Specific curiosity as a cause and consequence of transformational creativity

Kazjon Grace & Mary Lou Maher
Department of Software and Information Systems
University of North Carolina at Charlotte
Charlotte, NC 28203 USA
{k.grace,m.maher}@uncc.edu

Abstract

This paper describes a framework by which creative systems can intentionally exhibit transformational creativity. Intentions are derived from surprising events in a process based on specific curiosity. We argue that autonomy of intent is achieved when a creative system directs its generative processes based on knowledge learnt from within its creative domain, and develop a framework to elaborate this behaviour. The framework describes ways that transformation of the creative domain can arise: from learning, from a serendipitous situation, and as a result of intentional exploration. Examples of each of these kinds of transformation are then illustrated through examples in the domain of recipes.

Introduction

Significant effort has been devoted to developing computational models that can recognise creative artefacts, on the assumption such a capability could be used to generate creative artefacts if paired with an appropriate search algorithm. However, generate-and-test creative systems lack any kind of *autonomous intent*: they never decide to make a green artefact, or a loud one, or a happy one, unless such qualities are built into their externally-provided objective function. As classically formulated, a search function does not distinguish two points within its space in any way but by the objective, and thus has no intent that can be defined with the representations that define that space, only in how the resulting artefacts perform. Search functions that can modify their goals while searching (Gebser, Kaufmann, and Schaub 2009), or that search based on specific past experiences (Cully, Clune, and Mouret 2014) do exist, but, from a computational creativity perspective, there remains an unanswered question: under what conditions should a system decide to modify its search?

At first this lack of autonomous intentions in our systems' search processes may fail to seem problematic: we are not constrained by cognitive plausibility. There is no inherent reason why intentionality, while clearly a quality of human creators, should be required in their digital analogue. Our goal is systems which produce output that would be considered creative, regardless of the processes involved. On closer inspection, however, autonomy of intention may not be so easily discarded from creativity. Intent is intrinsically tied to definitions of art and creativity (Dewey 2005), where the debated questions concern not whether an artefact's creator had intent, but whether that intent should be privileged over observers' interpretations (Best 1981). Intention is seen among human creators as critical both to the production and consumption of creative artefacts – evidence that argues for its role in appreciative as well as generative computational processes.

Autonomy of intent also provides critical information for use in framing. A creative system's ability to construct framing narratives for its work – considered critical to any computationally creative construct (Charnley, Pease, and Colton 2012) – stems from its ability to provide *justification* for creative decisions. Without autonomy of intent these justifications can only be driven by external objectives (e.g. "I wanted to make the artefact seem brighter"), not intrinsic motivations (e.g. "I was exploring how colour influenced brightness"). Human creators make the decision to explore a particular set of concepts, and follow that exploration to its resolution by way of creative expression. Framing, as the channel by which a creative system can convince its audience of its creative autonomy, should explain such explorations. Previous models of intent in framing have been based on information extrinsic to the creative domain, such as the day's top news stories (Krzeczkowska et al. 2010), but we argue that without learning how to connect such external knowledge to the creative domain (e.g. through analogy), then such intent cannot be autonomous.

How, then, can a creative system derive intent from its knowledge about the creative domain? On what basis should it transform its inspiring set and own past creations into contextual constraints on its search process? For one possible answer we turn to cognitive studies of how human designers think during the process of designing, and how their search for a creative solution affects itself. Human designers do not sequentially analyse a problem, synthesise solutions to it and then evaluate those solutions, but instead switch between those processes iteratively (Schön 1983), finding new problems as frequently as they find new solutions (Weisberg 1993). This co-evolution of problem-framing alongside problem-solving becomes more evident in expert designers (Cross 2004), and – more critically for our purposes – has been shown to produce more valuable output (Getzels and Csikszentmihalyi 1976). A cognitive protocol analysis of

sketching architects found that not only did they regularly unexpected discover features in their own drawings, but that those discoveries often led to reformulation of the design task (Suwa, Gero, and Purcell 1999). These reformulations led in turn to more unexpected discoveries, evidence that this cycle of intentionality and exploration is beneficial, if not central, to human creativity. We seek to capture this cycle in the computational model presented in this paper.

We propose that the inspiration for a computational model of intentional creativity can come from the iterative process of defining the creative task and solving it in parallel. We propose that intentions are not created *de novo*, but that they arise from a drive to explore what the system has observed but not understood, both from its own output and that of other creators. The catalyst for this exploratory behaviour is unexpectedness: a creator being surprised by an artefact, and forming the intention to explore some part of the design space in return. We refer to this as a kind of *specific curiosity*, after the distinction between specific and diversive curiosity first articulated by Berlyne (1966).

We frame our model for specific curiosity as an extension of Wiggins (2006) framework for describing exploratory and transformational creativity. With that symbolic representation we can then describe how transformational creativity leads to surprise, and how surprise can in turn lead to further creativity.

Transformational creativity, surprise and their effects on behaviour

This paper describes a model of autonomous intent in creative systems, drawing on theories of evaluating creativity, psychological studies of curiosity and cognitive studies of how designers respond to unexpected discoveries. We introduce each of those literatures here.

