The importance of applying computational creativity to scientific and mathematical domains Position Paper

Abstract

Science and mathematics are currently vastly under-represented in the computational creativity (CC) community. This is at best a wasted opportunity, and at worst a significant problem for the field. We discuss why the CC community should apply their work to scientific and mathematical domains, and argue that this would be mutually beneficial for the domains in question, demonstrating our position throughout the paper by reference to examples. We propose a research programme for building closer relationships between CC and the scientific community, focusing on understandability of science and the role that AI is currently playing in scientific research.

Introduction

Despite the best efforts of successive ICCC organising committees and the computational creativity (CC) community, CC has always attracted substantially more interest from researchers in artistic than scientific and mathematical domains. In their 2017 study of application domains in CC, (Loughran and O'Neill 2017) found that, of 16 categories, papers on Maths, Science and Logic accounted for only 3% of the 353 papers on CC across 12 years. Of course, some work is domain independent, or at least not easily assigned to an academic discipline, such as the body of work on CC and curiosity (for instance (Grace et al. 2017)). Even taking this into account, it is clear that science and mathematics are vastly under-represented in our community.

There are potentially many reasons why this may be the case. Firstly, AI researchers in scientific domains may well be doing creativity-related work in other contexts but not engaging with the CC community. Automated reasoning (usually deductive reasoning in mathematics) and automated scientific discovery (usually inductive reasoning in a scientific domain) are both thriving subfields of AI, with internationally recognised journals as outlets for publication and engagement; certainly these will contain work relevant to our field but couched in different terminology with different methodologies. Secondly, it may be that other, practical, priorities in scientific domains have led to a focus on techniques such as search, data-mining and automated deduction. Since these generate results of interest to domain experts, the more difficult, fluid and tenuous concept of creativity may be seen as unnecessary, risky or simply not a priority. This may particularly be the case given the various "AI winters" in the twentieth century (the second of which ended in 1993, just six years before the first workshop on CC), and the need for AI to "prove itself" (Crevier 1993). Thirdly, it is very easy to be a hobbyist game designer or artist or composer (or a lapsed one), thus being an AI researcher by day with a side interest in another domain. Many CC researchers are deeply involved in the domains in which they work. By contrast, it is harder to be an AI researcher and also an occasional physicist, or vice versa. Fourthly, CC researchers may consider that even if generation is possible within scientific domains, evaluation is too difficult. How we should evaluate our work and our systems has always been a contentious – albeit important – issue in CC, with few proposed evaluation metrics and the majority of researchers still arguing for value along the lines of "we/people liked the system's output" or "we/people couldn't distinguish the system's output from human produced work" (Jordanous 2014). It might be the case that in science, the main evaluation metric – "is it true?", or "does it work?" – is considered simply too expensive or difficult to demonstrate. Even if evaluation is possible, we may be more prone to dismissing initial results as uninteresting in science than in artistic domains. For instance, we may get a greater sense of progress from a system working in game design which outputs a new (rather basic) game, than one working in geology which outputs a new (rather basic) result.

We argue here that neglecting scientific and mathematics domains in CC is at best a wasted opportunity, and at worst a significant problem for the field. Rapid advances in ML mean that the status of generative work in CC may be untenable, with further problems of authenticity and possible saturation in the arts. We call this the *Identity Problem*. We believe that it is essential to the health of our field that we reach out as a community at this stage, both to domain experts in science and maths and to those in related AI areas. The benefits of doing so will go both ways. As AI is used more and more in science, there is greater dependence on blackbox machine learning systems. While providing greater predictive power, this often comes at the cost of understanding. We call this the *Understandability Problem* and argue that it will become a big issue in science, which we will have to address. Twenty years of thinking about computational

creativity has provided us with valuable tools for thinking about these problems. This paper is a call to arms to CC researchers to apply their work to science, thus solving both the identity problem and the understandability problem.

Loughran and O'Neill argue that "tackling scientific, logical or realistic issues could help bring the reputation of CC away from a purely aesthetic domain towards developing solutions for real world problems." (Loughran and O'Neill 2017, p. 7) and that "It is imperative that the field remains balanced as it grows and that we remember to reflect on all areas of growth." (*Ibid.*). In this paper we support and present further arguments for this position, alongside practical recommendations for doing so.

What is science?

