Bayesian Modelling of the Well-Made Surprise

Patrick Chieppe, Penny Sweetser and **Eryn Newman**
Australian National University
Canberra ACT 2600 Australia
patrick.chieppe@anu.edu.au, penny.kyburz@anu.edu.au, eryn.newman@anu.edu.au

Abstract

The "well-made" surprise is a narrative pattern of setting up and executing a surprise in a way that is generally perceived as enjoyable and rewarding. It leverages biases in human cognition to manipulate the audience's state of belief, and is commonly found in western culture as early as Aristotle's Poetics. We propose a novel framework to model the audience's beliefs of a narrative world using approximate Bayesian inference over Markov Logic Networks. We operationalise three qualitative attributes of the well-made surprise (consistency, divergence and certainty) as quantitative functions of the outputs of inference. This work follows the paradigm from computational narrative of operationalising qualitative concepts from literary theory in order to model and generate narratives, either autonomously or cooperatively with a human author. We demonstrate the proposed framework on ten short narratives, and test it with a study on 91 participants. We find that for consistency and divergence, a change in the model's prediction corresponds with a significant change in the participants' rating. Our results suggest that the proposed framework may have meaningful predictive power and potential for future applications to narrative generation, plot analysis, and computer-aided creativity.

Introduction

Computational narrative is a long-standing research field focusing on modelling the building blocks of a narrative in a machine-processable structure, most often with the goal of analysing existing narratives or generating novel ones (Kybartas and Bidarra 2017; Valls-Vargas, Zhu, and Ontanon 2017). A common paradigm in generational narrative is to apply some existing algorithmic framework to an operationalisation of a concept from literary theory, such as suspense (Cheong and Young 2008), surprise (Bae and Young 2014) and conflict (Ware 2002). In contrast, other recent examples in the field build upon natural language processing advances using neural networks (Radford et al. 2019) to directly process and generate natural language narratives (Yao et al. 2019; Fan, Lewis, and Dauphin 2018) and have gained public popularity, for instance with the indie game AI Dungeon (Hua and Raley 2020). These rely on the ready availability of large datasets rather than human-encoded models, solving scalability issues but losing transparency and decodability of the model's internal workings in doing so.

Tobin (2018) describes the well-made surprise as a common pattern in western narratives, dating as far back as the well-made tragedy in Aristotle's Poetics, from which they borrow the term. It describes a surprise, or an unexpected event or occurrence in a narrative, that is accompanied by an experience of insight, also called an "Aha" experience (Topolinski and Reber 2010; Skaar and Reber 2020). In this particular scenario, it is the formation of a new understanding of the narrative, associated with suddenness, ease and fluency of processing, certainty in the new understanding and positive affect, which lead to overall enjoyment of the surprise.

Tobin's theory of the well-made surprise deals primarily with literature and film, but it's a construct extensible to any storytelling medium. Thus, for the remainder of this paper we will not assume any particular medium when referring to an author (the person or people crafting the narrative and the surprise), an audience (the person or people experiencing the narrative) and a text (the artifact through which the narrative is conveyed from the author to the audience).

Tobin also details a variety of techniques by which an author may construct a well-made surprise, largely leveraging common biases in human cognition (Evans 1989) to deliberately construct misunderstandings or misinterpretations about the narrative leading up to the surprise, while still enabling the "correct" meaning of the text to be recognised and accepted in retrospect. Tobin argues that it is especially because well-made surprises exploit these biases that they produce an experience of insight. Examples include minimising the audience's attention towards certain information (Emmott and Alexander 2014) and shifting the frame of reference from an objective telling of the events to a character's impression of them using presupposition clauses to mask falsehood as truth (Loftus and Zanni 1975; Bredart and Modolo 1988). Many of these details of the techniques are specific to a medium, but in general, their intended effect is to manipulate how the audience processes information and builds a mental model of the narrative, steering them towards a state in which later events can best deliver a satisfying, insightful surprise.

In this work, we investigate the applicability of Tobin's theory as a modelling tool in the field of computational creativity. We identify three main areas of applications for such a model:

- *Narrative analysis*: Improving the understanding of existing narratives by providing a new analytical lens through which to model surprises (Valls-Vargas, Zhu, and Ontanon 2017).
- *Computer-aided creativity*: Aiding authors in the process of writing a satisfying plot by identifying features such as plot holes and well-made surprises (Kapadia et al. 2015), similarly to how formal methods in software modelling can aid software developers verify their abstractions (Guttag and Horning 1980).
- *Generative narrative evaluation*: Evaluating the output of other narrative generation tools, for example as a search heuristic or as a validation metric. Tobin (2018, pp. 54-55) highlights that the pattern of the well-made surprise is often pleasant to experience even when familiar with it, which is a very attractive property for narrative generation, which often struggles with overcoming repetitiveness (Alabdulkarim, Li, and Peng 2021).

We believe that there is unexplored potential in the computational modelling of the theory of the well-made surprise. There exists significant work in modelling surprise and other related concepts in computational narrative (Bae and Young 2014; Cheong and Young 2008), as well extensive study into the properties of the "Aha" experience in cognitive psychology and neuroscience (Skaar 2019; Skaar and Reber 2020; Chu and MacGregor 2011), and the theory of the well-made surprise points out an important link between surprise and insight. However, no previous work that we are aware of has attempted to bring all of the above together.

