What If A Fish Got Drunk?
Exploring the Plausibility of Machine-Generated Fictions

Maria Teresa Llano[1], Christian Guckelsberger[1], Rose Hepworth[1]
Jeremy Gow[1], Joseph Corneli[1] and **Simon Colton[1,2]**
[1]Computational Creativity Group, Goldsmiths, University of London, UK
[2]The Metamakers Institute, Falmouth University, UK
m.llano@gold.ac.uk

Abstract

Within the WHIM project, we study *fictional ideation*: processes for automatically inventing, assessing and presenting fictional ideas. Here we examine the foundational notion of the *plausibility* of fictional ideas, by performing an empirical study to surface the factors that affect judgements of plausibility. Our long term aim is to formalise a computational method which captures some intuitive notions of plausibility and can predict how certain types of people will assess the plausibility of certain types of fictional ideas. This paper constitutes a first firm step towards this aim.

Introduction

In Llano et al. (2016), we define a successful fictional idea as *"one that presents a character, event or scenario that transforms or distorts the 'real' world in the imagination of the reader without requiring him or her to leave it entirely"*. In the WHIM project (an acronym for The *What-if Machine*), we are undertaking the first large-scale study of how software can invent, evaluate and express fictional ideas with real cultural value (www.whim-project.eu). We have identified *plausibility* as one of the key dimensions of fictionality, and so investigating questions of plausibility is important for the aims of the WHIM project. Unfortunately, plausibility resists a simple definition. Here, we explore the factors that support the perception of a machine-generated fictional idea as plausible or implausible.

Plausibility in fictional scenarios is different from notions of probability, which rely on modelling situations in terms of relative frequency, or the updating of prior distributions. Judgements about plausibility in fictional situations involve a process of *interpretation*, where the reader makes – perhaps implicit – subjective decisions about the underspecified fictional universe. For example, in the absurdist play *Rosencrantz and Guildenstern are Dead*, Rosencrantz bets "heads" on a coin flip 92 times in a row, and wins each time. If presented as a factual news story, this would likely both be judged as implausible and mathematically improbable (albeit no more improbable than any other session of 92 coin flips). However, within the context of the play, we're invited to consider a fictional world in which this highly improbable chain of events is plausible, i.e., it actually happens.

Following an overview of prior research, we propose some candidate factors to capture how people assess the plausibility of fictional statements. We then report on how these were used in an exploratory study, where 20 participants were asked to categorisation a set of fictional ideas – both machine- and human- generated – into four plausibility categories and interviewed about their judgements. We then examine these judgements from two distinct perspectives: (1) a grounded theory analysis that identifies several factors they considered relevant to plausibility; (2) a multidimensional scaling analysis that suggests several factors that can explain differences between how fictional ideas were assessed. Finally – as per Rothbauer (2008) – we triangulate the outcomes of these analyses and our initial theories, leading to a final set of factors. We conclude with a discussion of the relevance of this work for further research on fictional ideation and Computational Creativity in general.

Background

Connell and Keane (2004) carried out an empirical study of plausibility. They first evaluated the *concept coherence* of a set of events written as two connected sentences, e.g.,

The bottle fell off the shelf. The bottle smashed.

These were classified according to different types of inferences; the sentence above references a *causal inference*, while the sentence pair

The bottle fell off the shelf. The bottle was pretty.

references an *attributal inference*. Their results supported the received view that concept coherence is important in plausibility judgements, and showed that different inference types differentially affect plausibility. A second experiment evaluated *word coherence* rather than concept coherence, but this experiment found no reliable effect of word coherence on plausibility.

Connell and Keane note that their studies used "the term plausible interchangeably with other descriptions such as appropriate, sensible, or makes sense." Our approach differs considerably from theirs, as they focused on statements whose two constituent parts have a strong conceptual relation. In contrast, as the fictional statements evaluated in this paper have not been conceived with such restrictions, this has allowed for a more comprehensive analysis.

Lombardi, Nussbaum, and Sinatra (2015) have sought to outline a primarily theoretical model for plausibility judgements. In particular, they examine the role such judgements

play in *conceptual change*, and come to understand plausibility as meaning: 'what is perceived to be potentially truthful when evaluating explanations'. They cite Nicholas Rescher's observation that a statement deemed to be plausible, or potentially truthful, indicates that there has been a 'highly provisional and conditional epistemic inclination towards it' (Rescher 1976).