The concept of science is not a straightforward one. The division of the origins of learning and systematic production of new knowledge into disciplines as we know them tends to takes into account at least some of the following: methodologies, objects of study (which can be shared with other disciplines), a body of accumulated knowledge (which is generally not shared with other disciplines), theories and concepts, terminology and an institutional manifesto (so that it can reproduce itself) (Krishnan 2009, p. 9). Sciences include *Natural sciences*, which are subdivided into *physical sciences* (chemistry, physics, astronomy), *life sciences*, or biology (zoology, botany, ecology, genetics) and *earth science* (geology, oceanography, meteorology, palaeontology); *Social sciences* (psychology, sociology, economics, law, political science); *Formal sciences* (mathematics, logic, theoretical computer science, statistics); and *Applied sciences*, which are subdivided into *engineering* (computer science, civil engineering, electrical engineering, mechanical engineering); *health sciences* (medicine, dentistry, pharmacy) and *agriculture*. The number and variety of sciences makes generalisations difficult, and core values vary accordingly. However, values commonly associated in particular with the (rather unhelpfully named) "hard sciences" include repeatability, reproducibility, predictability, generality and understandability. This last value is particularly cherished: for instance, Roger G. Newton sums it up as "The primary aim of most physical scientists is to understand and explain the workings of Nature." (Newton 2000, p. 4).

The arts are possibly even harder to define. Indeed, Gallie specifically uses "Art" as an example of an essentially contested concept (Gallie 1956). This is a concept, the definition of which is "not resolvable by argument of any kind", and "the proper use of which inevitably involves endless disputes about their proper uses on the part of their users." (Gallie 1955 1956, p. 169). Julie Van Camp, writing in the context of United States Congressional policy on arts education, provides the following extensional definition: [1]

> The term 'the arts' includes, but is not limited to, music (instrumental and vocal), dance, drama, folk art, creative writing, architecture and allied fields, painting, sculpture, photography, graphic and craft arts, industrial design, costume and fashion design, motion pictures, television, radio, film, video, tape and sound recording, the arts related to the presentation, performance, execution, and exhibition of such major art forms, all those traditional arts practiced by the diverse peoples of this country. (*sic*) and the study and application of the arts to the human environment.[2]

	Science	**Arts**
Aesthetic:	truth	beauty
Approach:	problem-driven	artefact-driven
Task:	analytic	generative
Terminology:	discover	create
Status:	objective	subjective
Goal:	knowledge	self-expression

Table 1: Possible differences between the sciences and the arts

As a starting point – generalising (and controversially) – we could suggest some of the differences between the sciences and the arts as shown in Table 1. Of course, the real-world everyday lived experience of *doing* science or *doing* art is far more complex than Table 1 would suggest. Studies of interpretations of seismic data in geology, for instance, show the large number of different expert interpretations of the same seismic section, highlighting the subjectivity involved (Bond et al. 2007). These interpretations are used to analyse subsurface geology, and form the basis for many exploration and extraction decisions. Even in cases where interpreters report that an interpretation is relatively straightforward, there are significant differences in interpretation, leading to significantly different predictions, for instance about gross pore volume or gross rock volume (Rankey 2003). While clearly objectivity is the goal here, such studies may suggest that this aspect of geological practice is closer to visual art interpretation than it is to some other scientific domains.

Similarly, studies of the backstage production of mathematics show that beauty is often a guiding value (Inglis and Aberdein 2015); there is a high level of disagreement amongst experts about the validity of certain proofs (Inglis et al. 2013); and proofs and theories are often considered to be constructed rather than discovered (Lakatos 1976). Less structured knowledge such as our ability to reason logically has been shown to be highly context dependent (for instance, participants in the Wason Selection Task were unable to solve a logical problem at an abstract level but could solve it correctly when it was framed in a familiar context (Wason and Shapiro 1971)); constructing grounding metaphors to the physical world and abstract linking metaphors found to be fundamental to our understanding and construction of mathematical knowledge (Lakoff and Núñez 2000); and even the language in which reasoning occurs affecting our

[1] A visualisation of the fields of knowledge based on Wikipedia's list of academic disciplines and sub-disciplines can be found here: http://www.thingsmadethinkable.com/item/fields_of_knowledge.php

[2] http://web.csulb.edu/ jvancamp/361_r8.html

preconceptions, perceptions and assumptions (Dehaene et al. 2008; Barton 2009). An analogous story could be told in the arts; for instance, in some contexts paintings are criticised for being beautiful, with the goal being truth, or knowledge (Derrida 1987).

Dibbets expresses the relationship between arts and sciences well:

> But in the end, we all do very much of the same. All scientists, artists, composers and writers are intensively occupied imagining something that does not yet exist. They find themselves at the borders of areas where up to then hardly anyone found himself, trying to solve problems that are incomprehensible to others, trying to answer questions no one has ever asked. Here, they share a vision on things that are not yet real. (Dibbets 2002, p. 1)

Some of these interdisciplinary features are recognised in curriculum design and teaching featuring transferable skills, in which one skill may be learned within a scientific context and developed or employed in an arts context, or vice versa (see for instance (Gaff and Ratcliff 1996)). Indeed, the STEM to STEAM movement (expanding the acronym Science, Technology, Engineering, and Mathematical disciplines to include Art) explores the role of arts integration, collaboration, and experience centered learning in knowledge creation (Ghanbari 2015). Of course, the need for so many interdisciplinary initiatives (and related concepts such as transdisciplinarity, pluridisciplinarity, and multidisciplinarity) may suggest that some traditional discipline boundaries are no longer drawn in a helpful way. The evolving role and functionality of AI systems further complicates things. The focus of AI researchers, particularly in machine learning, is often on the skills they hope to simulate rather than a particular domain in which they are usually employed. This may be more productive approach than the typical CC focus on domain over skill.

