Towards an Algebraic Approach to Theory and Concept Evaluation

Pietro Galliani

Free University of Bozen-Bolzano
pgallian@gmail.com

Abstract. I introduce a theoretical framework for reasoning about the values of theories (understood as sets of elements of random conceptual algebras) and about the values of concepts with respect to theories. Then I define *theory formation games* and argue that they provide an adequate theoretical framework for the evaluation of strategies and algorithms for computational creativity; and, finally, I briefly discuss two possible practical applications of this framework, namely the automatic search for Natural Deduction rule systems for logics and the construction of falling-rule-list-plus-definitions classification models.

1 Introduction

One problem common to computational models of creativity, in mathematics or in other areas, is that it is often less than clear how such models, or their products, are to be evaluated [17, 19, 14, 18, 10]. There exist models of concept formation that have been shown to be, in principle, capable of reproducing non-trivial conceptual leaps, of the kind that – if produced spontaneously by a human being – would surely be regarded as creative: for instance, in [3] it is shown how the notion of *prime ideal* can arise from the blending of two conceptual spaces representing *commutative rings with unity* and *integer numbers* respectively. However, the same systems are equally able to generate, by the same techniques, concepts that a human would instantly recognize as spurious and of little value. It is not currently clear which criteria humans use for such an evaluation: while some conditions that, let us say, a *conceptual blending* operation [9] should satisfy in order for its result to be meaningful have been proposed (e.g. [7, 6]), it is fair to say that, at the moment, models of creativity and concept formation are relatively well suited for the formal *explanation* of creative concept formation but not as useable for tools for the assisted (let alone autonomous) creation of novel and useful concepts in mathematics or in other disciplines.

In this work, I present a simple, abstract, algebraic formalization of the notion of the *value* of a theory $\mathbf{T}$ (or of a concept $\mathtt{c}$ given the theory $\mathbf{T}$). This formalization does not involve the modelling of researchers as individual agents with utilities and possible actions, as it is often done in the area of social simulation (see e.g. [1, 11, 21]): while such approaches can often be the source of very valuable insights, the high number of possible choices involved in their specification poses serious difficulties to the analysis of their implications. Finally,

I briefly discuss two potential applications of this framework to the problems of building Natural Deduction systems and of improving the interpretability of *falling rule list classification models* [2].

2 Basic definitions

Concepts, mathematical or non-mathematical, do not live in isolation; and, to a large degree, creative activity can be understood as the generation of novel concepts through the *combination* of known ones. However, as remarked in the introduction, not all such combinations are feasible, let alone fruitful. This justifies the following definition:

Definition 1 (Conceptual Algebra). *A* conceptual algebra $\mathcal{A}$ *is a pair* $(\mathbf{A}, \cdot)$*, where*

- $\mathbf{A}$ *is a set of* concepts*;*
- $\cdot$ *is a partial operation over* $\mathbf{A}$*.*[1]

In the above definition, the algebra $\mathbf{A}$ is not required to be commutative. In this way, we can distinguish between two possibly distinct roles (active/passive) of concepts in a concept combination: a concept can either be applied *to* another concept to modify it, or be modified by the application of another concept. To make an example, if we take the concept of "prime ideal" as a combination of the concepts "prime number" and "ideal", it is clear that it is the notion of primality that is being applied to the notion of ideal and not vice versa; and indeed, the resulting concept belongs to the conceptual domain of ideals rather than to the one of numbers. Quite arbitrarily, in a concept combination $\mathtt{a} \cdot \mathtt{b}$ the left operand $\mathtt{a}$ is assumed to take the active role and the right one $\mathtt{b}$ is assumed to take the passive one.

A theory $\mathbf{T}$ is then defined as a set of concepts in a conceptual algebra. It is important to emphasize here, for the following analysis to be meaningful, that a theory is not understood as a set of *axioms*: instead, it is taken to be the set of all *concepts* - be them statements, rules, heuristics, or proofs - that are part of the "received knowledge" about a subject. For instance, the Riesz Representation Theorem is certainly part of the theory of modern functional analysis, even though it cannot meaningfully be considered one of its axioms.[2]

Definition 2 (Theory). *A theory over a conceptual algebra* $\mathcal{A} = (\mathbf{A}, \cdot)$ *is a finite subset* $\mathbf{T} \subseteq \mathbf{A}$*.*

[1] That is, it is a subset $\cdot^{\mathcal{A}}$ of $\mathbf{A} \times \mathbf{A} \times \mathbf{A}$ such that $(\mathtt{a}, \mathtt{b}, \mathtt{c}), (\mathtt{a}, \mathtt{b}, \mathtt{c}') \in \cdot^{\mathcal{A}}$ implies $\mathtt{c} = \mathtt{c}'$. As usual, we write $\mathtt{a} \cdot \mathtt{b} = \mathtt{c}$ for $(\mathtt{a}, \mathtt{b}, \mathtt{c}) \in \cdot^{\mathcal{A}}$.