In sum, there seem to be myriad criteria upon which people may base judgements of plausibility. Some factors in an individual's judgement appear to be highly subjective, often being heavily influenced by personal circumstances, religious belief, cultural background, or political sympathies. Indeed, while most people feel they have an understanding of plausibility, that understanding is almost always destabilised when an individual tries to apply consistent criteria to analyse the plausibility of various sample statements.

Plausibility is closely connected with the notion of *interpretation*: that is, a given interpretation of a given scenario is deemed "plausible" if potentially valid, under a given set of assumptions. Interpretation of conventional symbols is highly constrained. However, creative interpretations may be almost endlessly fanciful. To take one example: the science fiction author Philip K. Dick suggests in several of his written works that we continue to live in biblical times, evidence of which can only be accessed by visionary experience (Dick 1995). This interpretation of the world is grounded in the data of personal perception and reflection. Nevertheless, most people would find such an interpretation implausible if presented as anything other than fiction.

Eco argued that the world, history, and texts have constraints on their plausible interpretations, despite the wide range of *possible* interpretations. He posits that fiction and reality intersect in the following way: "We can make true statements about fictional characters because what happens to them is recorded in a text, and a text is like a musical score" (Eco 2009). Music shows us that a creative work can take on a life of its own through interpretation. Fictional characters can also "become individuals living outside their original scores," or, to put it more formally: "a fictional character is a semiotic object."

We can thus make chains of interpretations about fictional characters and other elements of fictional worlds. In the first instance, the validity of such interpretations is not "grounded" in real-world facts, but in the fictive notions of the fictional world – subject also to the perceptions, beliefs, and other features of the interpreting agent. When instigating behaviour (including storytelling behaviour), certain interpretations may be predicted, based in part on a preliminary interpretation of those agents who are expected to perceive the behaviour (Kockelman 2012).

We believe it is important for Computational Creativity researchers to tackle issues related to fictional interpretations, in particular to ask what kinds of interpretations are *useful* (Eco 2006) – rather than merely true. Although subjectivity plays a role, keeping in mind Eco's remarks on limits of interpretation, we think that the reader's perception of a text's plausibility will often draw on the text's objective features, and we develop this theme below.

Candidate Factors

As a first step, we conducted an introspective study to identify an initial set of factors that may be involved in human plausibility judgements. These candidate factors helped guide the design of the exploratory study, described below. Eight fictional statements were used. Four were from the What-If Machine and four were summaries of well known literary works, included to foster the generalisability of the findings. The machine-generated statements were selected for quality and diversity, to showcase a range of potentially relevant factors.

Three of the authors independently read the statements and rated each sentence 1–5 in terms of how plausible they were (1= low plausibility and 5 = high plausibility). They also wrote a commentary for each statement, describing their rationale for that score, their scoring process (including whether they had revised a score during analysis), and a set of labels that described relevant properties, dimensions or features. By comparing our individual answers, we found a common set of factors that appeared to affect our plausibility judgements, listed below.

Complexity The level of elaboration of the idea in terms of the amount of narrative detail that it is composed of. Our intuition is that a larger number of statements, or narrative details, used to compose an idea reduces its plausibility. An illustrative example is the statement:

> *"What if there was a poor orphan girl who was abused by her aunt, sent away to school where conditions were harsh, before becoming a governess and marrying her employer, who was already married to a mentally ill woman whom he has locked up in his house?" (1)*

Each part is rather plausible, but their conjunction renders the overall idea less plausible. *We hypothesise that there is a negative correlation between complexity and plausibility.*

Universality The scope of an idea, in terms of how general people think it is intended to be. In other words, whether the scenario in the fictional idea applies to one, a few or all members of a group. The intuition behind this is that an idea that is generalised to a large number of members of a group is less plausible than an idea that only involves one member. For instance, from the statement:

> *"What if there was a young girl who went through a rabbit hole and found herself in a strange and mysterious land where animals could talk and everyone is mad?" (2)*

The implicit universally quantified sentences "animals could talk" and "everyone is mad" decrease the plausibility of the idea. *We hypothesise that there is a negative correlation between universality and plausibility.*

Openness How open to subjective interpretation an idea is perceived to be. Our intuition is that if an idea that is composed of statements that are ambiguous or not specific, for which the reader can provide different interpretations or scenarios, is perceived as more plausible. As an example, take the statement:

> *"What if there was a young man who kept a painting of himself which aged while he himself stayed young?" (3)*

Here, "stayed young" could be interpreted both in terms of not looking old or actually not ageing, while the painting that "aged" could have been painted in highly impermanent materials. The possible explanations (natural youthfulness and cheap paint, or a bizarre medical condition) differ strongly in their plausibility; if there is a choice, a subject might choose the more plausible explanation. *We hypothesise that there is a positive correlation between openness and plausibility.*