The Identity Problem in Computational Creativity

Recent developments in other areas of AI – principally machine learning (ML) – have led to astonishingly rapid progress in generative processes. Research in Constructive Machine Learning has led to impressive generative results in both the arts and sciences, including painting, music, poetry, gaming, drug design, and gene design – usually in collaboration with domain experts. Our concern is that the sheer size and combined resources of the ML community may render generative work in CC untenable, potentially leading to an identity crisis in the field.

CC has long been seen as more than "mere generation".[3] Celebrating and automating other aspects of creative acts in addition to generation – such as making aesthetic judgements, producing framing information (background information about the work) and finding new meta-level processes – is partly what distinguishes us from other AI fields. As generative results in neighbouring areas of AI become more sophisticated, we may wish to focus on these other aspects of the creative act. Adding scientific domains to our repertoire will further strengthen our communal identity and enhance our value to other AI researchers and to domain experts in science.

There is also the question of whether CC output might run up against natural boundaries in some areas of the arts. For instance, it is possible that in highly expressive domains, such as poetry, computationally produced poems will not be taken seriously as valuable, given the lack of authenticity of life experiences they have. This was discussed in (Colton, Pease, and Saunders 2018), in which the authors argue that a lack of authenticity is a looming issue in the arts. The authors argue that:

> As the quality of outputs increases, we can envisage an uncanny valley stretching out, where audiences marvel at the value of the products from creative systems, while despairing at the lack of authenticity in the process and in the nature of the originator. (*Ibid.*, p7)

It is likely that authenticity is not so inherently valuable in the sciences.

Furthermore, it *may* be the case that as the novelty and backstory of computer-generated art wears off, society questions whether we want more computer-produced paintings or poems. The question as to whether we still want more computer-generated science or mathematics, however, seems less likely to be asked: we *always* want more science and mathematics. We suggest this only hesitantly. At the turn of the 20th century photography emerged as a new technology, causing an explosion in productivity of art. Flooding the market with images forced artists to redefine their value and led to the creation of modern art, transforming individual self-expression. Subareas of photography have themselves developed as unique art forms, such as wildlife photography and photojournalism. Art has further been transformed through digital technology by filters and editing. We can take inspiration from this: advances in AI can saturate old ways of thinking, but naturally open up new ones. If high quality computationally produced art becomes common-place, art as we know it will be transformed forever: a lot of concepts in art might collapse, but at the same time new concepts which are currently unpredictable might emerge. Of course, applying CC to science may equally saturate certain fields or kinds of work. We raise this here to begin a conversation on where CC in the arts may eventually lead, and as a further potential concern about focusing all our energy on the arts.

The Understandability Problem in Science

Roger Newton's quote above[4] about the primary aim of physical scientists being to understand and explain nature

[3] The slogan at the 2012 ICCC conference was "scoffing at mere generation for more than a decade." - although this has been challenged, for instance by (Ventura 2016).

[4] And many analogous quotes in mathematics, for instance: "Mathematics is not about numbers, equations, computations, or algorithms: it is about understanding" (William Thurston, quoted in (Cook 2009, p. 76)).

is uncontroversial, but difficult to unpack. Ever since the entirety of our collective scientific knowledge became too large for a single polymath to comprehend, we have had to outsource our understanding to others. The institutionalised ways in which trust of others' understanding and progress is handled started with the early universities, and developed with the invention of the printing press, academic journals, the peer review process and so on. Knowledge and understanding is a social process, as argued in (Martin and Pease 2013), but even in the human-only case, this gets complicated. The longest proof in history, of the Classification of Finite Simple Groups (CFSG), is over 10,000 pages, spread across 500 or so journal articles, by over 100 different authors from around the world, and took 110 years to complete. What does understanding mean here? Perhaps a handful of people understand the proof in its entirety, and when they die it is not obvious that any single person will ever again understand the entire proof (in part because it may be replaced with a simpler proof).

In the example of the CFSG, it is considered sufficient that someone once understood the proof. However in the ongoing case of the *abc* conjecture, this is not the case. This conjecture – proposed in 1985, on relationships between prime numbers – is considered to be one of the most important conjectures in number theory (more significant than Fermats Last Theorem; in fact Fermats Last Theorem would be a corollary of the proof). A proof would be "one of the most astounding achievements of mathematics of the twenty-first century." (Goldfeld, in (Ball 2012)). In 2012 Shinichi Mochizuki – a mathematician with a good track record, having proved "extremely deep" theorems in the past (Conrad in (Ball 2012)) – produced a 500-page proof. The problem is that the techniques and mathematical objects which Mochizuki has developed to use in his proof are so new and strange that it would take a reviewer or mathematical colleague most of their career to understand them, before they were able to understand and verify the proof. Despite some efforts from Mochizuki and a handful of his followers to make his work accessible, currently his proof has neither been published nor accepted by mainstream mathematicians, for the simple reason that they don't understand it.