Tobin's work bridges the disciplines of narrative theory and cognitive psychology, and in addition doesn't require deep familiarity with either field to understand. We combine this with the approach from computational narrative to operationalise literary theory and cognitive psychology concepts in an effort to bring computer science into the mix and take the first step towards a novel cross-disciplinary computational model of surprise.

We study the mental model that the audience builds of the narrative through the theory of Bayesian probability (Cox 1946; Jaynes, Jaynes, and Bretthorst 2003), focusing on their knowledge or beliefs about the story world and the inferences they perform throughout the narrative (McKoon and Ratcliff 1992; Graesser, Singer, and Trabasso 1994). From such a model, we aim to operationalise key qualities of the well-made surprise by expressing them as functions of the model, based on related ideas from logic and information theory. We implement a probabilistic framework of the well-made surprise and implement three such operationalisations, which we evaluate on a study with 91 participants. We find that the model's predictions agree with participant ratings for two of the three operationalisations, and identify several strengths and weaknesses of the proposed framework.

Background

The Bayesian theory of probability (Cox 1946; Jaynes, Jaynes, and Bretthorst 2003) has seen extensive use in computer science as the theoretical basis for probabilistic models (Russell and Norvig 2010, pp. 510-546), as well as applications in both cognitive sciences (Griffiths, Kemp, and Tenenbaum 2008, pp. 85-138) and literary theory (Kukkonen 2014). Under the Bayesian framework, probabilities represent a degree of belief in a hypothesis, with $P(x) = 1$ representing a certain fact, and $P(x) = 0$ an impossibility. As new data is acquired, existing beliefs are updated using Bayes' theorem. In the context of experiencing a narrative, new beliefs are added to the model as needed in order to make sense of the narrative (McKoon and Ratcliff 1992; Graesser, Singer, and Trabasso 1994), and the resulting model is a combination of information that the narrative has provided and of the audience's own background knowledge.

A Markov Logic Network or MLN (Richardson and Domingos 2006) is a probabilistic model that encodes a joint probability distribution over the truth value of all ground atoms in its domain. Like other Bayesian models it allows for inference, or computing the probability $P(A|B)$ for some A to be true given some known prior B, both being logic formulas. While exact inference is intractable in the general case, efficient approximate inference algorithms have been developed (Riedel 2005; Geier and Biundo 2011; Niu et al. 2012; Van den Broeck 2013).

A MLN is defined as a set of weighted first-order logic statements defined over a finite domain. MLNs afford the expressive power of first-order logic alongside the ability to model uncertainty, both in the sense of information that is varying degrees of plausible rather than absolutely true or false, and in the sense of contradictory information. We find these to be valuable properties in the modelling of the well-made surprise. Partial, uncertain and contradictory information is extremely common in surprising narratives, and the ability to reason about such imperfect information is an important part of understanding well-made surprises, where an initially unexpected outcome is made sense of and obvious in hindsight. In addition, MLNs' expressive power proves especially useful due to the exploratory nature of this work, allowing a wide range of concepts to be modelled.

MLNs can be seen as a template from which a ground Markov Random Field or MRF (Murphy 2012) can be built. The ground MRF is a bipartite graph of all ground atoms and all groundings of all rules, where an edge exists between an atom and a rule if the atom appears in the rule. This interpretation is especially useful for visualising the structure of a MLN and the flow of information during inference, as shown later in this paper.

Literature review

In the field of computational narrative, there are many examples of systems designed to generate stories guided by some operationalised narrative concept. Bae and Young (2014) use AI planning to model flashbacks and foreshadowing in order to construct surprising narratives with explainable causes, and present a methodology to adapt their model to narrative analysis of surprises. Arinbjarnar (2005; 2008) propose an interactive murder mystery plot generation engine based on Bayesian networks which combines ideas from Propp's (1968) morphology of the Russian folktale with accepted genre conventions from mystery writers. Riedl and Bulitko (2012) and Arinbjarnar, Barber, and Kudenko (2009) survey a large body of work on interactive narratives.

Bayesian methods and especially Bayesian networks have seen extensive use in the modelling of uncertainty and knowledge, on both real and fictional narratives and on a very wide variety of topics. These include evidence in legal cases (Vlek et al. 2013), workplace injury narrative coding (Lehto, Marucci-Wellman, and Corns 2009; Measure 2014; Taylor et al. 2014), the visual perception of surprising events while watching television (Itti and Baldi 2009) and how emotion appraisals are transmitted across retellings of a story (Breithaupt, Li, and Kruschke 2022). There are many more examples, see Canaj, Biba, and Kote (2018) for a more thorough survey. Skaar (2019) studies in detail several aspects of the "Aha" experience using Bayesian statistics.

While Markov Logic Networks are less prominent in the literature than Bayesian networks, they have seen several successful applications. Singla and Mooney (2011) train a MLN of a plan from observed actions, Ohwatari et al. (2014) model interpersonal relationships between characters and Patil et al. (2018) use MLNs to identify characters with multiple referring aliases.

Applications to interactive narratives are especially relevant to our research, as the algorithmic infrastructure driving the telling of an interactive narrative can naturally start closer to the world of logic and Bayesian modelling than more traditional media, potentially allowing for a smoother and more direct modelling process. Rowe and Lester (2010) modelled user knowledge in an interactive narrative using dynamic Bayesian networks, while Ha et al. (2012) apply MLN structure learning to their user's goals based on the actions they take in a narrative world.