[2] In fact, it is questionable whether a notion such as "the axioms of functional analysis" has a meaningful and unique denotation to begin with: axiomatic systems can be very useful tools to *systematize* mathematical knowledge, but they are not necessarily epistemically prior to it.

How valuable would such a theory be? In our framework, we do not have any access whatsoever to properties of individual concepts such as their "simplicity" or "elegance": "internal" properties affecting the values of individual concepts could be added to our framework easily enough, but for now it is preferable to avoid such complications and treat concepts as individually indistinguishable. The only difference between concepts in this framework, thus, lies in which further concepts can be obtained (or "derived") from them through applications of the concept combination operator $\cdot$.

The following analysis of theory value is based on three **guiding principles:**

Guiding Principle 1: Storing and using theories involves costs (e.g. paper, disk space, effort required for researchers to achieve proficiency with them...). These costs grow approximately linearly with the number of concepts contained in the theory.

Guiding Principle 2: The more concepts can be derived from a theory, the more valuable it is.

Guiding Principle 3: Concepts that can be derived from a theory in few steps contribute more to its value than concepts that can be derived from it only in many steps.

In accordance with the above principles, I now propose a definition for the value $v(\mathbf{T})$ of a theory $\mathbf{T}$ in a conceptual algebra $\mathcal{A}$. First, however, I need a couple of auxiliary concepts:

Definition 3 (n-th level span and edge of a theory). *Let $\mathcal{A} = (\mathbf{A}, \cdot)$ be a conceptual space and let $\mathbf{T} \subseteq \mathbf{A}$ be a theory in it. Then for all $n \in \mathbb{N}$, its n-th level* span $\langle \mathbf{T} \rangle_n$ *and its n-th level* edge $[\mathbf{T}]_n$ *are defined recursively as follows:*

- $\langle \mathbf{T} \rangle_0 = \mathbf{T}$;
- *For all* $n \in \mathbb{N}$, $[\mathbf{T}]_n = \{\mathtt{c} \in \mathbf{A} \backslash \langle \mathbf{T} \rangle_n : \exists \mathtt{a} \in \langle \mathbf{T} \rangle_i, \mathtt{b} \in \langle \mathbf{T} \rangle_j$ *s.t.* $\mathtt{a} \cdot \mathtt{b} = \mathtt{c}$ *and* $i + j = n\}$;
- *For all* $n \in \mathbb{N}$, $\langle \mathbf{T} \rangle_{n+1} = \langle \mathbf{T} \rangle_n \cup [\mathbf{T}]_n$.

In other words, $\langle \mathbf{T} \rangle_n$ is the set of all concepts that can be derived from those in $\mathbf{T}$ within n concept combination steps,[3], while $[\mathbf{T}]_n$ is the set of all those that cannot, but could be derived in one extra step.

As per **Guiding Principle 2**, all concepts in $\langle \mathbf{T} \rangle = \mathbf{T} \cup \bigcup_{k=0}^{\infty} [\mathbf{T}]_k$ contribute to the value $v(\mathbf{T})$ of $\mathbf{T}$. Furthermore, as per **Guiding Principle 3**, the concepts in $\mathbf{T}$ —which are already in the theory as given— contribute individually more to $v(\mathbf{T})$ than those in $[\mathbf{T}]_0$, which require one concept formation step to be obtained; and, in general, the individual contributions of concepts in $[\mathbf{T}]_i$ are greater than those of concepts in $[\mathbf{T}]_j$ whenever $i < j$. As no further information regarding

[3] It may be worth pointing out, once more, that these are **not** meant to be deduction steps in a specific formal system. The concepts contained in $\mathbf{T}$ may correspond to statements; but they may also correspond to definitions, heuristics or so on.

the contributions of concepts to the value of theory is available, a conservative choice is to let —by choice of unit— the concepts in $\mathbf{T}$ individually provide a unit contribution to the value of the theory,[4] and to let the concepts in $[\mathbf{T}]_i$ contribute for $v_i \in [0,1]$ each, where $1 > v_0 > v_1 \ldots > v_n \ldots$ Additionally, it is reasonable to require v_n to tend to zero for $n \to \infty$: indeed, the contribution of a concept to the value of theory becomes negligible as the length of the derivation necessary to obtain it from concepts in the theory grows extremely large. Thus, and by analogy with the notion of discounted reward in reinforcement learning [22], a simple (and by no means unique, but reasonable enough until and unless evidence to the contrary is found) choice would be to let $v_k = \lambda^{k+1}$ for some fixed discount rate $\lambda \in [0,1]$.