Crowd-sourced mathematics, in which open conjectures are solved collaboratively via online communities, has been used for around ten years now by a subset of the mathematical community as a new way of producing mathematics through collaboration and sharing (Gowers and Nielsen 2009). Neilsen argues that this has resulted in "amplifying collective intelligence" in his book *Reinventing Discovery* (Nielsen 2011). It has certainly resulted in some original and significant new proofs (for instance, the proofs of the Bounded Gaps Between Primes and the Bounded intervals with many primes, in the 2014 Polymath8 project (Castryck et al. 2014; Polymath 2014; **?**)). Here it is perfectly possible (in fact it would be surprising if it were not the case) for a person to be a co-author but not fully understand the proof in their own paper.

Adding computers to the social process, to form a combination of people, computers, and mathematical archives to create and apply mathematics – a "mathematics social machine" (Martin and Pease 2013) – further complicates matters. Take *automated theorem proving*, the task of deciding whether a given formal statement follows from a given set of premises (Sutcliffe and Suttner 2001). The least informative approach would be to produce merely a "yes", "no" or "unknown" response. Not only is this devoid of *explanation*, but it also hides the effects of any bugs; requiring the user to either trust the results, or verify the implementation.

This can be mitigated by having the system instead generate a *proof object*: a formal argument for *why* a given statement follows or does not follow. Once generated, a proof object's validity can be checked without requiring any knowledge of how it was created, thus avoiding the need to trust or verify the (potentially complicated) search and generation procedures. Theorem provers which produce proof objects that are trivial to check by independent *proof checker* programs (which are themselves easily verified, due to their simplicity) satisfy the *de Bruijn criterion* (Barendregt and Wiedijk 2005); examples are Coq (Barras et al. 1997) and Isabelle/HOL (Nipkow, Paulson, and Wenzel 2002).

Proof objects are not a complete solution to understandability, since they can still be quite inscrutable to human users. This often depends on how closely the chosen formal system is able to encode the user's ideas: for example, the formal proof of the Kepler Conjecture was performed using a system of Higher Order Logic (HOL) (Hales et al. 2015) whose proof objects (natural-deduction style derivations), whilst tedious, are in principle understandable to a user experienced with both the software and problem domain. The same cannot be said of the Boolean Pythagorean Triples problem, a statement of Ramsey theory involving the structure of the natural numbers. Rather than taking a high-level approach like HOL, (2016) analysed sets $\{0 \dots n\}$ for larger and larger n, encoding these restricted versions of the problem into the language of boolean satisfiability (SAT), and found that the problem is unsatisfiable for $n = 7825$, and hence for the natural numbers as a whole. In this case, the proof object demonstrates this unsatisfiability using 200 terabytes of propositional logic clauses (compressable to 68 gigabytes). Not only is this far too much for any human to comprehend, but the concepts used in the proof (boolean formulae) are several layers removed from the actual problem statement (natural numbers, subsets and pythagorean triples).

Whilst "low level" formalisms like SAT are less understandable or explanatory for users, they are far more amenable to automation than more expressive logics. Despite the proof for the Boolean Pythagorean Triples problem being many orders of magnitude larger than that of the Kepler Conjecture, the latter is well beyond the ability of today's automated theorem provers due to its encoding in HOL. Instead, it took 22 collaborators 9 years just to formalise the proof (Hales had previously produced a less formal proof, hundreds of pages long and accompanied by unverified software; yet another reminder that human-generated artefacts are not necessarily understandable either).

The situation is even worse in automated scientific discovery. The widespread use of *machine learning* (ML) to

find patterns in scientific data has been criticised by Genevera Allen in her recent talk at the 2019 Annual Meeting of the American Association for the Advancement of Science (AAAS)[5]. She highlighted accuracy and reproducibility issues with scientific discoveries made by machine-learning techniques. Understandability is another huge problem with such discoveries; often with a tradeoff between the generality of an approach and how easily its resulting behaviour can be understood. The understandability problem concerns this last issue.

Proposed solutions

Here we propose three different approaches that the CC community can take to address the understandability problem. Each approach, at the same time, would solve the saturation problem, by focusing on science. We end each subsection with a recommendation towards a new Research Programme in CC.