Qualities of the well-made surprise

Tobin (2018, Chapter 5) identifies several qualities that define the "Aha" experience, and by extension the well-made surprise, which we elaborate on and adapt to our approach in the following sections. Their work focuses on four qualities that are required for an experience of realisation to be the basis of a well-made surprise (suddenness, certainty/confidence, ease/fluency and pleasure/enjoyment). We specify and formalise an additional three which are based on our interpretation of concepts Tobin alludes to throughout their work (coherence, consistency, and divergence).

Coherence

Coherence is a measure of the logical flow in the surprise. An incoherent surprise is unrelated to the rest of the narrative and is confusing even in hindsight. "Cheap" twist endings (Marie-Laure Ryan 2009) often fall into this category, failing to justify their existence in the story world beyond resolving a plot element or dismissing a contradiction (e.g. "it was all a dream", so it doesn't have to make sense).

Consistency

Consistency is the degree to which to which the surprise is compatible with the rest of the story leading up to it. A consistent surprise is plausible given all of the information presented by the story thus far, and the audience is able to integrate it into their understanding of the story world without any unexplainable contradictions emerging. Stories often uphold this by masking contradictions behind a character's subjective impression of events, reframing what originally appeared as factual to be misguided, misunderstood or fabricated.

Divergence

Divergence is the magnitude of the knowledge revision caused by the reveal. A divergent surprise will have deeper, further reaching implications in the plot, and force the audience to revise their understanding of earlier events. This extends the notion of how surprising any single event is (i.e. its probability) with the outcome of the inferences that the new information triggers in the audience.

Suddenness

Suddenness is the speed at which the audience arrives at a new understanding after their previous one is revised. A sudden surprise will cause the audience to revise their understanding and adopt a new one within a short span of time.

Inevitability

Inevitability is the degree to which the final understanding is intuitive, satisfying and (in hindsight) obvious, compared to the initial understanding. This can take on a variety of shapes, such as a character's actions being reframed to be more in line with their motivations, or a previously unimportant detail (a "Chekhov's gun") gaining new meaning. This is closely related to Tobin's ease/fluency concept, but we adopt the term "inevitability" from other parts of their work. We chose this name to focus on the knowledge and reasoning side of the concept (a surprise that can be explained and reasoned about into a likely occurrence in hindsight), rather than the psychological idea of cognitive fluency (the quality of thoughts that are easy to mentally process), although the latter would be an interesting avenue for future research (Oppenheimer 2008).

Certainty

Certainty is the degree to which the new understanding appears as certain and undoubtable, naturally fitting into the story world in such a way that it answers questions and fills gaps in knowledge besides the subject of the surprise.

Enjoyment

When the other qualities hold, we expect the surprise to be enjoyable. Due to the highly subjective nature of the experience, there is a fine line between accepting the surprise as insightful and rejecting it as a cheap writing trick. This becomes more evident the more ambitious the surprise is at unravelling the audience's previous understanding. For

	Sentence	Encoding
1	Katie just had a very long week at work.	$WorkHard(Katie, Weekdays)$
2	*One cannot work hard without working.*	$\neg DoWork(x,t) \rightarrow \neg WorkHard(x,t)$
3	*One is working if they are working hard.*	$WorkHard(x,t) \rightarrow DoWork(x,t)$
4	She couldn't wait for the weekend, she had made plans to relax and watch her favorite tv series.	$WantToWatchShows(Katie, Saturday)$
5	*Being denied a wish can make someone unhappy*	$WantToWatchShows(x,t) \land \neg WatchShows(x,t) \rightarrow Unhappy(x,t)$
6	As Saturday morning rolled around, she woke up to a call from her boss.	$Call(Boss, Katie, Saturday)$
7	He asked her if she could come over to work.	$AskToAtWork(Boss, Katie, Saturday)$
8	*One wouldn't go to work on a Saturday unless their boss asked.*	$\neg AskToAtWork(Boss, y, Saturday) \rightarrow \neg AtWork(y, Saturday)$
9	*If one goes to work, it's to do work.*	$AtWork(x,t) \rightarrow DoWork(x,t)$
10	*Katie couldn't work and watch her shows at the same time.*	$\neg\,(DoWork(Katie,t) \land WatchShows(Katie,t))$
11	*One cannot be happy and unhappy.*	$\neg\,(Happy(x,t) \land Unhappy(x,t))$
12	**She happily agreed and had a great time.**	$AtWork(Katie, Saturday) \land Happy(Katie, Saturday)$
13	Her boss had noticed how hard everyone worked last week, and threw a party at the office.	$(WorkHard(x, Weekdays) \land AtWork(x, Saturday)) \rightarrow Party(x, Saturday)$
14	*One doesn't party and do work.*	$\neg\,(Party(x,y) \land DoWork(x,y))$
15	*Surprise parties make people happy*	$Party(x,t) \rightarrow Happy(x,t)$

Table 1: Example encoding of a story. The reveal is in bold. Background rules are in italics.

instance, surprises relying on unreliable narrators that completely change the perspective of the story from a factual retelling of events to the fallible perception and interpretation of a character can have a polarising effect on their audience.

Proposed model

We view a well-made surprise as composed of three phases: setup, reveal and explanation. During the setup, the audience forms an understanding of the story world, which we call the *flawed* understanding. Then, the reveal is a sudden, surprising event which prompts the audience to question the flawed understanding, and begin forming a new one. The explanation is a final, optional phase in which the story guides the audience towards an improved understanding of the story world, which we call the *truth* understanding.