Finally, as per **Guiding Principle 1**, working with a theory $\mathbf{T}$ carries costs which grow linearly with its size $|\mathbf{T}|$: thus, $v(\mathbf{T})$ also contains a term $-C|\mathbf{T}|$, where $C \in \mathbb{R}_{\geq 0}$.

We thus obtained the following definition, which is the central definition of this work:

Definition 4 (Value of a theory in a conceptual algebra). *Let $C \in \mathbb{R}_{\geq 0}$ be the* cost per concept *and let $\lambda \in [0,1]$ be the* concept derivation discount rate*. Then for every conceptual algebra $\mathcal{A} = (\mathbf{A}, \cdot)$ and every theory $\mathbf{T} \subseteq \mathbf{A}$, the value $v_{C,\lambda}(\mathbf{T})$ is given by*

$$v_{C,\lambda}(\mathbf{T}) = (1 - C)|\mathbf{T}| + \sum_{k=0}^{\infty} \lambda^{k+1} |[\mathbf{T}]_n|.$$

The value of a concept given a theory can then be defined simply as the effect that adding (or removing) it would have on the value of a theory. More precisely:

Definition 5 (Value of a concept given a theory). *Let $\mathbf{T}$ be a theory in a conceptual algebra $\mathcal{A} = (\mathbf{A}, \cdot)$, let $\mathsf{c} \in \mathbf{A}$ be a concept of $\mathcal{A}$, let $C \in \mathbb{R}_{\geq 0}$ and let $\lambda \in [0,1]$. Then*

$$v_{C,\lambda}(\mathsf{c}|\mathbf{T}) = v_{C,\lambda}(\mathbf{T} \cup \{\mathsf{c}\}) - v_{C,\lambda}(\mathbf{T} \backslash \{\mathsf{c}\}).$$

Note that this definition makes sense regardless of whether the concept c is or is not part of $\mathbf{T}$ already; and that, moreover, it follows easily from it that if $\mathbf{T} = \{\mathsf{c}_1 \ldots \mathsf{c}_n\}$ then $v_{C,\lambda}(\mathbf{T}) = \sum_{t=1}^{n} v_{C,\lambda}(\mathsf{c}_t | \{\mathsf{c}_i : i < t\})$.

3 Towards a Formal Model of Creativity

The simple model presented above can be used as a component of a formal, abstract framework through which to evaluate the *creativity* of an algorithm.

[4] In other words, the unit of value corresponds to the amount of value that a single concept would contribute to the theory if added directly to it, *before considering costs and possible derivations.*

The idea, in brief, is that a procedure is creative insomuch as it can generate valuable theories in an unknown (but – at a cost – explorable) conceptual algebra. In what follows, I present a decision-theoretical formalization of this intuition.

Definition 6 (Theory Formation Games). *Let $\mathcal{A} = (\mathbf{A}, \cdot)$ be a conceptual algebra. Furthermore, let $\beta_0 \in R_{\geq 0}$ be a* budget*, let $\rho, \eta \in \mathbb{R}_{\geq 0}$ be the costs associated to* combination *and* random exploration *respectively, and – as usual – let C and λ be the memorization cost and the concept derivation discount rate respectively. Then a theory formation game position is a tuple $p = (\mathbf{K}, \mathcal{F}, \beta)$, where*

1. *$\mathbf{K} \subseteq \mathbf{A}$ is a set of* known concepts*;*
2. *$\mathcal{F}$ is a set of* known facts*, of the form $\mathtt{a} \cdot \mathtt{b} = \mathtt{c}$ or of the form $\mathtt{a} \cdot \mathtt{b} = \downarrow$ for $\mathtt{a}$, $\mathtt{b}$ and (if applicable) $\mathtt{c}$ in $\mathbf{K}$;*
3. *$\beta \in \mathbb{R}_{\geq 0}$ is the* remaining budget*.*