Causality The level of connectivity between the components that make up an idea. In other words, how naturally the statements that make up an idea lead coherently from one to the other. The intuition behind this is that plausibility increases when the statements of an idea are clearly connected so as to serve as supporting arguments themselves. To illustrate this, take the example:

"What if there was a little doctor who couldn't take a pulse?" (4)

Without an explanation as to why the doctor is unable to perform the common task of taking a pulse, this statement will likely score low for plausibility. *We hypothesise that there is a positive correlation between causality and plausibility.*

Familiarity The level of awareness of the overall scenario relative to known ideas. Although this is a subjective factor, the intuition behind it is that our perception of plausibility is affected by common themes, scenarios and characters that figure more commonly in culture. Statement (1) illustrates this intuition. Well-known character stereotypes such as *an orphan girl*, *an evil aunt* and *a mentally ill woman* render the statement more plausible. *We hypothesise that there is a positive correlation between familiarity and plausibility.*

Feasibility How well the elements within an idea fit within the overall scenario. The intuition behind this is that plausibility increases if an element; e.g., a character, is better suited to one situation than another. To illustrate, the statement:

"What if there was a little cat who learned how to use a phone?" (5)

Would rank lower in plausibility if instead of *a cat*, the subject was an inanimate object, for instance *a cooker* due to the affordances of the subjects. *We hypothesise that there is a positive correlation between feasibility and plausibility.*

An Exploratory Categorisation Study

To further explore the factors underlying human plausibility judgements, we conducted a categorisation study where the above candidate factors guided the selection of the stimuli for, and design of, the study. In the study, participants were asked to assign machine- and human-generated fictional ideas into different categories of plausibility. We collected both quantitative and qualitative data from these sessions, which were then separately analysed and interpreted: 1) the raw categorisation results were used to calculate explorative statistics; 2) participant think aloud commentaries and post-task interviews formed the basis of a grounded theory analysis (Adams, Lunt, and Cairns 2008), identifying key factors in their categorisation process. 3) The categorisation results were transformed into similarity data for a multidimensional scaling (MDS) analysis (Borg and Groenen 2005) of the fictional ideas, to collect more evidence on the underlying dimensions which influenced the plausibility categorisations; By studying the same plausibility judgements qualitatively and quantitatively, we hoped to triangulate the results to arrive at a final set of factors. A categorisation task with physical cards, as opposed to ordinary Likert-scale rating, was deliberately chosen to promote think-aloud comments and the comparison of stimuli. These methods are well-suited to exploring complex and poorly conceptualised domains, e.g., see Wallraven et al. (2009) or Gow et al. (2010).

Stimuli We used 28 fictional ideas in total, consisting of 18 ideas generated by The What-If Machine (three from each of the six categories the system currently supports: "Disney", "Metaphors", "Utopian/Dystopian", "Alternative Scenarios", "Kafkaesque" and "Musicals"), 7 ideas summarising well-known fictional literature works ("Literary Fiction"), and 3 ideas that used known fictional characters or worlds ("Fiction in Fiction"). We also selected a subset of six ideas (from the 28 already selected) for the participants to verbally elaborate on in more detail. A selection of stimuli from each category can be found in Table 1.

Method Participants took part in the study individually and were all read the same introductory material. Each session was audio recorded for later analysis. We first asked them to sort the 28 stimuli, provided as paper cards, into four plausibility categories:

1. **Highly implausible:** describe scenarios that have very little grounding in your experience of reality.
2. **Slightly implausible:** describe scenarios that have a low degree of grounding in your experience of reality.
3. **Slightly plausible:** describe scenarios that are somewhat grounded in your experience of reality.
4. **Highly plausible:** describe scenarios that have a high degree of grounding in your experience of reality.

An "**I don't understand**" category was also provided. In contrast to our introspective study, we chose four categories to eliminate the neutral choice. Participants were not told that some of the statements had been written by software, nor asked if they recognised those from human-authored narratives.

We asked participants to think aloud while performing this task, i.e., to articulate their categorisation process and rationale. For some participants (see below) this was followed by open-ended questions where these issues were probed in greater depth, focusing on the six statements which we had pre- selected, or others highlighted during the categorisation study. Finally, we asked some participants explicit questions about our candidate factors (described above), to determine if they considered them relevant. For instance, regarding *complexity* we asked: "Do you believe that a complex fictional statement; that is, with a large amount of conditions, makes the plausibility higher, lower or neither?". Each such question was accompanied by an example statement.