Three long-lost cousins: novelty, transformational creativity and surprise

Novelty (Newell, Shaw, and Simon 1959; Saunders and Gero 2001), surprise (Macedo and Cardoso 2001; Grace et al. 2014) and domain transformation (Boden 2003; Wiggins 2006) are three core ideas around which the debate on how to computationally recognise creative artefacts has revolved. In Grace and Maher (2014) we outlined how each of those three could be connected to the notion of unexpectedness, establishing one possible way to compare them in a common language.

Novelty was, to the authors' knowledge, first floated alongside value by Newell, Shaw and Simon (1959), forming the closest thing to a broadly-accepted definition for creativity that we have today. Novelty and value are proposed as necessary and complementary aspects of creativity: a solely valuable artefact is merely good, while a solely novel artefact is merely weird. Novelty is typically conceptualised as difference from that which is known (Sternberg and Lubart 1999), and usually operationalised by a distance measure between a new observation and past experiences. An alternate view of novelty is based on the degree to which observing an artefact helps an agent to understand the world (Schmidhuber 2010), proponents of which criticise the distance-based approach as attributing overly high novelty to noise.

Boden (2003) proposed another solution to the problem of distinguishing meaningful novelty from noise by focusing on impact. *Transformational* creativity is based on the degree to which an artefact changes the creative domain to which it belongs. This is suggested by Boden to be a more significant form of creativity than the combination of "mere" novelty and value, which she considers the result of *exploratory* creativity. Wiggins (2006) formalises Boden's definition of transformational creativity and provides a general description of a creative system that is capable of it, although he questions Boden's strict hierarchical superiority of transformation over exploration.

The authors have previously proposed unexpectedness and surprise as an alternative formulation of novelty (Grace et al. 2014), although we are far from the first to do so (Macedo and Cardoso 2001). Unexpectedness is the degree to which observing an artefact violates (i.e. opposes) an agent's confident predictions about the world. The flexibility of this approach is in the source of predictions, which may be relationships within the artefacts, trends derived from the domain's history, or other sources of knowledge. Novelty can be described from this perspective as a form of unexpectedness based on the predicting that the domain will continue as it has in the past. Surprise is an affective response to unexpectedness: unexpected artefacts induce surprise in their observers. Transformational creativity can be described as a quantification of surprise based on how much a new artefact changed domain knowledge. This connection was described in Baldi and Itti (2010), who used an information theoretic perspective to connect measuring surprise by (un-)likelihood to measuring it by impact on knowledge.

Throughout this paper we adopt the viewpoint that these three notions are intimately connected, constituting complementary perspectives on how a creative artefact can be meaningfully different from those that preceded it. We argue that the evaluative processes of creative systems should possess the ability to detect all of the above aspects of meaningful difference, and that any one of them – in conjunction with value – can indicate creativity.

Curiosity and the pursuit of novelty

Curiosity is an overloaded term in psychology, referring both to a trait possessed by different people to different degrees, as well as to motivating state that drives its experiencers to seek novel stimuli (Berlyne 1966). The latter definition, curiosity as a state, has been proposed as a motivator for computational creative systems (Saunders and Gero 2001; Merrick and Maher 2009), based on the principle that novelty-seeking (alone or alongside value) will drive exploration towards creative solutions.

Berlyne distinguishes state-curiosity along two axes: perceptual vs epistemic and specific vs diversive. Perceptual curiosity is the drive towards novel sensory stimuli, and has been observed in a variety of animals of different cognitive capabilities. Epistemic curiosity is the drive to acquire novel knowledge This conceptual curiosity can be modelled by

systems that learn a conceptual space and measure novelty within it, rather than measuring between artefacts at the level of sensory input (Saunders and Gero 2001). The distinction between creativity at the sensory and knowledge-levels has been drawn within computational creativity by Smith and Mateas (2011), who refer to the latter as "rational curiosity".

The specific/diversive division has received less attention in computational creativity. Specific curiosity is the search for observations that explain or elaborate a particular goal concept. Diversive curiosity, on which most computational models of curiosity have focussed, is the search for new information without any specific targets. While the search for a specific concept can be modelled by search, the challenge is how to trigger specific curiosity: when and why should a creative system become specifically curious? This is related to the broader issue of creative autonomy (Jennings 2010; Saunders 2011). In this paper we develop a model of specific curiosity that uses surprise as a way to address this challenge.

How surprises affect designing

Cognitive studies of human creators – particularly in the field of design – have shown that surprise significantly impacts the creative process. Designing has been described as a "reflective conversation with the medium" (Schön 1983), meaning that designers iteratively synthesise new additions to their emerging design and then reflect on their effects. Expressing creative artefacts through rough yet external representations – usually referred to as sketches in the case of human designers – is a critical component of the creative process as it allows designers to observe changes they did not consciously make (Schon and Wiggins 1992; Goldschmidt 1991). Through this externalisation a designer may perceive an emergent shape, discover a new relationship between components, or construct an analogy to past designs. Several computational creativity systems have adopted this cyclical reflective approach in whole or in part, including the search-bias transformation in DeLeNoX (Liapis et al. 2013), the interpretation-driven mapping of Idiom (Grace, Gero, and Saunders 2015) and the expectation-based reinterpretation of Kelly and Gero (Kelly and Gero 2014).