We model each story as a pair of MLNs, corresponding to the flawed and truth understandings. To demonstrate our modelling process, we wrote ten short stories of four to six sentences each, each story focusing on one of the modelled qualities. For each story, we wrote two variants, one being high in the associated quality, and one low. We wrote the stories such that the difference between the two variants is as minimal as possible to produce the desired difference in the associated quality, while also producing as small a difference as possible in all the other qualities. During the writing process, we categorised stories as high or low in each quality using our subjective judgement.

For each story, we identify one sentence as the reveal, every sentence preceding it as the setup, and every sentence (if any) following it as the explanation. We then encoded each sentence as one or more rules, which are either encoding information stated explicitly in the story, or background knowledge that we assume the audience will draw from in order to make sense of the sentence. Rules and atoms are shared by both the flawed and truth models where possible, as we will later define functions over shared atoms.

See Table 1 for an example encoding of a story written to have high certainty. In the story, Katie is hoping to have a relaxing weekend (4) but is suddenly asked to come to work (7). The audience might expect her to either not abide the request ($\neg AtWork(Katie, Saturday)$), or to begrudgingly do so and be unhappy with the result (due to 5, 9 and 10). The reveal (12) is unexpected because neither holds (due to 11), and is then explained by referring back to the fact that Katie worked hard during the week (1).

We run approximate inference over both the truth and flawed models, using the Alchemy 2 implementation of MLNs (Kok and Domingos 2005) with the default MaxWalkSat and Gibbs sampling approximate inference algorithm described by Richardson and Domingos (2006).

Atoms

1	$DoWork(Katie, Weekdays)$
2	$WorkHard(Katie, Weekdays)$
3	$AskToGoToWork(Boss, Katie, Saturday)$
4	$WatchShows(Katie, Saturday)$
5	$WantToWatchShows(Katie, Saturday)$
6	$DoWork(Katie, Saturday)$
7	$AtWork(Katie, Saturday)$
8	$Happy(Katie, Saturday)$
9	$Unhappy(Katie, Saturday)$
10	$WorkHard(Katie, Saturday)$
11	$Party(Katie, Saturday)$

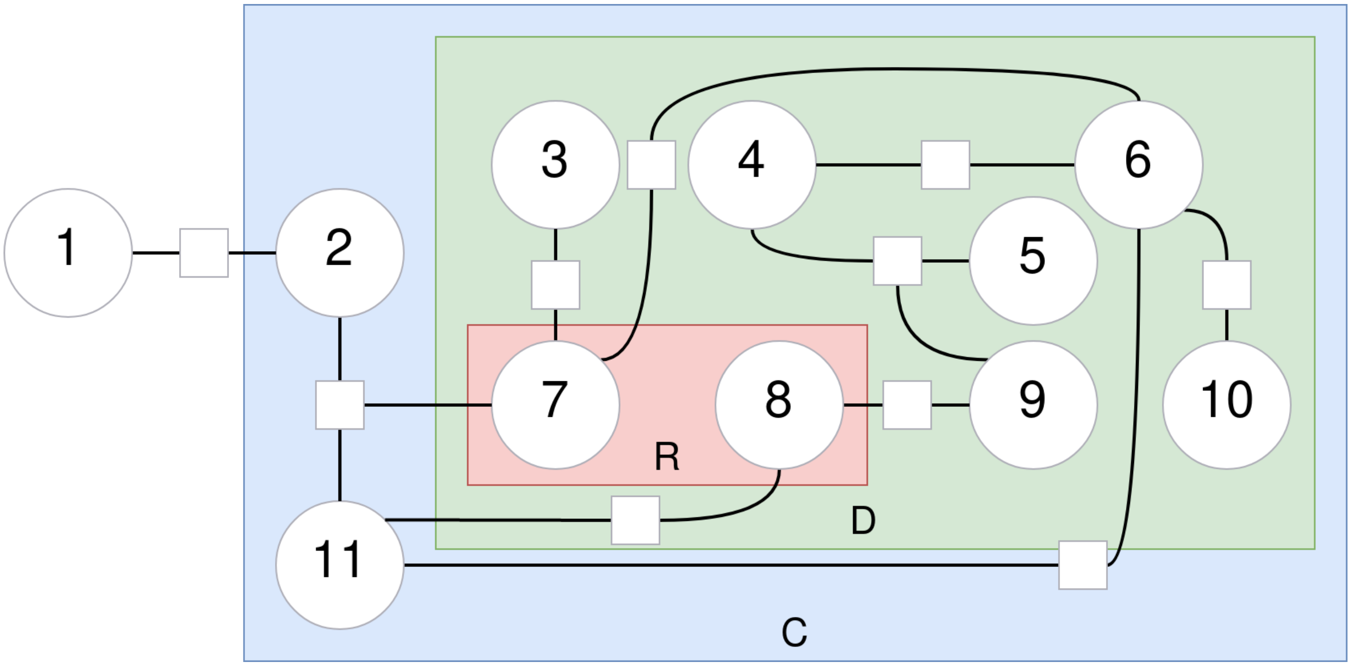

Figure 1: Partial example ground network with reveal (R), divergence blanket (D) and certainty blanket (C) highlighted. Circles are atoms, squares are rules.