Id	Mean	Var.	Ag.	NAs	Stimulus	Category
5	2.00	1.44	14	0	What if we could give life to a being created by combining the body parts of dead people?	Literary Fiction
8	0.39	0.72	16	1	What if a zombie rugby-tackled a ghost and broke his leg?	Fiction in Fiction
12	0.38	0.78	14	3	What if there was a little pen who forgot how to write?	Disney
14	2.50	0.62	15	1	What if ignorant fools were to overcome mistakes, establish cults and become knowledgeable gurus?	Metaphors
19	1.69	1.03	9	3	What if the world suddenly had lots more assassins? Then there would be more antidotes, since assassins use the poisons that require antidotes.	Utopia / Dystopia
20	0.44	0.61	15	1	What if there was an old fish, who couldn't swim anymore, which he used to do for relaxation, so decided instead to get drunk?	Alternative
21	0.94	1.31	11	2	What if there was an old car that could be used as the space for holding a star?	Alternative
24	0.17	0.26	17	1	What if a bicycle appeared in a dog pound, and suddenly became a dog that was able to drive an automobile?	Kafkaesque
26	2.72	0.21	18	1	What if a wounded soldier had to learn how to understand a child in order to find true love?	Musicals
28	2.06	0.43	14	2	What if a janitor needed to suppress a rebellion in order to gain admiration?	Musicals

Table 1: A selection of stimuli, with response mean and variance. Ag.= participants agreeing with most common response. NAs = times classified as "Don't understand".

Participants In total, 20 participants took part in the study, although one participant's data was excluded (see below). Of the remaining 19, 4 participants were female and 15 male. 8 participants were in the age range 18-24 years old, 9 were 25-34, and 2 were 35-44. 4 of them specified A levels as their current level of education, 7 had a first degree, 6 a higher degree, and 2 a doctorate. 7 participants were fluent in English, 11 were native speakers, and one self-rated as "intermediate", but was considered fluent. We assumed that the lack of demographic diversity would have limited impact on our results, although future studies could make some provision for variations related to gender, age or educational background, e.g., cultural references. Participants did not have familiarity with our work on plausibility prior to taking part the study.

All participants were paid £10 and undertook the categorisation experiment. Only 12 were asked the open-ended questions and questions about the candidate factors. This allowed us to constrain the amount of data collected for the grounded theory analysis, while satisfying representativeness for the quantitative analysis. One participant was a very distinct outlier in terms of categorisation mean and variance, as they classified most statements into either "highly implausible" or "I don't understand". They were perfectly aware of the meaning but didn't agree with the logic of the statement. Their think aloud data also suggested that they did not engage with the task as requested. We therefore excluded this participant from the analysis that follows.

Categorisation Results

Of the 532 judgements made, the most common were "highly implausible" (34%) and "highly plausible" (25%), followed closely by "slightly plausible" (23%). The least common responses were "slightly implausible" (11%) and "don't understand" (7%). In the analysis below, we sometimes interpret these ordinal responses (excluding "don't understand") as interval data from 0 (highly implausible) to 3 (highly plausible). Table 1 shows the response mean and variance for a selection of stimuli.

By Participant All participants used the entire range of responses. There were notable individual differences: 4 participants had median response of "highly implausible", 8 had "slightly plausible", with the remaining 7 medians falling in-between. The variance for each stimuli provides a measure of agreement between participants: the mean variance was 0.93 (min 0.21, max 1.49), suggesting quite a high level of disagreement. However, if we ignore the distinction between *highly* and *slightly* and merge categories to *plausible/implausible/don't understand*, we actually see many participants agreeing with the modal (most popular) category: for 68% of stimuli, at least two-thirds agree. This shows at least a weak consensus was often present.

Participants used the "I don't understand this statement" category a median of 1 times, indicating comprehension was not a problem for most participants. Only one participant claimed to understand all stimuli and, at the other extreme, two didn't understand six stimuli. Using Spearman's ρ, there is a medium negative correlation ($\rho_S(28) = -0.4, p = 0.09$) between not understanding and use of "Highly implausible" and a medium positive correlation ($\rho_S(28) = 0.35, p = 0.1$) between not understanding and that participant's mean plausibility. This suggests there may be some confusion between "don't understand" and "implausible", which should be addressed in the design of future studies.