This iterative process of "seeing" (perceiving an emerging design) and "moving" (making a change to it) allows designers to read more off a sketch than they originally put there (Schon and Wiggins 1992). Though the term has since been corrupted beyond recognition, this was the original meaning of *design thinking*: an iterative, reflective, solutions-focussed strategy as opposed to a step-by-step, analytical problem-focused one (Lawson 2006). In a "think aloud" cognitive protocol study where architects were observed designing, unexpected discoveries were bi-directionally causally connected to reformulation of the design goals, i.e.: surprises led to transformation of the problem, and transformation of the problem led to surprises (Suwa, Gero, and Purcell 1999). These results with human creators suggest that surprise-triggered specific curiosity might be useful for encouraging transformative creativity in artificial creative systems. In the remainder of this paper we develop a framework for how that behaviour could be operationalised.

Unexpectedness-triggered specific curiosity: A model of transformation-seeking behaviour

We adopt the creative systems framework from (Wiggins 2006) to describe our model of unexpectedness and specific curiosity. Wiggins' framework describes a creative system in terms of a search process that traverses a conceptual space to generate artefacts, coupled with a metacognitive search process that traverses the space of all possible conceptual spaces. The resulting system is capable of both exploratory and transformational creativity, with the latter represented as exploration at the meta-level. The following symbols define the core of the framework, although readers are encouraged to familiarise themselves with the original, which affords each definition far greater depth:

- $\mathscr{U}$ is the *universe*, the space of all possible distinct concepts that make up all possible representations of artefacts in the current creative domain.
- $\mathscr{L}$ is the *ruleset language*, the set of all possible rules that act on concepts the creative system can construct.
- $[\![.]\!]$ is the *definition interpreter* that takes a subset of $\mathscr{L}$ and acts on a set of concepts, yielding real numbers in [0,1]. This is used to apply a rule set to a set of concepts, assigning a value to each.
- $\mathscr{R} \subseteq \mathscr{L}$ is a *constraint ruleset*, by which the system defines the scope of the conceptual space (within $\mathscr{U}$).
- $\mathscr{C}$ is a *conceptual space* is the current subset of $\mathscr{U}$ permitted by $\mathscr{R}$. i.e., $\mathscr{C} = [\![\mathscr{R}]\!](\mathscr{U})$.
- $\mathscr{T} \subseteq \mathscr{L}$ is a *traversal ruleset*, by which the system explores $\mathscr{C}$.
- $\mathscr{E} \subseteq \mathscr{L}$ is an *evaluation ruleset*, by which the system evaluates proposed concepts.
- c_{in} is the *input set*, a totally ordered subset of $\mathscr{U}$ that reflects the list of artefacts known to the system, in the order of the system's observation of them.
- c_{out} is the *output set*, a totally ordered subset of $\mathscr{U}$ that reflects the output of the creative system after a particular generative iteration.
- $\langle\!\langle .,.,. \rangle\!\rangle$ is the *generation interpreter* that takes three subsets of $\mathscr{L}$, the rules that define the conceptual space $\mathscr{R}$, the rules that define how to traverse that space $\mathscr{T}$, and the rules that assign value to members of that space, $\mathscr{E}$ and acts on the set of all previously observed artefacts to generate a new set of artefacts. i.e., $c_{out} = \langle\!\langle \mathscr{R},\mathscr{T},\mathscr{E} \rangle\!\rangle (c_{in})$.
- $\mathscr{L}_{\mathscr{L}}$ is the *meta-level ruleset language*, the set of all possible rules that act on rulesets (i.e., on $\mathscr{L}$) the creative system can construct.
- $\mathscr{R}_{\mathscr{L}} \subseteq \mathscr{L}_{\mathscr{L}}$ is a *meta-level constraint ruleset*, by which the system defines the scope of the meta-conceptual space of possible rules that can be part of $\mathscr{L}$.

$\mathscr{T}_{\mathscr{L}} \subseteq \mathscr{L}_{\mathscr{L}}$ is a *meta-level traversal ruleset*, by which the system explores the space of possible rules for $\mathscr{L}$.

$\mathscr{E}_{\mathscr{L}} \subseteq \mathscr{L}_{\mathscr{L}}$ is a *meta-level evaluation ruleset*, by which the system evaluates proposed rulesets for their ability to generate valuable concepts.

The differentiation of $\mathscr{R}$, the rules defining the conceptual space, from $\mathscr{T}$, the rules defining the search process which acts on that space, is a significant addition to Boden's notion of transformational creativity. With this distinction Wiggins can describe two kinds of transformational creativity: $\mathscr{R}$-transformation of the space of possible concepts, and $\mathscr{T}$-transformation of the search process for generating new concepts. $\mathscr{R}$-transformation, closest to Boden's original conceptualisation of transformative creativity, concerns the redefinition of what a creative system considers possible. $\mathscr{T}$-transformation concerns the redefinition of how a creative system creates.