This process approximates sampling from the joint probability of all possible worlds defined by the model, and for each ground atom and rule it keeps track of the number of sampled worlds in which they are true. $P(x)$ is then the output probability of x, defined as the proportion of worlds in which x is true among all sampled worlds. We define $P_f(x)$ and $P_t(x)$ to be $P(x)$ in the flawed and truth model respectively.

Operationalisations

From the qualities described earlier in the paper, we operationalise three: consistency, divergence and certainty. Partial work for the operationalisation of coherence was completed, but was not included in the final model. Similarly, inevitability and suddenness are out of the scope of the current model and analysis. Enjoyment involves many subjective factors to the point that it cannot be expressed simply in terms of knowledge and belief, and is not modelled in this work.

Consistency

For the scope of this work, we limit our analysis to instances in which facts directly stated by the narrative contradict each other, and we operationalise this quality as a satisfiability check of the conjunction of all the hard rules in the truth model.

Divergence

We follow an approach similar to Itti and Baldi's (2009) "wow" unit of surprise to quantify the total amount of surprise generated by the reveal.

$$\frac{1}{|D|}\sum_{x \in D} KL(P_t(x) \| P_f(x)) \tag{1}$$

Where KL is the Kullback-Leibler divergence (Kullback and Leibler 1951). We define the divergence blanket D as the set of atoms in common to the flawed and truth ground networks and that are conditionally dependent on the reveal, conditioned on all atoms with known value. In other words, D is the set of all ground atoms that can be reached starting from the reveal, traversing any ground rule edge, and stopping at any atom with known value. D captures the notion of the chain of reasoning that the audience performs to predict the reveal. It aims to capture not only how surprising the reveal is, but also how much this surprise flows backwards through logical links and prompts revision of previously believed information.

Statement	Value
Divergence	
The story is surprising	Positive
The story is not surprising	Negative
The story is not predictable	Positive
The story is predictable	Negative
Consistency	
The story doesn't contradict itself	Positive
The story contradicts itself	Negative
The story made sense	Positive
The story didn't make sense	Negative
Certainty	
The ending is satisfying	Positive
The ending is not satisfying	Negative
The surprise doesn't feel cheap	Positive
The surprise feels cheap	Negative

Table 2: Evaluation statements

Answer	Positive	Negative
Strongly disagree	-1.0	1.0
Somewhat disagree	-0.5	0.5
Neither agree nor disagree	0.0	0.0
Somewhat agree	0.5	-0.5
Strongly agree	1.0	-1.0

Table 3: Likert scale conversion key

It should be noted that $KL(P(x) \| Q(x))$ is not defined when $P(x) = 0$ and $Q(x) \neq 0$, but since we encode hard rules as an arbitrarily large weight rather than an actually infinite weight for computation reasons, output probabilities are never exactly 0 or 1. This has a similar effect to adding a small prior to all probabilities.

Certainty

Shannon entropy (Shannon 1948) is a commonly used measure of uncertainty, and we track its overall change across all modelled information when transitioning from the flawed to the truth model.

$$\frac{1}{|C|} \sum_{x \in C} H(P_f(x)) - H(P_t(x)) \qquad (2)$$

Where H is the Shannon entropy. We define the certainty blanket C as the set of atoms in common to the flawed and truth ground networks and that are conditionally dependent on the reveal in either the flawed or truth ground network, conditioned on all atoms with known value. This is defined similarly to D, but note that $C \subseteq D$, as it includes new information that the flawed interpretation had no knowledge of (the audience hadn't thought of it), but that is still relevant to reasoning about the reveal in retrospect. In Figure 1, $\mathit{WorkHard}(\mathit{Katie}, \mathit{Weekdays})$ is not in D since before knowing about the surprise party, Katie's hard work during the week only relates with her desire for a restful weekend. The same atom is in C since it's used in the explanation. $\mathit{DoWork}(\mathit{Katie}, \mathit{Weekdays})$ is in neither, as any rules leading from it to the reveal first go through known atoms.

Evaluation

We evaluate our framework with an exploratory study on a small set of stories, with a total of 91 undergraduate participants recruited through the Australian National University's School of Psychology's Research Participation Scheme.

In this evaluation, we used a fully within-subjects design, focusing on factors that made it into the final framework. We had a 3 (quality: consistency, divergence, certainty) by 2 (variant level: low, high) design. While participants read a total of 10 stories, we focused our analysis on only 7 of them, as 3 stories focused on an operationalisation of coherence that was not included in the final framework. For each quality we operationalised, we used multiple stories to eliminate item specific effects—that is, participants read 3 different stories varying in consistency (low, high). For each of the 7 stories, participants read a version that was high or low on a target quality. For example, for consistency, participants saw a total of 3 stories, in a high and low level of consistency. See supplemental materials for how consistency, divergence and certainty were manipulated as high or low across each story. Each participant was shown all 14 story variants in random order, subject to the restriction that the two variants (low, high) of the same story were always presented one after the other, again in random order. After reading each of the 14 total story versions, participants were asked to evaluate each story across ratings presented in Table 2, as well as rating comprehension of each story. All ratings were answered on a 5-point Likert scale, which were converted according to the key in Table 3 and averaged for each quality. Relatively more positive values as displayed in Table 4 indicate higher ratings of the key dependant variable (e.g. consistency). Answers associated with low comprehension scores (< 0.25) were filtered out as outliers, but the same significant pattern of results is found with those outliers included. Note that in Table 4, we limit our analysis to the key dependant variable of interest—for stories where we varied consistency, we focus our analysis on consistency as per Table 4. Note that other values may be of interest for further analysis, such as interactions between qualities.