The initial position *of the game is $(\emptyset, \emptyset, \beta_0)$,*[5] *and the possible moves given a given position $(\mathbf{K}, \mathcal{F}, \beta)$ are defined as follows:*

- *If $\beta \geq \rho$ then, for all $\mathtt{a}, \mathtt{b} \in \mathbf{K}$, $\mathtt{a} \cdot \mathtt{b} = \mathtt{?}$ is a possible move. If $\mathtt{a} \cdot \mathtt{b}$ is defined in $\mathcal{A}$ and is some $\mathtt{c} \in \mathbf{A}$, then the successor position is $(\mathbf{K} \cup \{\mathtt{c}\}, \mathcal{F} \cup \{\mathtt{a} \cdot \mathtt{b} = \mathtt{c}\}, \beta - \rho)$; otherwise, it is $(\mathbf{K}, \mathcal{F} \cup \{\mathtt{a} \cdot \mathtt{b} = \downarrow\}, \beta - \rho)$.*
- *If $\beta \geq \eta$ then $\mathtt{??}$ is a possible move. The successor positions are then of the form $(\mathbf{K} \cup \{\mathtt{c}\}, \mathcal{F}, \beta - \eta)$, where the concept $\mathtt{c}$ is an arbitrary concept in $\mathbf{A}$.*
- *For all $\mathbf{T} \subseteq \mathbf{K}$, $\mathtt{Choose}(\mathbf{T})$ is always a possible move which ends the game. $\mathbf{T}$ is then the* output theory*.*

A strategy *for such a game is a function σ from positions $(\mathbf{K}, \mathcal{F}, \beta)$ to legal moves. A* complete play *of such a strategy is then a sequence $p_0 p_1 \dots p_n$, where p_0 is the starting position $(\emptyset, \emptyset, \beta)$, for each $i = 1 \dots n$ it is the case that p_{i+1} is a possible successor of position p_i under the move $\sigma(p_i)$, and $\sigma(p_n)$ is of the form $\mathtt{Choose}(\mathbf{T})$. The* value *of such a play is then simply the value $v_{C,\lambda}(\mathbf{T})$ of the output theory $\mathbf{T}$.*[6]

The value of a strategy σ is then defined simply as the expectation of the value of its complete plays, when the elements $\mathbf{c}$ added to $\mathbf{K}$ during $\mathtt{??}$ moves are chosen randomly from the uniform distribution[7] *over the set $\mathbf{A}$ of all possible concepts.*

---

[5] Of course, this can be changed if one wishes to consider games starting from nonempty sets of known concepts and facts.

[6] It is easy to see that no infinite-length plays are possible in theory formation games, as $\mathtt{a} \cdot \mathtt{b} = \mathtt{?}$ and $\mathtt{??}$ moves decrease the available budget β until only play-ending moves $\mathtt{Choose}(\mathbf{T})$ remain available.

[7] Or from another suitable distribution. Note that the selected $\mathtt{c}$ may be in $\mathbf{K}$ already: this reflects the fact that, if most of the concepts relevant to a certain subject are known already, it is rarely productive to go fishing for the others in a random, undirected way.

In the above definition, theory formation games are specified with respect to a *fixed* conceptual algebra $\mathcal{A}$. However, it is not usually the case that the structure of a conceptual algebra is known in advance; and indeed, there would not be any reason for the positions of concept formation games to contain a set $\mathcal{F}$ of known facts if only one conceptual algebra was considered possible. But as we will now see, it is easy to extend theory formation games to *distributions* of algebras:

Definition 7 (Theory Formation Games: Values of Strategies with Distributions). *Let $\mathcal{P}(\mathcal{A})$ be a probability distribution over conceptual algebras $(\mathbf{A}, \cdot)$ all having the same domain $\mathbf{A}$,*[8]*. Moreover, let β, ρ, η, C and λ be as before and let σ be a strategy. Then the value of σ with respect to the distribution $\mathcal{P}(\mathcal{A})$ is the expectation of the value of σ with respect to a random algebra $\mathcal{A}$ chosen according to the distribution $\mathcal{P}(\mathcal{A})$.*

In brief, the value of a strategy represents its ability to find valuable theories in an unknown conceptual algebra, when exploring said algebras – either by computing the combination of two concepts through $\mathtt{a} \cdot \mathtt{b} = \mathtt{?}$ moves or by attempting to find possibly novel concepts through `??` moves – carries a cost. This framework is rather reminiscent of *bandit algorithms* [23]: but whereas the main problem that bandit algorithms need to solve is to find the optimal balance between the *exploitation* of known options and the *exploration* of novel ones, in the case of theory formation games the chief difficulties that a strategy must address lie first in balancing the *search* for novel concepts with the *investigation* of the properties of the known ones, and then in selecting – on the basis of the information gathered – a set of concepts which is expected to be of optimal value.