The "human-like computing" approach

Human-like computing research is a research programme developed for a UK funding initiative by the Engineering and Physical Sciences Research Council.[6] This programme "aims to endow machines with human-like perceptual, reasoning and learning abilities, which support collaboration and communication with human beings.", with one of the stated motivations to "inspire new forms of computation based on human cognition, especially on tasks where humans currently exhibit superior abilities."[7]

We believe that a focus on more human-like computing would result in systems whose output more closely resembled human produced work. If this were the case, then we hypothesise that there would be greater understandability. In the context of science and mathematics, there is much theoretical work on how people work in these domains and how knowledge is constructed. For instance, the philosophy of informal mathematics and the study of mathematical practice and cultures are thriving communities with annual research events and a good number of published books and papers. We can further this work by building human-like computing systems which model theoretical findings.

The CC community is particularly well equipped to work in this research programme. The FACE evaluation model (Colton, Pease, and Charnley 2011) is based on the multiple aspects involved in the human creative act. These include aesthetic judgements; concept development; contextual, framing information; and meta-level processes. This might be reflected in automated theory formation approaches to Automated Reasoning, which consider a far wider view of mathematical knowledge production than the traditional narrow focus of Automated Theorem Proving, including the automatic generation and evaluation of new conjectures, concepts, examples explanations, and so on (see (Pease, Colton, and Charnley 2013) for an example). In some ways, given the closeness of some (aspects of some) sciences to artistic domains, this may be low hanging fruit for system developers to apply their systems to scientific and mathematical domains.

Recommendation 1: Apply your system to scientific domains.

The "Framing " approach

Enhancing software with explanatory functionality would also help to mitigate the understandability problem. The "F" from the FACE model stands for *Framing*, and we advocate a dual-approach of software generating framing information alongside an artefact, problem solution or new data pattern (proof objects can be seen as a limited form of framing information). We foresee this being an increasingly important area of research in CC, with an increasing level of sophistication: from explanation to justification to argument and dialogue with a user about the value, method of production, motivation etc. behind output. How framing information should be developed is a research programme in its own right. For now, we discuss greater and lesser understandability in terms of describing the processes underlying the generative act and consider these for ML approaches.

Many ML approaches can be characterised as constructing a computer program (or "model") consisting of two parts: an overall structure or *architecture*, which remains fixed; and a set of adjustable *parameters*, which are inferred or "learned" from data (e.g. *training data* of desired input/output examples). One particularly simple architecture is the *decision tree*: nested boolean queries of the input, often used for *classification* (Safavian and Landgrebe 1991). These queries are parameters, and are chosen based on how efficiently they separate the classes given in the training data. Decision trees usually perform poorly compared to other ML algorithms, but are nominally understandable since their behaviour on a given input traces a single path through these queries, which could be turned into framing information such as "Class x was chosen because y was greater than z...". The *random forest* approach gives better performance by combining many decision trees and having them vote on the overall outcome (Breiman 2001), although such ensemble behaviour is more difficult to reason about than that of a single tree, and is hence harder to frame in an understandable way. One approach might be to find patterns in the votes, such as "Class x was chosen because most trees looking at features y and z voted for it".

Recent research has focused on highly expressive classes of models such as *differentiable programming* (Wang et al. 2018), whose architectures output not only a (numerical) answer, but also partial derivatives with respect to the parameters; and *probabilistic programming*, which samples from a *distribution* conditioned on the training data. Both frameworks allow arbitrary architectures, specified via Turing-complete languages, and provide efficient, composable methods for optimising the parameters (e.g. Stochastic Gradient Descent and Markov Chain Monte Carlo, respectively) to minimise arbitrary loss functions (e.g. output error

[5]https://eurekalert.org/pub_releases/2019-02/ru-cwt021119.php

[6]http://hlc.doc.ic.ac.uk

[7]https://epsrc.ukri.org/newsevents/pubs/human-like-computing-strategy-roadmap/

for the training data).

With such expressive formalisms, the conflict between the generality of a model and its understandability becomes clear. Task-specific architectures require fewer parameters than general-purpose approaches, perform well with little training data, and are amenable to descriptive framing information. For example, hand-written characters can be classified based on a single example if we allow our model to assume the given images are generated by pen strokes (Lake, Salakhutdinov, and Tenenbaum 2015), and these models may allow descriptions such as "Character *c* was chosen because there appear to be *x* long strokes, *y* curved strokes, etc.". Likewise the parameters of a 3D scene (such as object position and lighting) can be inferred from images if a ray-tracer is embedded in the model (Li et al. 2018); embedding a physics engine enables predictions about these scenes, which are useful e.g. for robot controllers (Degrave et al. 2016). Whilst more complicated than the previous examples, such a controller could (in principle) justify its actions based on interpretations of the model, such as "The motor was engaged because the pendulum appeared to be falling to the left".

However, the specificity that makes these implementations understandable also makes them unsuitable for any other task. The choice of such high-level, task-specific components is performed by the user, and encodes some of their domain knowledge into the structure of the solution, such that it doesn't have to be learned. This is similar to how high-level logics can represent relevant domain concepts (like natural numbers and sets), yet proof methods making use of this have limited reusability due to the difficulty of automating such high-level reasoning.