Results

For consistency and divergence, a paired-samples two-tailed t-test showed significant change in the mean of participant answers between the two variants of a story ($p < 0.001$), in the same direction as predicted by the model. For certainty, the answers showed less marked change in one of the stories ($p = 0.077$).

Discussion

The results suggest that the framework has meaningful predictive power for the modelled qualities, and in general the approach of using information theoretical functions to model the well-made surprise shows promise.

Title	Variant	Predicted	Mean	Std dev	t-value	p-value
		Consistency				
The Macaroni	Low	0	−0.5443	0.3968	−7.859	< 0.001
	High	1	−0.0016	0.5449		
Sarah's Walk	Low	0	−0.6094	0.4857	−13.762	< 0.001
	High	1	0.5688	0.522		
Catherine at the Beach	Low	0	0.0111	0.5935	−4.902	< 0.001
	High	1	0.3436	0.4554		
		Divergence				
Emma's Move	Low	0.0083	−0.4253	0.3502	−12.389	< 0.001
	High	1.7592	0.3287	0.3804		
Jimmy and the Candy	Low	0.1484	−0.3657	0.3831	−14.297	< 0.001
	High	0.9032	0.4645	0.3101		
		Certainty				
Katie's Weekend	Low	0.0985	−0.3224	0.3669	−6.347	< 0.001
	High	0.1094	−0.0015	0.4052		
Peter Plays Pool	Low	0	−0.2214	0.3241	−1.792	0.077
	High	0.1425	−0.1296	0.3244		

Table 4: Results of model evaluation. Note that the predicted values are not in the same units as the collected data.

The lack of a common unit of measure between model output and collected data makes it difficult to quantify its precision, and a methodology for normalising model outputs to an accepted scale would greatly improve its verifiability.

The result for the last story under certainty ("Peter Plays Pool") may be partially explained by the questions for certainty being very vague statements about the quality of the surprise and of the insight experience, so more specific questions might yield more useful results. This result still highlights how subjective the overall quality of a surprise can be, even for a very short narrative.

Future Work

Future studies should explore further generalisation of the current findings to more general categories of narratives, especially longer narratives and existing corpora of real-world narratives containing well-made surprises. Tobin (2018) touches upon many literary examples throughout their work which future research should strive towards being able to model. The design of future studies should also take into account the ability to generalise across items. Our study's design manipulated each story in an unique way, largely limiting analysis to individual story variation pairs. These questions could also be examined in a between-subjects design, where people do not have the relative comparison across story versions. These are fruitful avenues for future research.

The weakest part of the current framework is matching a model to a narrative. Due to the high flexibility of MLNs, any narrative (even very short ones) can have a wide range of subjective encodings, and two very similar models may produce different outputs. This is a general criticism often raised towards Bayesian modelling in cognitive sciences (Marcus and Davis 2013; Tauber et al. 2017), and is a consequence of the combination of model flexibility, subjective human modelling, and outputs that are sensitive to model formulation. Model consensus procedures such as those used by Trabasso and Sperry (1985) should be used by future research using hand-written models. Another option is to pursue model extraction from questionnaires (Graesser, Robertson, and Anderson 1981). The approach is still inherently not scalable in the context of open narrative generation, and is likely better suited to aid in narrative analysis or as a computer-aided writing tool.

As an alternative to hand-written models, this flexibility also means that MLNs' modelling language subsume many existing structured representations of narratives, and we suggest the development of conversion procedures from existing narrative models to the proposed framework. In particular, the operationalised qualities may find use as heuristics to evaluate the output of generative models which learn their domain representation from existing data (Li et al. 2013) or publicly available corpora (Guan, Wang, and Huang 2019; Swanson and Gordon 2012). Conversely, existing generative frameworks could be adapted to produce narrative variations suitable for use in future studies (Porteous et al. 2010; Piacenza et al. 2011).

It may be possible to extend consistency to a continuous quantity by adapting a MLN weight learning algorithm, such as the voted perceptron (Richardson and Domingos 2006) or the scaled conjugate gradient (Lowd and Domingos 2007). Since MLN weight training is based around computing the optimal weights for each rule given a dataset, we may be able to learn new weights on the samples obtained from inference. Intuitively, conflicting information will cause the respective rules to be false more often than their original weights would imply, and thus result a lower trained weight.

Divergence and certainty are defined over a subset of the marginals, which varies in size depending on model formulation and verbosity. Furthermore, atoms are included in the respective blankets if any rule links to them, with no regard for how important each atom is in the inference process. Future work could draw from research into recall and importance of events (Trabasso and Sperry 1985; Trabasso and van den Broek 1985) to improve them.

The other qualities that haven't been operationalised yet (coherence, suddenness) should also be investigated and modelled in future work. Some, like inevitability, may benefit from being further decomposed into constituent parts in order to be more easily modelled.

Conclusions

We presented a novel cross-disciplinary modelling framework for the well-made surprise. The proposed framework takes the first step in a cross-disciplinary effort to bring the literary theory of the well-made surprise into the world of computer science, drawing from the field of cognitive science along the way to inform the design of the models and research direction. We believe the framework to have potential in the field of computational narrative and creativity, and identify three main areas of promising application as narrative analysis, computer-aided creativity and generative narrative evaluation. We supported our claims with a pilot study, and examined ways in which the framework may be improved and further developed.