By Stimuli Almost all the stimuli provoked the full range of responses, confirming that assessing plausibility is a highly subjective task. The mean response for each stimuli

ranged from 0.17 (Stimuli 24) to 2.84. The variance ranged from 0.21 (Stimuli 26) to 1.49. We compared the plausibility ratings between the different stimulus groups described above. A Kruskal-Wallis one-way analysis of variance indicated that the plausibility ratings between the eight groups were significantly different $H(7) = 80, p < 1e - 13$. We then performed a series of Wilcoxon rank sum post-hoc tests with Bonferroni correction to determine which of the groups are significantly different. The p-values and group means are listed in Table 2. It shows, amongst others things, that statements from the categories "Musicals" and "Metaphor" were rated highest in plausibility ($\mu = 2.33, \mu = 1.76$), and differ significantly from the categories that were considered highly implausible, namely "Kafkaesque" and "Utopia/Dystopia" ($\mu = 0.56, \mu = 1.06$).

Think Aloud Results

To understand the factors which contributed to participants' plausibility judgements, we performed a grounded theory analysis of the think aloud data. Grounded theory is a qualitative research method that is used to build, validate and expand theories from data, in order to reach "a theoretical formulation of the reality under investigation" (Corbin and Strauss 1990). Our analysis validated four of our initial hypothesised factors as influential within our participants' judgements: *openness*, *familiarity*, *causality* and *feasibility*; the other two, *complexity* and *universality*, were concluded as non-influential. An additional factor, *perception of reality* was identified. Furthermore, for each of the supported factors, we identified a set of properties that represent the different ways the participants talked about the factors, as well as dimensions describing values these properties can hold. These results are summarised in Table 3.

Participants often based their judgement on how ***Familiar*** they were with the content; either from *experience* (own or by others) or *knowledge* they have acquired from different mediums. An illustrative example is Statement (1), quoted earlier to highlight the Complexity dimension. Two participants said the following:

> "Doesn't go with things in this time and day but people's lives are complicated [...]"

> "you see similar situations in the news [...] these are different personalities that actually exists [...]"

Consequently, this statement was often classified in the plausible spectrum (10 as highly plausible and 8 as slightly plausible). Familiarity at the level of *cultural recognition* also affects plausibility judgements. For instance, despite the fact that the statement: *What if a zombie rugby-tackled a ghost and broke his leg?*, contains fictional characters, as these are well-known concepts that form part of our culture, participants would hesitate about their plausibility value (even if eventually most decided the statement was not plausible).

Openness was also a recurrent factor we identified from the recorded sessions. Often, participants would try to make sense of the statements, saying things such as:

> 'Maybe because my brain [is trying to give] sense to sentences.'

> 'Where there is more room for interpretation, it is more easy to be black or white.'

Ambiguity and *context* played an important role for this factor. We found that key concepts appearing in a statement made a significant difference in plausibility judgements when these could be interpreted in different ways, and there was not enough context to narrow down the intended meaning. This led participants to stick with their favourite interpretation and provide their judgement accordingly. To illustrate, regarding the statement '*What if there was an old car that could be used as the space for holding a star?*', participants would often ask if the concept *star* meant the astrological object or a celebrity, with most of the participants selecting the former and consequently placing this statement within the implausible spectrum. A similar reasoning was common with statement (4) above, for which participants would consider the concept of the *little doctor* as being either a child or a doctor short in height. Most participants chose the former interpretation and placed the statement in the plausible spectrum.

Feasibility was also one of the factors used by the participants when judging plausibility. In particular, we found that they would consider if the *likelihood* of the statement would form a usual or unusual scenario to decide on its plausibility. This was often seen in statements like (1), where the co-occurrence of all the elements of the statement was seen as unusual – but still plausible. One participant said:

> 'it's quite a complicated story but elements of the story makes it feel more real.'

Additionally, feasibility was also accounted for based on the use of *stereotypes* and how the individual parts of the statement fit together with a stereotypical construct. To illustrate, take the statement: *What if the world suddenly had lots more dictators? Then there would be less neediness, since dictators abuse the victims that demonstrate neediness*, for which a handful of participants focused on the contradiction between the concept of *dictators*, which has negative connotations, and the concept of *less neediness*, which has positive connotations.

Specific keywords, in particular *attributes* of the concepts in the statement, were also a decisive property when judging the statement based on its feasibility. For instance, the use of the adjective *little* in statement (4) made the plausibility higher, since participants interpreted the scenario as a *child playing doctor who is not able to actually take a pulse*, which in their view was completely feasible.