The definition of a creative domain – as captured by Wiggins' $\mathscr{R}$ – is a socially grounded construct. While it is useful from the perspective of defining transformation across a creative domain to think of that construct as stable across all members of a society, in practice this knowledge must be learnt by each member. In Boden's original model, the definition of the creative domain is agreed amongst all participants, and this knowledge is not expected to be constructed through exposure to the domain. Wiggins hints at the social nature of $\mathscr{R}$, but does not distinguish individual and societal transformation of the conceptual space. To model the influence of artefacts created by others on a system's behaviour, we must capture this distinction: we will use $\mathscr{R}$ to refer to an individual creative system's definition of the space, but one could imagine a broader, socially grounded *historical-*$\mathscr{R}$ of the sort Boden describes emerging from the cross-pollination of ideas and norms.

Our intent is to capture specific curiosity – intentional pursuit of further transformation along a search trajectory incited by a particular transformative example – within an expansion of this framework. To achieve this, we need to expand Wiggins' formalisation in four ways:

- To enable a creator to be surprised by its own output, as in Schön (1983), a creative system must externalise and re-perceive its creations as part of the generative process.
- To incorporate the influence of other creators, the input to a creative system's generative process must include all artefacts it has observed, not just its own creations.
- To model the probabilistic nature of expectations, the conceptual space should be a fuzzy set of *probable* concepts, not a crisp set of *possible* concepts.
- To separate unexpectedness from inexplicability, the system should be aware of its confidence in the predicted likelihood of any concept being in the conceptual space.

These changes capture the situated, social, and expectation-based nature of creative systems, allowing us to use Wiggins' formalisation to explore the question of when, where and why transformative creativity occurs.

Surprise as $\mathscr{R}$-transformation

We now formally describe the above expansion of the framework. The literature on design cognition describes how creators can be surprised by their own creations. For this to be possible in an artificial creative system those creations must be represented in a way that contains additional information not used to create them. To reflect this we add a step to the post-generation process of the creative system. First, $\langle\langle .,.,. \rangle\rangle$ is used to generate a new set of outputs, c_{out}, from the current inputs, and then instead of those outputs being directly appended to c_{in} for the next iteration, they are first reified via a function r, which maps from a concept to an externalised representation of that concept which we call an "artefact", and then re-perceived by a function p, which maps from an artefact back to a concept in $\mathscr{U}$. The nature of perception, reification and the space of possible artefacts is beyond the scope of this paper.

To capture a society of creative systems that influence each others' work, we must amend the generative step of Wiggins' formalisation: instead of applying the interpreter $\langle\langle .,.,. \rangle\rangle$ to just c_{in}, the ordered set of that system's own past creations, we must apply it to all an ordered set of all concepts the system has previously observed, regardless of their source. We assume our creative system is part of a society of creative systems that are all producing artefacts within the same domain (by which we mean they share at least $\mathscr{U}$). Each creative system possesses an additional ordered set of concepts, c_{obs} that it has observed but did not create. Different societies may have different structures in which creative systems are exposed to each others' work in more or less selective ways, but c_{obs} is generated by applying the perception function p defined above to some subset of the artefacts externalised by other creative systems. If c_{obs} is non-empty before a creative system has generated any concepts of its own, then those pre-existing known artefacts are the system's *inspiring set* (Ritchie 2001). We can now describe the generation step in our amended formalisation, applying the interpreter to the union of creations and observations, and afterwards reifying and re-perceiving the output, i.e. $c_{out} = p(r(\langle\langle \mathscr{R},\mathscr{T},\mathscr{E} \rangle\rangle(c_{in} \cup c_{obs})))$.

Wiggins suggests that the output of the interpretation function for $\mathscr{R}$ (a real number in $[0,1]$) be converted to a boolean value indicating membership in $\mathscr{C}$. We propose instead that $\mathscr{C}$ be considered a fuzzy set, with the output of the interpreter defining a membership function $l : \mathscr{U} \rightarrow [0,1]$ that indicates the likelihood of observing each concept as part of the domain. This transforms Wiggins' space of *possible* artefacts into a space of *probable* artefacts, and lets us capture all the rich relationships between concepts that influence their mutual likelihoods. We derive this interpretation from our previous work on expectation, novelty and transformation, see Grace and Maher (2014) for details.

We introduce into our framework a notion of confidence. This serves to differentiate unexpectedness (a violation of confident expectations) from ignorance (Ortony and Partridge 1987). To achieve this we replace the $[\![.]\!]$ interpreter from Wiggins with a modified version, $(\!|.|\!)$, which differs only in that it returns a 2-tuple of real numbers in [0,1] for each artefact to which it is applied. The first, as in $[\![.]\!]$, is

the truth value, which becomes the value of the likelihood function l that defines the artefact's membership in the resulting set. The second value is the system's confidence, with 0 indicating a complete lack of confidence and 1 indicating complete certainty. This confidence becomes another function $c : \mathscr{U} \rightarrow [0, 1]$. We use $(\!|.|\!)$ when generating the conceptual space with $\mathscr{R}$, as in:

$$\mathscr{C} = (\!|\mathscr{R}|\!)(\mathscr{U})$$

As a result our $\mathscr{C}$ is a fuzzy set of concepts with a membership function l defining the likelihood of observing each concept in $\mathscr{U}$, as well as a similar confidence function c defining the system's confidence in each artefact's likelihood. These functions are compiled from the first and last elements, respectively, of the tuples output by $(\!|.|\!)$. That is, for each concept $a \in \mathscr{U}$, given $(\!|\mathscr{R}|\!)(\{a\}) = (a_l, a_c)$ and assuming $a_l > 0$:

$$a \in \mathscr{C}, l(a) = a_l, c(a) = a_c$$

From this perspective, Wiggins' $\mathscr{R}$ becomes the creative system's *expectations about the creative domain*. This connection between conceptual space membership and expectation allows us to describe the influence of surprise on creative search. In our amended framework, $\mathscr{R}$-transformation is commonplace and necessary, a natural effect of creative systems acquiring the knowledge they need to competently model the society's rules about the domain through their own experience.

A creative system experiences *expectation failure* when the conceptual representation of a newly observed artefact has a low a-priori likelihood in the conceptual space. We can then distinguish two kinds of artefact that cause expectation failure: inexplicable ones, where the system is not confident of its predicted low likelihood, and unexpected ones, where it is. An unexpected artefact a_u is one for which:

$$a_u \in (c_{in} \cup c_{obs}), l(a_u) \approx 0, c(a_u) \approx 1$$

Complementarily, for an inexplicable artefact a_i:

$$a_i \in (c_{in} \cup c_{obs}), l(a_i) \approx 0, c(a_i) \not\approx 1$$

Only in the first case can we say that the agent's expectations were violated – in the absence of a confident prediction the system was merely ignorant. Both inexplicable and unexpected artefacts should by rights induce a transformation of the domain knowledge in $\mathscr{R}$, as well as potentially transformations of $\mathscr{T}$. Those transformations can be considered a result of creativity if the artefact(s) that caused them are valuable under $\mathscr{E}$. Given our definition of unexpectedness in terms of $\mathscr{R}$ we can restate how our expanded formalisation captures the dyad of novelty and value. The rules in $\mathscr{E}$ will be concerned with the evaluation of artefacts' performance, quality, style, and other components of value, and some portion of $\mathscr{T}$ will use those evaluations to direct search. Contrastingly, some other subset of $\mathscr{T}$ will be concerned with novelty seeking: evaluating the dissimilarity of new artefacts to existing ones using measures of novelty, surprise and transformativity. We refer to this novelty-seeking subset as $\mathscr{T}_n \subset \mathscr{T}$. These latter traversal rules will be based on the likelihoods, confidences, and transformations of $\mathscr{L}$ associated with artefacts.

We do not seek to resolve the disputes surrounding the definitions of novelty, surprise or transformation, only suggesting that $\mathscr{T}_n$ could contain metrics for any or all of those, but we do require that for any creative system $\mathscr{T}_n \neq \emptyset$.

Any artefact valued by both $\mathscr{T}_n$ and $\mathscr{E}$ can be considered p-creative. This generative act is *serendipitous* if the search process possessed no specific intent to create that artefact or anything like it. An artefact discovered to be transformative by $\mathscr{T}_n$ after its creation was not the result of a directed search, for the system cannot know how its knowledge will be transformed by new observations. This places limits on a creative system's ability to generate framing about its creative output: serendipity defies satisfying explanation.

In the next section we use our definitions of inexplicable and unexpected artefacts to describe different possible kinds of transformational creativity. We also propose how a system might adopt constraints on its future generation in response to unexpectedness, and thereby intentionally seek out further unexpected discoveries.

Specific curiosity as a consequence of surprise

A system that has observed inexplicable artefacts will attempt to learn: to improve its (clearly insufficient) knowledge of $\mathscr{U}$. We consider *learning* to be a creative system's response to the inexplicable, and it is our first possible kind of $\mathscr{R}$-transformation. Learning can be expressed as the application of $\mathscr{T}_{\mathscr{L}}$ to produce new $\mathscr{R}$ and/or $\mathscr{T}$ in response to inexplicable artefact(s) in c_{in} or c_{obs}. While the mechanisms of learning will be specific to the rules in $\mathscr{L}_{\mathscr{L}}$, we can describe its effects: it attempts to transform $\mathscr{R}$ such that the likelihood of previously observed artefacts increases.

A system that has observed unexpected artefacts will be surprised. We consider *artefact-induced surprise* to be a creative system's response to unexpected artefacts, and it is our second kind of $\mathscr{R}$-transformation. Artefact-induced surprise can be expressed as the application of $\mathscr{T}_{\mathscr{L}}$ to produce new $\mathscr{R}$ and/or $\mathscr{T}$ in response to unexpected artefact(s) in c_{in} or c_{obs}. Learning occurs from unexpected objects as it does from inexplicable ones, producing $\mathscr{R}$-transformations that increase the expected likelihood of previous observations.