Author Contributions

Patrick Chieppe was responsible for writing the manuscript and all other research work not otherwise attributed to the other authors below.

Penny Sweetser advised on the overall course of research and writing, provided frequent feedback and considerably helped shape the direction of this work.

Eryn Newman contributed to the evaluation design and provided initial feedback on the manuscript.

Acknowledgements

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

Supplementary material

Supplementary material including the stories, models, survey and dataset are available at:

`https://github.com/Palladinium/iccc22`

References

Alabdulkarim, A.; Li, S.; and Peng, X. 2021. Automatic Story Generation: Challenges and Attempts. In *Proceedings of the Third Workshop on Narrative Understanding*, 72–83. Stroudsburg, PA, USA: Association for Computational Linguistics.

Arinbjarnar, M.; Barber, H.; and Kudenko, D. 2009. A critical review of interactive drama systems. In *AISB 2009 Symposium. AI and Games.*

Arinbjarnar, M. 2005. *Murder She Programmed: Dynamic Plot Generating Engine for Murder Mystery Games*. Ph.D. Dissertation, Reykjavík University.

Arinbjarnar, M. 2008. Dynamic Plot Generating Engine. *Proceedings of the Workshop on Integrating Technologies for Interactive Stories (INTETAIN 2008).*

Bae, B. C., and Young, R. M. 2014. A computational model of narrative generation for surprise arousal. *IEEE Transactions on Computational Intelligence and AI in Games* 6(2):131–143.

Bredart, S., and Modolo, K. 1988. Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica* 67(2):135–144.

Breithaupt, F.; Li, B.; and Kruschke, J. K. 2022. Serial reproduction of narratives preserves emotional appraisals. *Cognition and Emotion* 1–21.

Canaj, E.; Biba, M.; and Kote, N. 2018. Bayesian Networks: A State-Of-The-Art Survey. *CEUR Workshop Proceedings* 2280:31–40.

Cheong, Y. G., and Young, R. M. 2008. Narrative generation for suspense: Modeling and evaluation. *Lecture Notes in Computer Science* 5334 LNCS:144–155.

Chu, Y., and MacGregor, J. N. 2011. Human Performance on Insight Problem Solving: A Review. *The Journal of Problem Solving* 3(2):119–150.

Cox, R. T. 1946. Probability, Frequency and Reasonable Expectation. *American Journal of Physics* 14(1):1–13.

Emmott, C., and Alexander, M. 2014. Foregrounding, burying and plot construction. In Stockwell, P., and Whiteley, S., eds., *The Cambridge Handbook of Stylistics*. Cambridge: Cambridge University Press. 329–343.

Evans, J. S. B. T. 1989. *Bias in human reasoning: Causes and consequences.* Essays in cognitive psychology. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. *arXiv preprint arXiv:1805.048331.*

Geier, T., and Biundo, S. 2011. Approximate online inference for dynamic Markov logic networks. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* 764–768.

Graesser, A. C.; Robertson, S. P.; and Anderson, P. A. 1981. Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology* 13(1):1–26.

Graesser, A. C.; Singer, M.; and Trabasso, T. 1994. Constructing inferences during narrative text comprehension. *Psychological Review* 101(3):371–395.

Griffiths, T. L.; Kemp, C.; and Tenenbaum, J. B. 2008. Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(6):811–823.

Guan, J.; Wang, Y.; and Huang, M. 2019. Story ending generation with incremental encoding and commonsense knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence1* 33(1):6473–6480.

Guttag, J., and Horning, J. J. 1980. Formal specification as a design tool. In *Proceedings of the 7th ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '80*, 251–261. New York, New York, USA: ACM Press.

Ha, E. Y.; Rowe, J. P.; Mott, B. W.; and Lester, J. C. 2012. Goal Recognition with Markov Logic Networks for Player-Adaptive Games. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* 2113–2119.

Hua, M., and Raley, R. 2020. Playing with unicorns: AI dungeon and citizen NLP. *Digital Humanities Quarterly* 14(4):1–27.

Itti, L., and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision Research* 49(10):1295–1306.

Jaynes, E. T.; Jaynes, E. T. J.; and Bretthorst, G. L. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Kapadia, M.; Falk, J.; Zünd, F.; Marti, M.; Sumner, R. W.; and Gross, M. 2015. Computer-assisted authoring of interactive narratives. *Proceedings of the 19th Symposium on Interactive 3D Graphics and Games, i3D 2015* 85–92.

Kok, S., and Domingos, P. 2005. Learning the Structure of Markov Logic Networks. In *Proceedings of the 22nd International Conference on Machine Learning*, 441–448.

Kukkonen, K. 2014. Bayesian narrative: Probability, plot and the shape of the fictional world. *Anglia* 132(4):720–739.

Kullback, S., and Leibler, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Kybartas, B., and Bidarra, R. 2017. A Survey on Story Generation Techniques for Authoring Computational Narratives. *IEEE Transactions on Computational Intelligence and AI in Games* 9(3):239–253.

Lehto, M.; Marucci-Wellman, H.; and Corns, H. 2009. Bayesian methods: A useful tool for classifying injury narratives into cause groups. *Injury Prevention* 15(4):259–265.

Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, 598–604.

Loftus, E. F., and Zanni, G. 1975. Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society* 5(1):86–88.

Lowd, D., and Domingos, P. 2007. Efficient weight learning for Markov logic networks. *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)* 200–211.

Marcus, G. F., and Davis, E. 2013. How Robust Are Probabilistic Models of Higher-Level Cognition? *Psychological Science* 24(12):2351–2360.

Marie-Laure Ryan. 2009. Cheap Plot Tricks, Plot Holes, and Narrative Design. *Narrative* 17(1):56–75.

McKoon, G., and Ratcliff, R. 1992. Inference during reading. *Psychological Review* 99(3):440–466.

Measure, A. 2014. Automated Coding of Worker Injury Narratives. *Joint Statistical Meetings* 2124–2133.

Murphy, K. P. 2012. Undirected Graphical Models (Markov Random Fields). In *Machine Learning: A Probabilistic Perspective*. MIT Press. chapter 19, 661–705.

Niu, F.; Zhang, C.; Ré, C.; and Shavlik, J. 2012. Scaling inference for Markov logic via dual decomposition. *Proceedings - IEEE International Conference on Data Mining, ICDM* (1):1032–1037.

Ohwatari, Y.; Kawamura, T.; Sei, Y.; Tahara, Y.; and Ohsuga, A. 2014. Estimation of character diagram from open movie database using Markov logic network. In *CEUR Workshop Proceedings*, volume 1312, 124–127.

Oppenheimer, D. M. 2008. The secret life of fluency. *Trends in Cognitive Sciences* 12(6):237–241.

Patil, S.; Pawar, S.; Hingmire, S.; Palshikar, G.; Varma, V.; and Bhattacharyya, P. 2018. Identification of Alias Links among Participants in Narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 63–68. Stroudsburg, PA, USA: Association for Computational Linguistics.

Piacenza, A.; Guerrini, F.; Adami, N.; Leonardi, R.; Teutenberg, J.; Porteous, J.; and Cavazza, M. 2011. Generating story variants with constrained video recombination. *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops* 223–232.

Porteous, J.; Benini, S.; Canini, L.; Charles, F.; Cavazza, M.; and Leonardi, R. 2010. Interactive storytelling via video content recombination. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference* 1715–1718.

Propp, V. I. 1968. *Morphology of the Folktale*, volume 9. University of Texas Press.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; and Others. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1.

Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine Learning* 62(1-2 SPEC. ISS.):107–136.

Riedel, S. 2005. Improving the Accuracy and Efficiency of MAP Inference for Markov Logic. *Network* 468–475.

Riedl, M. O., and Bulitko, V. 2012. Interactive Narrative: An Intelligent Systems Approach. *AI Magazine* 34(1):67.

Rowe, J. P., and Lester, J. C. 2010. Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. *Proc. 6th Annu. AI Interact. Digital Entertain. Conf.* 57–62.

Russell, S. J., and Norvig, P. 2010. *Artificial intelligence: a modern approach*. Upper Saddle River, N.J: Prentice Hall, 3rd ed. edition.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3):379–423.

Singla, P., and Mooney, R. J. 2011. Abductive markov logic for plan recognition. *Proceedings of the National Conference on Artificial Intelligence* 2:1069–1075.

Skaar, Ø. O., and Reber, R. 2020. The phenomenology of aha-experiences. *Motivation Science* 6(1):49–60.

Skaar, Ø. O. 2019. *Moments of Brilliance: Understanding the Aha-experience through Bayesian Statistics*. Ph.D. Dissertation, University of Oslo.

Swanson, R., and Gordon, A. S. 2012. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):1–35.

Tauber, S.; Navarro, D. J.; Perfors, A.; and Steyvers, M. 2017. Bayesian Models of Cognition Revisited: Setting Optimality Aside and Letting Data Drive Psychological Theory. *Psychological review* 124(4):410–441.

Taylor, J. A.; Lacovara, A. V.; Smith, G. S.; Pandian, R.; and Lehto, M. 2014. Near-miss narratives from the fire service: A Bayesian analysis. *Accident Analysis and Prevention* 62(2014):119–129.

Tobin, V. 2018. *Elements of Surprise: Our Mental Limits and the Satisfactions of Plot*. Harvard University Press.

Topolinski, S., and Reber, R. 2010. Gaining insight into the ”Aha” experience. *Current Directions in Psychological Science* 19(6):402–405.

Trabasso, T., and Sperry, L. 1985. Causal Relatedness and the Importance of Narrative Events. *Journal of Memory and Language* 24(1894):595–611.

Trabasso, T., and van den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24(5):612–630.

Valls-Vargas, J.; Zhu, J.; and Ontanon, S. 2017. From computational narrative analysis to generation: A preliminary review. *ACM International Conference Proceeding Series* Part F1301.

Van den Broeck, G. 2013. *Lifted Inference and Learning in Statistical Relational Models*. Ph.D. Dissertation, KU Leuven.

Vlek, C.; Prakken, H.; Renooij, S.; and Verheij, B. 2013. Modeling crime scenarios in a Bayesian network. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law - ICAIL '13*, 150. New York, New York, USA: ACM Press.

Ware, S. G. 2002. A Plan-Based Model of Conflict for Narrative Reasoning and Generation. 52(1):1–5.

Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; and Yan, R. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7378–7385.