We also found that, although ***causality*** was not a strong factor in the decision making process, it was present on some occasions. In particular, finding *arguments* in favour or against particular elements of a statement had an influence in plausibility judgements. For instance, the statement: *What if the world suddenly had lots more assassins? Then there would be more antidotes, since assassins use the poisons that require antidotes*, links the concept of *assassins* with the concept of *poisons*, and this itself to the concept of *antidotes*. Although this statement was built through well-attested associations, specifically that assassins use poisons, and that poisons require antidotes, the intended strong link between assassins and poisons was used constantly as an argument against the plausibility of the statement. In contrast, from the statement: *What if there was a punishable man who had to learn how to eat a person in order to achieve his dream of becoming a criminal?*, the link between 'eating

Category	Literary Fiction	Fiction in Fiction	Disney	Metaphors	Utopia / Dystopia	Alternative	Kafkaesque	Musicals
Mean Plausibility	1.71	1.11	1.37	1.76	1.06	1.12	0.56	2.33
Fiction in Fiction	0.11394	-	-	-	-	-	-	-
Disney	1.00000	1.00000	-	-	-	-	-	-
Metaphors	1.00000	0.16271	1.00000	-	-	-	-	-
Utopia / Dystopia	0.07103*	1.00000	1.00000	0.08518*	-	-	-	-
Alternative	0.14113	1.00000	1.00000	0.19369	1.00000	-	-	-
Kafkaesque	5.9e-07**	0.48768	0.00455**	2.4e-06**	0.31448	0.44275	-	-
Musicals	0.09299*	3.5e-05**	0.00039**	0.23095	5.9e-06**	4.4e-05**	1.9e-11**	-

Table 2: p-values from pairwise comparisons of stimuli groups using Wilcoxon rank sum test and Bonferroni correction. Significance: * low ($\alpha < 0.1$) and ** high ($\alpha < 0.01$). The second row comprises plausibility means for all categories.

a person' and becoming a 'criminal' was seen as logically connected:

> 'I can imagine eating a person as an act of initiation for a person to be part of a gang...'

Interestingly, the idea of *unknowns* was also used as an argument to decide on a plausibility category. This is when a participant considered that he/she did not have enough knowledge to argue against or in favour of a particular scenario. An example of this was the statement: *What if the ministry of magic paid JK Rowling to write her books so we muggles would think magic is fiction?*:

> 'I don't know if there is a minister of magic [...] who knows?'

which some participants used as an argument to assign a higher plausibility value.

Lastly, ***perception of reality*** played a role for some individuals when making their judgements. This factor represents how people may account for different ways of perceiving reality within certain scenarios. To illustrate, a participant categorised statement (1) as highly plausible based on the following reasoning:

> "From my experience of reality, that might happen in some psychedelic state, a dream state, an imaginary state. I don't think it's right to count what happens in these states as any less real [...] the rabbit hole could be a doorway to other states. That's an idea I'm definitely open to."

another participant questioned the meaning of reality:

> '[...] what is reality? Different statements push different readings of what reality is: objective reality in terms of things that are physically possible for ever and ever, things that might be possible in the future with technology, things that kind of work in a fictional world, and things that don't work at all.'

Factor	Properties	Dimensions
Familiarity	Experience	Own/Others
	Knowledge	Cultural/Heard/Read/Seen
	Cultural recognition	Conceptual/Factual
Openness	Ambiguity	Most/Least plausible
	Context	Lack/Presence of
Feasibility	Likelihood	Usual/Unusual
	Stereotypes	Confirmation/Contradiction
	Attributes	Opened/Specific
Causality	Arguments	In favour/Against/Unknowns
	Context	Lack/Presence of
Perception of reality	Abstraction	Conceptual/Physical
	Cultural influence	Background/Beliefs

Table 3: Influential factors when judging plausibility.

This is a subjective factor, but the intuition behind it is that judgements of plausibility are affected by personal views of what can be considered to be real or not.

Within this factor, *abstraction* was found to be a common property. In this case, the overall scenario was considered as having a hidden meaning. To illustrate, the statement: *What if the world suddenly had lots more angels? Then there would be more barriers, since angels serve the gods that impose the laws that create barriers*, was abstracted by some participants:

> 'I don't believe in angels or God, but I think the government can use it as a tool to manage people.'

leading them to assign a higher plausibility value to the statement. Similarly, *cultural influence* played a role in how participants' perception of reality would affect plausibility. Take the statement: *What if respected senators were to retire from their senates, join gangs and become shady gangsters?*, which was implausible for many participants because it did not make sense with their notion of reality:

> 'there is no reason why a senator with power and money would choose to be a gangster'

while for others, this was a plausible scenario due to their cultural background, where this situation was feasible:

> '[...] in certain very corrupt countries it actually happens [...] when they are senators they belong to gangs, legal ones, but they do [...] and when they retire they keep being part of those clubs'

Likewise, the statement *What if a janitor needed to suppress a rebellion in order to gain admiration?* was classified as slightly plausible because:

> 'this is kind of a standard Hollywood plot really, I can imagine that being played by Tom Cruise [...] it has high degree of grounding in my experience of reality [...] not my experience of reality, my experience of Hollywood film making'

suggesting the participant considered the fictional world of Hollywood films as a type of reality.