At the other end of the spectrum are general purpose architectures, like *artificial neural networks* (differentiable programs capable of universal function approximation (Funahashi 1989)). These are compositions of a large number of identical sub-expressions ("neurons"), whose parameters ("weights") scale their input values, and hence the contribution of each sub-expression to the whole. Such architectures encode essentially no domain knowledge, requiring much larger training sets than task-specific algorithms in order to "learn" these details. So much of these general purpose models' behaviour comes from tuning their (many) parameters, that understanding or describing their high-level behaviour is difficult; indeed they are often treated as inscrutable "black boxes", akin to the large (un)SAT proof described above.

Understanding exactly how such programs make their decisions is an active area of research, known as *explainable artificial intelligence* (XAI) (Došilović, Brčić, and Hlupić 2018; Doshi-Velez and Kim 2017; Molnar 2019). Saliency maps are a popular form of framing information (Simonyan, Vedaldi, and Zisserman 2013), which reverse-engineer the factors which lead to a particular decision made by a model (for example judging the saliency of input pixels by how strongly they each effect the output prediction if adjusted). These methods appear intuitive, e.g. producing visualisations highlighting a particular object in a scene as the reason for its classification; yet this can obscure the difficulty of interpreting such high-dimensional decision boundaries. In particular, reasonable justifications (such as classifying an image as a butterfly, with high saliency given to those pixels which show the butterfly) can be fundamentally altered by imperceptible adjustments to the input (Ghorbani, Abid, and Zou 2017) (in this case choosing butterfly based only on the background vegetation). Similar adjustments can also change a model's output, leading to the field of *adversarial machine learning* (Goodfellow, Shlens, and Szegedy 2014); adjustments to even a single pixel can not only cause a system to mislabel an input, but to give high confidence to its erroneous result (Su, Vargas, and Sakurai 2019).

Attempts to understand the internal operation of models themselves include techniques like activation maximisation (Erhan et al. 2009): iteratively perturbing the input to maximise the value of a chosen internal component. The result is an input (usually an image) which provides a strong stimulus for that one component, and hence visualises the sorts of features that component has learned to detect. Similar techniques like deconvolution can also generate such imagery (Zeiler and Fergus 2014), but all require some degree of qualitative interpretation is to understand what the system may be focusing on. Alternative use cases for these methods, like Google's "DeepDream", perturb user-provided images to maximise the activation of certain neurons of a pre-trained model, which have learned to detect objects like faces or animals, or artwork following a certain style, etc. The resulting images resemble the original, but with an artificial form of pareidolia (Mordvintsev, Olah, and Tyka 2015).

Whilst much attention has been focused on making the outputs of ML systems more accurate and robust, there is also a need for framing information which explains more, is more understandable to users and less prone to misinterpretation.

Recommendation 2: Enhance your system with framing capabilities.

The "forgoing understandability" approach

It may be the case that, given the increase in power, generality and predictiveness that ML approaches give, and the increasing complexity and amount of scientific knowledge, we decide to forgo understandability in science. As a community we would be in a unique position to develop thinking on this, and to answer questions such as whether we should try to replace understandability with something else. We suggest identifying and engaging with stakeholder groups in science and mathematics to ensure that we develop in directions which will be fruitful and useful to society.

Another possible solution to the problems described here would be to forego understandability in the current sense, or rather to change our notion of what *kind* of thing we are aiming to understand. For instance, could a neural network itself be considered to be a scientific discovery, analogous to the discovery of a new plant? It may be that AI systems become objects of study in the same way as the human brain is currently an object of study, with methods and approaches from neural science, psychology, cognitive science and so on employed to understand an AI system and its behaviour

and interactions. There is an interesting analogy between ways in which we can "interrogate" a neural network, for instance via generating inputs aligned to deep features (by specifying a deep-level state, then "training" the input to get close), and how we use introspection and analysis to understand human learning. We're gradually becoming cognitive scientists and psychologists for the robots.[8] This is already an active research area, with (Jonas and K.P. 2017) offering a cautionary tale. Again, as a community we would be in a position to provide a unique perspective on this, having reflected on what constitutes an artefact and how they might be evaluated as novel or significant discoveries.

Recommendation 3: Produce philosophical work on what computational creativity should mean, and what science done with computers should entail.

Concluding Remarks

Deep learning and ML are making inroads everywhere: generative arts, processors, Go, machine vision, and so on, and we need to consider as a community where this leaves us. Our suggestion in this paper is to focus on science and mathematics, where we have much to contribute.