I leave to future work the discussion and comparison of various possible – and computationally feasible – strategies in concept formation games; here, I merely observe that the problems that such strategies must overcome to be successful appear to mirror very closely the ones that humans face when involved in activities such as mathematical research, and that therefore this framework is a promising abstract testbed for the study of creatively advantageous strategies.

4 Two Examples

Until now, the framework introduced in this work has been discussed from a very abstract perspective. In this section, I will instead show two practical examples of it, related to the problem of axiomatizing a logic and to the problem of learning *falling rule list* and *definitions* from a given dataset.

Let $\mathcal{L}$ be a logic with Modus Ponens and Contraction – for instance, propositional logic – and let the (infinite) algebra $\mathcal{A}$ be the set of all Natural Deduction rules

$$\frac{P_1 \dots P_n}{Q}$$

which are valid in $\mathcal{L}$, plus the following other types of objects:

[8] For simplicity, we can assume that $\mathbf{A} = \{1 \dots N\}$ for some integer N.

– For every $i \in \mathbb{N}$, a *premise selection operator* $\texttt{Sel}_i$;
– For every $i \in \mathbb{N}$ and any rule $r \in \mathcal{A}$ with at least i premises, an object

$$\texttt{Sel}_i(r) = \frac{P_1 \dots \widehat{P_i} \dots P_n}{Q}$$

which differs from r only in that its i-th premise is highlighted;
– For every $i, j \in \mathbb{N}$ with $i < j$, an object $\texttt{Eq}_{i,j}$;
– For every atomic expression A and (possibly complex) expression S, a *substitution operator* $\texttt{Sub}(A, S)$;

The concept combination rule is defined as follows:

– For all $i \in \mathbb{N}$, if $r \in \mathcal{A}$ has at least i premises then $\texttt{Sel}_i \cdot r = \texttt{Sel}_i(r)$;
– For all $i \in \mathbb{N}$, if $r, \texttt{Sel}_i(s) \in \mathcal{A}$ and the conclusion Q of r is the same as the highlighted premise $\widehat{P_i}$ of $\texttt{Sel}_i(s)$ then $r \cdot \texttt{Sel}_i(s)$ is the rule obtained by replacing its marked premise with r. In other words,

$$\frac{S_1 \dots S_k}{P_i} \cdot \frac{P_1 \dots \widehat{P_i} \dots P_n}{Q} = \frac{P_1 \dots P_{n-1} S_1 \dots S_k P_{i+1} \dots P_n}{Q};$$

– For all $i, j \in \mathbb{N}$ with $i < j$ and rules $r \in \mathcal{A}$ with at least j premises and such that its i-th and j-th premises are equal, then $\texttt{Eq}_{i,j} \cdot r$ is the rule obtained from r by removing its j-th premise. In other words,

$$\texttt{Eq}_{i,j} \cdot \frac{P_1 \dots P_i \dots P_j \dots P_n}{Q} = \frac{P_1 \dots P_i \dots P_{j-1} P_{j+1} \dots P_n}{Q}$$

whenever $P_i = P_j$;
– For all substitution operators $\texttt{Sub}(A, S)$ and all rules r, $\texttt{Sub}(A, S) \cdot r$ is the result of replacing all occurrences of A in the premises and conclusion of r with the expression S;
– For all $a, b \in \mathcal{A}$, if none of the above cases are applicable then $a \cdot b$ is undefined.

Now let $\mathbf{T}$ be a theory in our algebra – that is to say, a finite set of rules r, premise highlighting operators $\texttt{Sel}_i$, rules with highlighted premises $\texttt{Sel}_i(r)$, premise merging operators $\texttt{Eq}_{i,j}$ and substitution operators $\texttt{Sub}(A, S)$. Then, for a fixed choice of cost C and discount rate λ, the value

$$v_{C,\lambda}(\mathbf{T}) = (1 - C)|\mathbf{T}| + \sum_{k=0}^{\infty} \lambda^{k+1} |[\mathbf{T}]_n|$$

describes the tradeoff between the cost of memorizing such a set of rules and operators and their usefulness as tools to generate further rules; and the corresponding theory formation game describes the process of seeking such a set of rules when performing deductions or sampling random rules carries a cost.