Inspired by cognitive studies of reflection in human designers by Suwa et al (1999) and others we can now consider how surprise might affect a system's future generative behaviour (i.e. cause transformation of $\mathscr{T}$). Specific curiosity, as introduced earlier, is the deliberate pursuit of specific new knowledge or stimuli through the adoption of goals or constraints on behaviour. In the context of a creative system this is $\mathscr{T}$-transformation with the goal of exploring an unexpected stimulus, based on the hypothesis that (as observed in human designers), surprise begets further surprise. This can also be considered a form of active learning (Cohn, Ghahramani, and Jordan 1996), where the system actively tries to fill the gaps in its knowledge through generation.

To become specifically curious about an artefact is to seek to create more artefacts that embody the interesting things about it. We formalise this as follows: given an unexpected

artefact a_u we can determine the subset of rules that contributed to its confident low-likelihood prediction: $\mathscr{R}_{a_u} \subseteq \mathscr{R}$. These rules embody the domain knowledge that was violated by the perception of the new artefact, in that they produced a confident prediction that was proven wrong. This subset forms the basis of the system's specific curiosity, in that the system can use them to pursue artefacts that are unexpected according to just those rules. To define this we induce r, a relevance function over concepts that measures the complement of the likelihood of a concept occurring in a conceptual space defined exclusively by $\mathscr{R}_{a_u}$. Accordingly $r(a) \approx 1$ for any artefact a that would be considered unexpected according to the same rules as was a_u, including a_u itself. Conversely, any artefact that is not unexpected, or is unexpected due to other rules not in $\mathscr{R}_{a_u}$, would produce a lower value of r. We can then define specific curiosity about a_u as replacing $\mathscr{T}_n$ with a single rule that seeks artefacts for which $r(a) \approx 1$. This (temporarily) redirects the system's general (i.e. diversive) search for novel artefacts towards those that are unexpected according to the same rules as the one that caused the surprise.

By constructing a relevance function from the rules violated by the unexpected artefact we focus the system upon the parts of its own knowledge that produced the unexpected result. The results of this specific curiosity will vary based on the structure of the knowledge that was violated. If the rules define boundaries of the domain, the relevance function will value artefacts that break the same boundaries as the focus of curiosity. If the violated rules placed the focus in a new or rare category, the relevance function will value artefacts in that category. If the violated rules define an expected relationship between components of the artefacts' representation, the relevance function will value artefacts that break the same relationship in the same way as the focus. In each case the relevance function will value artefacts that are in some way similar to the one that caused surprise, but with that similarity determined by the system's knowledge.

The hypothesis driving this specific curiosity is that regions of the conceptual space that generate one unexpected artefact likely have the potential to generate more, and searching nearby has a greater chance to yield further unexpected (and therefore potentially creative) artefacts than searching elsewhere in the space. This behaviour aligns with the concept of creative autonomy and situational adaptation of goals described in (Jennings 2010).

In the following section we illustrate the above kinds of $\mathscr{R}$-transformation with examples from the domain of recipe generation.

A worked example of unexpectedness-triggered specific curiosity

As a hypothetical example of our unexpectedness-triggered reformulation approach, consider the creative domain of recipes. Culinary creativity has recently attracted attention in the computational creativity community (Morris et al. 2012; Varshney et al. 2013), and we draw upon it as a way of illustrating our model of specific curiosity.

Assume a hypothetical recipe generation system integrated with a large online recipe repository. The system has access to all the recipes posted by humans, and is tasked with supplementing that database with its own creations. Each recipe is an artefact represented by its ingredients and their quantities, the preparation steps, and metadata such as cooking time and user-applied tags. This is supplemented by behavioural information for each recipe: the full text and ratings of its set of user reviews. The system's task is to generate novel and valuable recipes, and submit them for human consumption and review. $\mathscr{E}$ is based on aggregated user ratings. $\mathscr{R}$ is based on domain knowledge represented by a set of predictive models that describe the likelihood of various combinations of ingredients, quantities, tags, categories, reviews and ratings occuring. We can now describe three ways that this implementation of our framework could encounter transformative creativity.

The first cause of $\mathscr{R}$-transformation is encountering an *inexplicable* recipe. This would be commonplace while the system developed its knowledge about the domain (as the pre-existing human-created recipes that form its inspiring set were added to its database). For example, assume that the system, early into its learning, encountered its first slow-cooked dish. The existing rules in $\mathscr{R}$ would assign a very low a-priori likelihood to a recipe with an eight hour cooking time, but having seen so few previous recipes of any kind it would also assign a low confidence to that prediction. The result would be learning – transforming $\mathscr{R}$ to incorporate the new range of observed cooking times. No surprise or specific curiosity would result – the system's understanding of the conceptual space improved as a result of observing new kinds of artefact that had been produced by others, a necessary and commonplace step of acquiring competency in a creative domain.