People are not naturally good at science. The history of science and scientific methodology, the length of time it takes to train a scientist and the high number of published research findings in science which are considered to be false or sub-standard[9] all hint at the difficulty of the scientific enterprise. This is partially due to political and institutional factors such as pressure to publish, conflicts of interest and a culture which is often more competitive than collaborative; but also partially due to the constant battle to avoid the large number of cognitive biases that adversely affect our reasoning and judgements (Haselton, Nettle, and Andrews 2005; Sutherland 2013). On the other hand, the arts – while also difficult to do well – do not usually go against our natural way of thinking, and can be seen as a celebration of our humanity. In many ways science should be an obvious application domain for computational creativity. This paper is a call to arms for the whole CC community, to apply their systems to scientific and mathematical domains, to enhance their systems with framing functionality, and to produce philosophical work on new directions in our field.

References

Asimov, I. 1950. *I, Robot*. Gnome Press.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* (533):452–454.

Ball, P. 2012. Proof claimed for deep connection between primes. *Nature* 489(7414).

[8]The term "robopsychologist" was coined by Isaac Asimov in (Asimov 1950) to describe the study of the personalities and behavior of intelligent machines.

[9]Meta-scientific studies suggest that 85% of biomedical research efforts are wasted (Macleod et al. 2014) and 90% of respondents to a recent survey in Nature agreed that there is a 'reproducibility crisis' (Baker 2016) (see (Munafò et al. 2017; Ioannidis 2005) for further details).

Barendregt, H., and Wiedijk, F. 2005. The challenge of computer mathematics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363(1835):2351–2375.

Barras, B.; Boutin, S.; Cornes, C.; Courant, J.; Filliatre, J.-C.; Gimenez, E.; Herbelin, H.; Huet, G.; Munoz, C.; Murthy, C.; et al. 1997. *The Coq proof assistant reference manual: Version 6.1*. Ph.D. Dissertation, INRIA.

Barton, B. 2009. *The Language of Mathematics: Telling Mathematical Tales*. Mathematics Education Library, Vol. 46. Springer.

Bond, C. E.; Gibbs, A.; Shipton, Z.; and Jones, S. 2007. What do you think this is? "Conceptual uncertainty" in geoscience interpretation. *GSA Today* 17(11):4–10.

Breiman, L. 2001. Random Forests. *Machine Learning* 45(1):5–32.

Castryck, W.; Fouvry, E.; Harcos, G.; Kowalski, E.; Michel, P.; Nelson, P.; Paldi, E.; Pintz, J.; Sutherland, A. V.; Tao, T.; and Xie, X.-F. 2014. New equidistribution estimates of zhang type. *Algebra & Number Theory* 8(9):2067–2199.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *2nd International Conference on Computational Creativity*.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proceedings of the Ninth International Conference on Computational Creativity*.

Cook, M. 2009. *Mathematicians: an outer view of the inner world*. Princeton University Press.

Crevier, D. 1993. *AI: The Tumultuous Search for Artificial Intelligence*. New York, NY: BasicBooks.

Degrave, J.; Hermans, M.; Dambre, J.; et al. 2016. A differentiable physics engine for deep learning in robotics. *arXiv preprint arXiv:1611.01652*.

Dehaene, S.; Izard, V.; Spelke, E.; and Pica, P. 2008. Log or linear? Distinct intuitions of the number scale in Western and Amazonian Indigene cultures. *Science* 320(5880):1217–1220.

Derrida, J. 1987. *The Truth in Painting*. Univ. of Chicago Press.

Dibbets, J. 2002. Interactions between science and art. *Cardiovascular Research* 56(3):330–331.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Došilović, F. K.; Brčić, M.; and Hlupić, N. 2018. Explainable artificial intelligence: A survey. In *41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 0210–0215. IEEE.

Erhan, D.; Bengio, Y.; Courville, A.; and Vincent, P. 2009. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univerist de Montral*.

Funahashi, K.-I. 1989. On the approximate realization of continuous mappings by neural networks. *Neural networks* 2(3):183–192.

Gaff, J. G., and Ratcliff, J. L. E. 1996. *Handbook of the Undergraduate Curriculum A Comprehensive Guide to Purposes, Structures, Practices, and Change - The Jossey-Bass Higher and Adult Education Series*. Wiley.

Gallie, W. B. 1955 – 1956. Essentially contested concepts. *Proceedings of the Aristotelian Society* 56:167–198.

Gallie, W. 1956. Art as an essentially contested concept. *The Philosophical Quarterly* 6(23):97–114.

Ghanbari, S. 2015. Learning across disciplines: A collective case study of two university programs that integrate the arts with stem. *International Journal of Education and the Arts* 16(7).

Ghorbani, A.; Abid, A.; and Zou, J. 2017. Interpretation of neural networks is fragile. *arXiv preprint arXiv:1710.10547*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gowers, T., and Nielsen, M. 2009. Massively collaborative mathematics. *Nature* 461(7266):879–881.