Complexity, as mentioned before, did not come across as an influential factor for plausibility. For instance, regarding the complexity of statement (1), a participant highlighted:

> 'All happening at once is unlikely but it's possible [...] that doesn't change the plausibility.'

instead, this factor was considered to sometimes make the statements more difficult to understand. *Universality*, on the other hand, was seen as a factor that would make a statement more interesting, but would not have a significant effect on its plausibility value, specially when the statement was placed in the implausible spectrum:

'if you are in the implausible categories, then it doesn't matter, one, many [...] we are talking about something that is not real, so it doesn't matter'

Multidimensional Scaling Results

We performed an multidimensional scaling (MDS) analysis to quantitatively derive a set of factors from the categorisation results, and to assess their influence on the overall judgment. Classic MDS maps measurements of (dis-) similarity among pairs of stimuli to distances between points in a geometric space (Borg and Groenen 2005, p. 3). In this space, each dimension can be considered a factor which influenced the initial similarity judgement. The meaning of a dimension is a matter of interpretation, based on the distribution of stimuli along it and their properties.

First, we had to determine the pairwise similarities between stimuli. The effort of collecting this data manually increases exponentially with the number of stimuli; we therefore followed a different approach suggested by Wallraven et al. (2009), where pairwise similarities are derived from a categorisation task. This approach allowed us to re-use our previously collected data, while implicitly grounding the similarities in plausibility judgements. We started with an empty similarity matrix, and increased the similarity value of two stimuli if they were put into the same plausibility group. This was repeated for all participants and normalised.

We then determined how many dimensions have to be used to approximate the data well enough by looking at how much variance in the data each dimension accounts for (Borg and Groenen 2005, pp. 247). We cut off at three dimensions, accounting for 78% of the variance, with the first dimension covering 57%. We then visualised each of these as a one-dimensional axis with the stimuli projected along it, allowing us to compare the relative distribution of the stimuli. These visualisations were given to four of the authors for interpretation, informed by the think-aloud results. A consensus interpretation of each dimension was then agreed on. These are summarised in Table 4, along with some examples of high and low scoring stimuli from Table 1.

On the first dimension, statements that showed a strong deviation from reality were grouped in one extreme. On the other extreme were statements that were more aligned with the rules of what it is commonly agreed as possible. The dimension was identified as *feasibility*. The second dimension was strongly associated with interpretability, i.e. with the stimuli's *openness to interpretation*. Interestingly, the stimuli in both extremes were found to have different interpretations; however, what separated one group from the other was how ambiguous the possible interpretations were assessed to be. In one extreme, an interpretation would allow for a more decisive judgement, while in the other, the interpretation would still be seen as not convincing. The third dimension was found to classify stimuli based on *familiarity*. Well known elements, similar stories, common characters and stereotypes were identified in one extreme, while the other extreme presented the same characteristics (i.e. familiar elements) used in contradictory ways.

Dim	Var	Example stimuli (Id) High	Low	Interpretation
1	57%	14, 26, 28	8, 12, 24	Feasibility
2	12%	5	28	Openness
3	9%	14, 26	19, 28	Familiarity

Table 4: The first three dimensions identified by MDS.

Future Work

This study provides evidence for three factors — feasibility, openness to interpretation, and familiarity — that contribute to judgements about the plausibility of fictional ideas. Understanding these factors is a necessary step towards further experimental investigation in this area. We plan to further test and refine this theory, and use it to design studies on the perception of machine-generated fictions.

We intend to model these factors computationally within the What-If Machine, to control the plausibility of the fictional ideas it generates. This could enhance the usefulness of the software and perhaps increase the cultural value of the ideas it produces. Although further experimentation is needed, we believe that metrics which predict values for each of the factors can be devised. Moreover, these could be used to predict the plausibility judgement that certain types of people will make for particular fictional statements.

