (a) Who developed the software to which you applied the EC?
3. Researcher to apply them:
   (a) Who applied the EC?
   (b) When did you apply the EC? (at what stage in the development cycle)

Ease of application

1. How long did it take to apply the EC?
2. How much expertise did it take to apply them?
3. How did you apply them?
4. How many people were involved in applying them? Did you measure agreement? If so, what was the result?
5. Were there any other practical considerations?
6. If you applied multiple EC, did you find any easier to apply than the others?

Usefulness of application

1. Did the results tell you anything?
2. Have you used the results so far? Do you expect to?
3. If you applied multiple EC, did you find any more useful than the others?

Faithfulness to notion of creativity

1. What did those criteria measure?
2. If you applied multiple EC, did you find that any captured your notion of creativity more than the others?

Is there anything else that you'd like to say about your experiences applying the criteria?

Part III: Evaluation of the criteria (reflections)

1. What properties should evaluation criteria for CC have?
2. Do any of the current evaluation criteria have those properties?
3. Are any of the current ones adequate?
4. Do you have thoughts on improving current criteria or developing a new set?