This framework can be easily modified in various ways: for instance, one could easily modify the definition of $[\mathbf{T}]_n$ by making the operators $\texttt{Sel}_i$, $\texttt{Eq}_{i,j}$

and $\texttt{Sub}(A, S)$ always available as left operands, regardless if they are in any $[\mathbf{T}]_i$, and furthermore require $\mathbf{T}$ to consist only of rules r. Then, it would also be advisable to replace the expressions $|[\mathbf{T}]_n|$ in the definition of $v(\mathbf{T})$ with a finer-grained notion of the contribution of $[\mathbf{T}]_n$ to the value of a theory, one which would likewise count only rules and – for instance – count them only up to substitution.

A more complex variation on this theme could be to adapt the framework for theory evaluation described in this work to *falling rule lists* ([24]).

Falling rule lists are classification models which can be represented as cascading sequences of `if ...then ...else ...` rules, to be applied in order to a given input until a match is found. Despite their apparent simplicity, in many cases their performances are comparable to these of less easily interpretable methods, and it was recently shown in [2] that certifiably optimal falling rule lists for categorical data[9] can be derived efficiently.

if (`age` = 23–25) ∧ (`priors` = 2–3) **then YES**
else if (`age` = 18–20) **then YES**
else if (`sex` = male) ∧ (`age` = 21–22) **then YES**
else if (`priors` > 3) **then YES**
else NO
end if

Fig. 1. A simple falling rule list for predicting recidivism, found by the CORELS algorithm [2]

One of the main advantages of falling rule lists lies in their ease of interpretability: given a falling rule list and an input, it is possible not only to extract a prediction but also to straightforwardly present a *justification* for it, that is, to produce a list of reasons why a given rule is applicable and no previous one is.

However, the interpretability of a falling rule list degrades rapidly with the increase of the number of features: while the falling rule list of Figure 1 is perfectly readable, an analogous falling rule list involving even just fifty different input features – a comparatively modest input dimensionality, in many modern applications – would be markedly less so.

A natural way to surpass this difficulty could consist in introducing *complex concepts*, defined in terms of combinations of input concepts which occur more than once in the left hand side of rules, and rewriting the falling rule list in order to make use of them, as in Figure 2: in many cases, this would be expected not only to improve the readability and interpretability of the model, but also to provide us with definitions for complex concepts which are relevant to the classification problem in exam.

[9] That is, data in which input features and outputs take values in finite, fixed domains.

	$T_2 \leftarrow A \wedge G$ $T_1 \leftarrow T_2 \wedge C$
if $(A \wedge B \wedge C \wedge E \wedge F \wedge G)$ **then YES**	**if** $(T_1 \wedge B \wedge E \wedge F)$ **then YES**
else if $(A \wedge C \wedge D \wedge G)$ **then YES**	**else if** $(T_1 \wedge D)$ **then YES**
else if $(A \wedge C \wedge G \wedge H)$ **then YES**	**else if** $(T_1 \wedge H)$ **then YES**
else if $(A \wedge F \wedge H)$ **then YES**	**else if** $(A \wedge F \wedge H)$ **then YES**
else if $(A \wedge G)$ **then YES**	**else if** (T_2) **then YES**
else NO	**else NO**
end if	**end if**

Fig. 2. Rewriting a falling rule list by introducing complex concepts: first we define T_1 as $A \wedge C \wedge G$ and rewrite rules 1–3 accordingly, then we define T_2 as $A \wedge G$ and use it in the definition of T_1 and rule 5. Does the rewrite improve readability? Are T_1 and T_2 valuable concepts? As in the framework discussed in this work, it depends on size–derivation length tradeoffs!

When generating definitions and falling rule lists, one must consider tradeoffs between the total size of the model, its predictive power, and the number of steps required to classify instances.[10] This is reminiscent of the tradeoff between representation size and derivation length which has been previously discussed in this work; and, as it will now be sketched, it can be reduced to it.

Very briefly, the elements of the algebra $\mathcal{A}$ in this case are definitions $X \leftarrow$ `Condition`, dataset instances d with their attributes, and falling rule lists indexed by dataset instances $d : \mathcal{R}$. The combination operator $\cdot$ can be used to combine two definitions, applying the former inside of the latter; or to combine a concept definition with a dataset instance which satisfies it, adding the defined concept to the instance; or to combine a dataset instance d with a rule list $d : \mathcal{R}$, causing the rule list to yield a success $d^\top$ (if its first rule is applicable to d and its output is correct), a failure $d^\bot$ (if its first rule is applicable to d but its output is incorrect), or returning the rule obtained by removing its first premise (if its first rule is not applicable to d). Figure 3 contains examples illustrating the behaviour of these rules.