The second cause of $\mathscr{R}$-transformation is an *unexpected* recipe. This occurs when the system makes confident predictions of the likelihood of observed recipes, but is still wrong, possibly as the result of a change in the behaviour of the other creative systems in the society (which, in this case, are the human submitters of recipes). Consider what would happen to the system's knowledge about the ingredient "ginger" if its inspiring set (i.e. the recipes in c_{obs} it used to populate $\mathscr{R}$ before generating any artefacts of its own) contained mostly Western recipes, and it developed confident predictions about that ingredient before being exposed to Eastern-inspired recipes. It would confidently expect that ginger was found mostly in sweet baked goods, alongside ingredients like butter, sugar and flour. Encountering a recipe for ginger-and-soy chicken would be highly surprising, causing it to adapt its domain knowledge to fit the new recipe. In this case the creative system had a robust, but incomplete model of the creative domain, and observed an artefact that it would consider p-creative, even though that artefact's creator may have considered it novel.

The third cause of $\mathscr{R}$-transformation is as a result of specific curiosity caused by an earlier surprising recipe. As another example, consider "chicken paprikash", a Hungarian-inspired dish that combines a roux-based sauce with curry-like spices (cumin, paprika and chili). This is an incongruous combination of ingredients and instructions, as the ma-

jority of roux-based sauces are flavoured with herbs, stocks and/or cheeses. Our creative system encounters this recipe, becomes surprised as in the ginger-and-soy chicken example, and uses that surprise to trigger specific curiosity. The rules in $\mathscr{R}$ that confidently assign a low likelihood to a recipe containing both the steps for a roux and the ingredients for a curry are extracted as $\mathscr{R}a_u$. A relevance function is then constructed from those rules that evaluates the degree to which a recipe violates them, and this function replaces the novelty-seeking rules in $\mathscr{T}_n$. The system begins generating recipes that violate these specific rules, such as a roux-based sauce with other unexpected ingredients (such as chocolate), or curries with unusual preparation steps (such as being baked into a pie). The authors feel compelled to mention that they are not chefs, but encourage readers to assume for the sake of argument that those new recipes are both novel and valuable. The observation of these new recipes would lead to additional $\mathscr{R}$-transformation, and this time that transformation can be said to have a deliberate cause. These artefacts were not created serendipitously, they were intentionally generated as the result of a targeted exploration of a specific region of the creative domain, and their discovery further transformed the conceptual space.

Specific curiosity can be triggered both from a creative system's own creations, or from those of the other creative systems within its society (here the human user-base of the recipe website). In the case of the chicken paprikash above, the specific curiosity episode was triggered by the observation of a surprising creative artefact generated by a human – other likely external curiosity-triggers in this domain could include the addition of bacon to sweet foods, the inclusion of leafy greens in smoothies, the rise of a new and novel "superfood", or a seasonally resurgent ingredient.

The creative system could trigger its own specific curiosity episodes by generating recipes that, once reified and reperceived, were considered surprising. Consider, for example, rules in our creative system's $\mathscr{T}$ that use computational analogy-making to map between two recipes and then transfer a new ingredient from the source to the target. An analogy could be constructed between a calzone and an omelette, as both consist of a base layer to which toppings are added before the base is folded over to create a filled final product. The rules for analogical transfer in $\mathscr{T}$ identify that the tomato paste spread on the calzone is missing from the omelette, and create a new recipe in which a tomato sauce is spread over the omelette before folding. This would be considered unexpected by the rules in $\mathscr{R}$ that pertain to omelettes, which would make confident predictions that a tomato-based sauce would be unlikely to be involved in an omelette recipe. The authors again remind the reader that we are definitely not chefs, but let us assume that the resulting sauced omelette was also considered valuable. Specific curiosity about that unexpected combination of ingredients and cooking methods would result in a transformation of $\mathscr{T}_n$ to specifically seek out further recipes involving unusual ingredients being added to omelettes during cooking. Generating new artefacts under this transformed search trajectory could lead to the recipes with further unexpected mid-omelette additions such as spices or fruits. These new creative artefacts would further transform the rules in $\mathscr{R}$ that pertain to omelette creation, and if they were also considered valuable according to $\mathscr{E}$ then they would constitute intentional transformative creativity.

Conclusions

We have described an extension to Wiggins' (2006) framework that captures the notions of unexpectedness, surprise and specific curiosity. This approach is motivated by the need for creative systems that can make autonomous evaluative decisions and exhibit intentional behaviour (Jennings 2010; Saunders 2012). The solution proposed in our framework draws on literature from design cognition which suggests that human creators are not only capable of self-surprise but that it is a significant driver of creative output. Based on this inspiration from cognition we model surprise based on violation of a creative system's learnt model of the conceptual space, and describe specific curiosity behaviours that explore surprising stimuli.

Within our framework we can distinguish three causes of transformational creativity: inexplicable artefacts, unexpected artefacts, and specific curiosity. If found in an artefact that was also valuable the first would not be creative (as the transformation resulted from a lack of sufficient knowledge to make predictions), the second would be serendipitous creativity (as the system stumbled upon it without any deliberate goal), and the last would constitute intentional creativity. Specific curiosity describes the iterative cycle between the $\mathscr{R}$-transformation that occurs when observing or creating an unexpected artefact, the $\mathscr{T}$-transformation that facilitates the resulting search for more, similarly surprising artefacts, and the resulting $\mathscr{R}$-transformation that heralds the success of that deliberate search. Our future work involves the development of systems like the one presented here as an example: creative machines capable of surprise, specific curiosity and autonomous intent.