Grace, K.; Maher, M.; Mohseni, M.; and Pérez y Pérez, R. 2017. Encouraging p-creative behaviour with computational curiosity. In *Proceedings of the 8th International Conference on Computational Creativity*.

Hales, T.; Adams, M.; Bauer, G.; Dang, D. T.; Harrison, J.; Hoang, T. L.; Kaliszyk, C.; Magron, V.; McLaughlin, S.; Nguyen, T. T.; et al. 2015. A formal proof of the Kepler conjecture. *arXiv preprint arXiv:1501.02155*.

Haselton, M. G.; Nettle, D.; and Andrews, P. W. 2005. The evolution of cognitive bias. In Buss, D. M., ed., *The Handbook of Evolutionary Psychology*. Hoboken, NJ, US: John Wiley and Sons Inc. 724–746.

Heule, M. J.; Kullmann, O.; and Marek, V. W. 2016. Solving and verifying the boolean pythagorean triples problem via cube-and-conquer. In *International Conference on Theory and Applications of Satisfiability Testing*, 228–245. Springer.

Inglis, M., and Aberdein, A. 2015. Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica* 23(1):87–109.

Inglis, M.; Pablo, J.; Mejia-Ramos; Weber, K.; and Alcock, L. 2013. On mathematicians' different standards when evaluating elementary proofs. *Topics in Cognitive Science* 5(2):270–282.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLOS Medicine* 2(8).

Jonas, E., and K.P., K. 2017. Could a neuroscientist understand a microprocessor? *PLoS Comput Biol* 13(1).

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*.

Krishnan, A. 2009. What are academic disciplines? Some observations on the disciplinarity vs. interdisciplinarity debate. Technical Report 03/09, National Centre for Research Methods.

Lakatos, I. 1976. *Proofs and Refutations*. Cambridge: Cambridge University Press.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Lakoff, G., and Núñez, R. 2000. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York: Basic Books.

Li, T.-M.; Aittala, M.; Durand, F.; and Lehtinen, J. 2018. Differentiable Monte Carlo ray tracing through edge sampling. In *SIGGRAPH Asia 2018 Technical Papers*, 222. ACM.

Loughran, R., and O'Neill, M. 2017. Application domains considered in computational creativity. In *Proceedings of the Eighth International Conference on Computational Creativity (ICCC-17)*, 197–204.

Macleod, M. R.; Michie, S.; Roberts, I.; Dirnagl, U.; Chalmers, I.; Ioannidis, J.; Al-Shahi Salman, R.; Chan, A.; and Glasziou, P. 2014. Biomedical research: increasing value, reducing waste. *Lancet* (383):101–104.

Martin, U., and Pease, A. 2013. Mathematical practice, crowdsourcing, and social machines. In Carette, J.; Aspinall, D.; Lange, C.; Sojka, P.; and Windsteiger, W., eds., *Intelligent Computer Mathematics*, volume 7961 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 98–119.

Molnar, C. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Mordvintsev, A.; Olah, C.; and Tyka, M. 2015. Deepdream-a code example for visualizing neural networks. *Google Research* 2(5).

Munafò, M. R.; Nosek, B. A.; Bishop, D. V. M.; Button, K. S.; Chambers, C. D.; Percie du Sert, N.; Simonsohn, U.; Wagenmakers, E.-J.; Ware, J. J.; and Ioannidis, J. P. A. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1(1):0021+.

Newton, R. G. 2000. *The Truth of Science: Physical Theories and Reality*. Cambridge, Massachusetts: Harvard University Press.

Nielsen, M. 2011. *Reinventing Discovery: The New Era of Networked Science*. USA: Princeton University Press.

Nipkow, T.; Paulson, L. C.; and Wenzel, M. 2002. *Isabelle/HOL: a proof assistant for higher-order logic*, volume 2283. Springer Science & Business Media.

Pease, A.; Colton, S.; and Charnley, J. 2013. Automated theory formation: The next generation. *IFCoLog Lectures in Computational Logic*.

Polymath, D. 2014. Variants of the selberg sieve, and bounded intervals containing many primes. *Research in the Mathematical Sciences* 1(12).

Rankey, E. C. 2003. Interpreter's corner – that's why it's called interpretation: Impact of horizon uncertainty on seismic attribute analysis. *The Leading Edge* 22(9).

Safavian, S. R., and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21(3):660–674.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 1–1.

Sutcliffe, G., and Suttner, C. 2001. Evaluating general purpose automated theorem proving systems. *Artificial intelligence* 131(1):39–54.

Sutherland, S. 2013. *Irrationality: The enemy within*. Pinter and Martin Ltd.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity*.

Wang, F.; Wu, X.; Essertel, G.; Decker, J.; and Rompf, T. 2018. Demystifying differentiable programming: Shift/reset the penultimate backpropagator. *arXiv preprint arXiv:1803.10228*.

Wason, P. C., and Shapiro, D. 1971. Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology* 23:63–71.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818833. Springer.