It is useful to remark that, according to the above mentioned rules, the generation of a success (or a failure) for a dataset instance corresponds in a natural way to the *explanation* of its classification in terms of the given rule list and definitions. For instance, consider the rule set and definitions of Figure 2 right, and the dataset instance $d : A, C, D, G, H, \mathbf{YES}$. Then a possible way to generate $d^\top$ could be given by first applying d to the rule list, thus removing the first rule as inapplicable; then applying the two definitions to d, thus obtaining $d : A, C, D, G, H, T_2, T_1, \mathbf{YES}$; and finally applying the first rule of the derived rule list (that is, the second rule of the original rule list) **if** $(T_1 \wedge D)$ **then YES**.

[10] The motivation for attempting to minimize this number does not rise from computational costs as much as from explainability issues: the more steps – that is, definition applications or discardings of inapplicable premises – are necessary to classify an instance, the less readable the explanation of its classification will be.

$$T_2 \leftarrow A \wedge G \quad \cdot \quad T_1 \leftarrow T_2 \wedge C \quad = \quad T_1 \leftarrow A \wedge G \wedge C$$

$$T_2 \leftarrow A \wedge G \quad \cdot \quad d : A, G, H \quad = \quad d : A, G, H, T_2$$

$$d : A, F, H, \mathbf{YES} \quad \cdot \quad d : \begin{array}{l} \textbf{if } (T_2 \wedge C \wedge H) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else if } (A \wedge F \wedge H) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else if } (T_2) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else NO} \\ \textbf{end if} \end{array} \quad = \quad d : \begin{array}{l} \textbf{if } (A \wedge F \wedge H) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else if } (T_2) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else NO} \\ \textbf{end if} \end{array}$$

$$d : A, F, H, \mathbf{YES} \quad \cdot \quad d : \begin{array}{l} \textbf{if } (A \wedge F \wedge H) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else if } (T_2) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else NO} \\ \textbf{end if} \end{array} \quad = d^{\top}$$

$$d : A, F, H, \mathbf{NO} \quad \cdot \quad d : \begin{array}{l} \textbf{if } (A \wedge F \wedge H) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else if } (T_2) \textbf{ then} \\ \quad \textbf{YES} \\ \textbf{else NO} \\ \textbf{end if} \end{array} \quad = d^{\bot}$$

Fig. 3. Concept combination in the algebra of falling rule lists, definitions and dataset instances. From top to bottom: combining two definitions, combining a definition and a dataset instance, combining a dataset instance with a rule list whose first rule is not applicable, combining a dataset instance with a rule list whose first rule is applicable (correct output), combining a dataset instance with a rule list whose first rule is applicable (incorrect output).

The length of such a derivation, therefore, can be used as a measure for the complexity of the explanation.

In general, a theory $\mathbf{T}$ can be defined as a rule list, indexed by all available dataset instances, and a set of definitions.[11] In order to measure its cost more accurately, we may weigh every rule list or definition in proportion to its size; and then, we may modify the computation of the $[\mathbf{T}]_n$ by making all dataset instances available as left operands when computing it (even if they are not in $\mathbf{T}$). Finally, we change the contribution $|[\mathbf{T}]_n|$ of $[\mathbf{T}]_n$ to the value of $\mathbf{T}$ by counting the number of successes and failures in $[\mathbf{T}]_n$ (and weighing them appropriately), instead of counting the number of elements in $[\mathbf{T}]_n$ (in other words, we assign positive value to the successes, negative to the failures, and zero to the definitions and rule lists in $[\mathbf{T}]_n$). Multiplying the contribution of $[\mathbf{T}]_n$ to the value of $\mathbf{T}$ by the factor λ^{n+1} then corresponds to valuing more highly successes (and penalizing more highly failures) if they can be obtained from the initial rule list and definitions in a small number of steps; thus, maximizing the value of a falling rule list plus definitions does not simply correspond to searching for a rule list which classifies accurately as many dataset instances as possible, but rather to seeking one such that, additionally, the instances which can be successfully classified admit as short an explanation (which is to say, in the above framework, a derivation) as possible. The choice of the cost and discount rate parameters, as well as the choice of the value of successes $d^{\top}$ and failures $d^{\bot}$, defines the tradeoff between the objective of classifying correctly as many instances as possible in as few steps as possible with the objective of having an acceptably compact set of initial rule lists and definitions.