References

Baldi, P., and Itti, L. 2010. Of bits and wows: a bayesian theory of surprise with applications to attention. *Neural Networks* 23(5):649–666.

Berlyne, D. E. 1966. Curiosity and exploration. *Science* 153(3731):25–33.

Best, D. 1981. Intentionality and art. *Philosophy* 56(217):349–363.

Boden, M. A. 2003. *The creative mind: Myths and mechanisms*. Routledge.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proceedings of the Third International Conference on Computational Creativity*, 77–82.

Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research*.

Cross, N. 2004. Expertise in design: an overview. *Design studies* 25(5):427–441.

Cully, A.; Clune, J.; and Mouret, J.-B. 2014. Robots that can adapt like natural animals. *arXiv preprint arXiv:1407.3501*.

Dewey, J. 2005. *Art as experience*. Penguin.

Gebser, M.; Kaufmann, B.; and Schaub, T. 2009. Solution enumeration for projected boolean search problems. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. Springer. 71–86.

Getzels, J. W., and Csikszentmihalyi, M. 1976. *The creative vision: A longitudinal study of problem finding in art*. Wiley New York.

Goldschmidt, G. 1991. The dialectics of sketching. *Creativity research journal* 4(2):123–143.

Grace, K., and Maher, M. L. 2014. What to expect when youre expecting: the role of unexpectedness in computationally evaluating creativity. In *Proceedings of the 4th International Conference on Computational Creativity, to appear*.

Grace, K.; Maher, M.; Fisher, D.; and Brady, K. 2014. A data-intensive approach to predicting creative designs based on novelty, value and surprise. *International Journal of Design Creativity and Innovation* (Ahead of Print).

Grace, K.; Gero, J.; and Saunders, R. 2015. Interpretation-driven mapping: a framework for conducting search and re-representation in parallel for computational analogy in design. *AI EDAM* 29(2):185–201.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kelly, N., and Gero, J. S. 2014. Interpretation in design: modelling how the situation changes during design activity. *Research in Engineering Design* 25(2):109–124.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation-with intent. In *Proceedings of the 1st international conference on computational creativity*, 20.

Lawson, B. 2006. *How designers think: the design process demystified*. Routledge.

Liapis, A.; Martınez, H. P.; Togelius, J.; and Yannakakis, G. N. 2013. Transforming exploratory creativity with delenox. In *Proceedings of the Fourth International Conference on Computational Creativity*, 56–63.

Macedo, L., and Cardoso, A. 2001. Modeling forms of surprise in an artificial agent. *Structure* 1(C2):C3.

Merrick, K., and Maher, M. 2009. Motivated reinforcement learing.

Morris, R. G.; Burton, S. H.; Bodily, P. M.; and Ventura, D. 2012. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *Proceedings of the 3rd International Conference on Computational Creativity*, 119–125.

Newell, A.; Shaw, J.; and Simon, H. A. 1959. *The processes of creative thinking*. Rand Corporation.

Ortony, A., and Partridge, D. 1987. Surprisingness and expectation failure: what's the difference? In *Proceedings of the 10th international joint conference on Artificial intelligence-Volume 1*, 106–108. Morgan Kaufmann Publishers Inc.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB01 Symposium*.

Saunders, R., and Gero, J. S. 2001. Artificial creativity: A synthetic approach to the study of creative behaviour. *Computational and Cognitive Models of Creative Design V, Key Centre of Design Computing and Cognition, University of Sydney, Sydney* 113–139.

Saunders, R. 2011. Artificial creative systems and the evolution of language. In *Proceedings of the Second International Conference on Computational Creativity*, 36–41.

Saunders, R. 2012. Towards autonomous creative systems: A computational approach. *Cognitive Computation* 4(3):216–225.

Schmidhuber, J. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *Autonomous Mental Development, IEEE Transactions on* 2(3):230–247.

Schon, D. A., and Wiggins, G. 1992. Kinds of seeing and their functions in designing. *Design studies* 13(2):135–156.

Schön, D. A. 1983. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books.

Smith, A. M., and Mateas, M. 2011. Knowledge-level creativity in game design. In *Proc. of the 2nd International Conference in Computational Creativity (ICCC 2011)*.

Sternberg, R. J., and Lubart, T. I. 1999. The concept of creativity: Prospects and paradigms. *Handbook of creativity* 1:3–15.

Suwa, M.; Gero, J.; and Purcell, T. 1999. Unexpected discoveries and s-inventions of design requirements: A key to creative designs. *Computational Models of Creative Design IV, Key Centre of Design Computing and Cognition, University of Sydney, Sydney, Australia* 297–320.

Varshney, L. R.; Pinel, F.; Varshney, K. R.; Bhattacharjya, D.; Schoergendorfer, A.; and Chee, Y.-M. 2013. A big data approach to computational creativity. *arXiv preprint arXiv:1311.1213*.

Weisberg, R. W. 1993. *Creativity: Beyond the myth of genius*. WH Freeman New York.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.