This is only a brief sketch, and more detailed presentation and analysis of this adaptation of the framework to the case of falling rule lists is left for further work. Here, it serves to illustrate a possible practical application of the framework presented here, one which in my opinion might be worth exploring further.

References

[1] Ahrweiler, P.: Modelling theory communities in science. Journal of Artificial Societies and Social Simulation 14(4), 8 (2011)

[2] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C.: Learning certifiably optimal rule lists for categorical data. stat 1050, 6 (2017)

[3] Bou, F., Schorlemmer, M., Corneli, J., Gómez-Ramírez, D., Maclean, E., Smaill, A., Pease, A.: The role of blending in mathematical invention. In: Proceedings of the Sixth International Conference on Computational Creativity June. vol. 55 (2015)

[4] Colton, S., Bundy, A., Walsh, T.: Agent based cooperative theory formation in pure mathematics. In: Proceedings of AISB 2000 symposium on creative and cultural aspects and applications of AI and cognitive science. pp. 11–18 (2000)

[5] Colton, S., Bundy, A., Walsh, T.: On the notion of interestingness in automated mathematical discovery. International Journal of Human-Computer Studies 53(3), 351–375 (2000)

[11] Nothing prevents us, in principle, from allowing multiple rule lists inside $\mathbf{T}$; but for simplicity, that possibility is not considered here.

[6] Confalonieri, R., Kutz, O., Troquard, N., Galliani, P., Porello, D., Peñaloza, R., Schorlemmer, M.: Coherence, similarity, and concept generalisation (2017), under review.
[7] Confalonieri, R., Plaza, E., Schorlemmer, M.: A process model for concept invention. In: Proc. of the 7th International Conference on Computational Creativity, ICCC16 (2016)
[8] Dyson, F.: A meeting with Enrico Fermi. Nature 427(6972), 297–297 (2004)
[9] Fauconnier, G., Turner, M.: The way we think: Conceptual blending and the mind's hidden complexities. Basic Books (2008)
[10] Galanter, P.: Computational aesthetic evaluation: Past and future. In: Computers and Creativity, pp. 255–293. Springer (2012)
[11] Gilbert, N.: A simulation of the structure of academic science. Sociological Research Online 2(2) (1997)
[12] Inglis, M., Aberdein, A.: Beauty is not simplicity: an analysis of mathematicians' proof appraisals. Philosophia Mathematica 23(1), 87–109 (2015)
[13] Inglis, M., Aberdein, A.: Diversity in proof appraisal. In: Mathematical Cultures, pp. 163–179. Springer (2016)
[14] Jordanous, A.: Evaluating evaluation: Assessing progress in computational creativity research. In: Proceedings of the second international conference on computational creativity (ICCC-11). Mexico City, Mexico. pp. 102–107 (2011)
[15] Martin, U., Pease, A.: Mathematical practice, crowdsourcing, and social machines. In: International Conference on Intelligent Computer Mathematics. pp. 98–119. Springer (2013)
[16] Mayer, J., Khairy, K., Howard, J.: Drawing an elephant with four complex parameters. American Journal of Physics 78(6), 648–649 (2010)
[17] McCormack, J.: Open problems in evolutionary music and art. In: Workshops on Applications of Evolutionary Computation. pp. 428–436. Springer (2005)
[18] McCormack, J., d'Inverno, M.: Computers and creativity: The road ahead. In: Computers and creativity, pp. 421–424. Springer (2012)
[19] Pease, A., Winterstein, D., Colton, S.: Evaluating machine creativity. In: Workshop on creative systems, 4th international conference on case based reasoning. pp. 129–137 (2001)
[20] Raman-Sundström, M.: The notion of fit as a mathematical value. In: Mathematical Cultures, pp. 271–285. Springer (2016)
[21] Righi, S., Takács, K.: The miracle of peer review and development in science: an agent-based model. Scientometrics pp. 1–21 (2017)
[22] Sutton, R.S., Barto, A.G.: Introduction to reinforcement learning, vol. 135. MIT Press Cambridge (1998)
[23] Vermorel, J., Mohri, M.: Multi-armed bandit algorithms and empirical evaluation. In: European conference on machine learning. pp. 437–448. Springer (2005)
[24] Wang, F., Rudin, C.: Falling rule lists. In: Artificial Intelligence and Statistics. pp. 1013–1022 (2015)