A heuristic approach to analyse whether a statement is open to interpretation can be based on the *concreteness scores* of its constituent keywords. This measures the level of ambiguity of these words and give an approximation of the concreteness of the overall scenario. We have formalised fictional statements within the WHIM project as short narratives composed of narrative points that are either linked through causal relations, assumed by the reader, or given by the knowledge base (Llano et al. 2016). This formalism allows us to represent each statement as a graph over which we can reason. For instance, analysing the connectivity within the graph may allow us to hypothesise the level of *contextual support* within the statement as a whole. Highly supported statements may be less open to interpretation.

Feasibility, on the other hand, could be accounted for through the use of techniques such as a distributional semantics vector space model (Mikolov et al. 2015). Specifically, how well the elements of a statement fit together could be measured by studying their *semantic similarity* as well as their shared contextual co-occurrences. Stereotypical properties of concepts can be mined from the web (Veale 2012). A similar method could be followed in order to assess stereotypes within a statement and compare the polarity between the stereotypes and the other elements in the statement.

Finally, although familiarity is a subjective factor, metrics could be defined by establishing links with the information in knowledge bases of common knowledge and narrative constructs. In this context, strongly linked data can be seen as connected to "known or familiar scenarios".

Progress in fictional ideation has general implications for Computational Creativity. In the problem solving paradigm of AI, intelligent tasks to automate are broken down into a series of problems to be solved, and there is a usually a

'right answer' to these problems whether local or global, or at least a fitness function relating to the potential value of solutions, which ordinarily captures notions of value from the real world. In the artefact generation paradigm of AI, however, an intelligence task to automate is considered as an invitation to create something of value in a potentially interesting way. Value can be externally imposed, but some Computational Creativity projects have allowed software to invent its own measures of value and to motivate these through framing (Charnley, Pease, and Colton 2012).

We can use the above observations on plausibility to set ourselves apart somewhat further from mainstream AI. In particular, Computational Creativity research could be seen as the sub-field focused on AI for *what could be* rather than *what is best*. Approaches to generate information about possible worlds naturally includes making discoveries about the real universe around us. However, it also includes the invention of imagined scenarios specifically constructed to reflect alternative realities. Such scenarios, like those produced by The What-If Machine, are valuable not because of their explicit reflection of reality, but because they force us to see the realities of our own existence in new and thought-provoking ways (in addition to simply providing entertainment). Analyses building on the work presented here could be influential in the advancement of the automatic generation of fictional universes and other creative works. Predicting how plausible (or not) people judge an idea to be will be a key part of automatically producing imagined scenarios.

Conclusions

We conducted a study in which participants categorised human authored and machine generated fictional statements (including the one paraphrased in the title of this paper), in terms of their plausibility. Unlike previous studies, in which the notion of plausibility had only operational significance, we explored the constituent factors of plausibility, in order to determine which are most influential. We found that the three most influential factors when judging plausibility are: *feasibility*, which determines how well the elements of an idea fit within the overall scenario, *openness* to subjective interpretation, and *familiarity*, which specifies the level of awareness of the overall scenario relative to known ideas. Our findings can serve as a theoretical grounding for future cognitive and computational studies involving plausibility, as well as informing wider discussions about perceptions of fictionality. We hope to build on this work to improve the cultural value of machine-generated fictions and to make fictional ideation a central part of Computational Creativity.

Acknowledgments

This research was supported by the European Commission via the WHIM (FP7 grant 611560), the EPSRC IGGI CDT (EP/L015846/1) and by EPSRC Leadership Fellowship grant EP/J004049/2.

References

Adams, A.; Lunt, P.; and Cairns, P. 2008. A qualititative approach to HCI research. In Cairns, P., and Cox, A., eds., *Research Methods for Human-Computer Interaction*. Cambridge University Press. chapter 7.

Borg, I., and Groenen, P. J. F. 2005. *Modern multidimensional scaling: Theory and applications*. Springer.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proc. 3rd International Conference on Computational Creativity*, 77–81.

Connell, L., and Keane, M. T. 2004. What plausibly affects plausibility? concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory & Cognition* 32:185–197.

Corbin, J. M., and Strauss, A. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13(1):3–21.

Dick, P. K. 1995. How to build a universe that doesn't fall apart two days later. In Sutin, L., ed., *The Shifting Realities of Philip K. Dick: Selected Literary and Philosophical Writings*. Vintage.

Eco, U. 2006. Weak Thought and the Limits of Interpretation. In Zabala, S., ed., *Weakening Philosophy: Essays in Honour of Gianni Vattimo*. McGill-Queen's U. Press.

Eco, U. 2009. On the ontology of fictional characters. *Sign Systems Studies* 37(1/2):82–97.

Gow, J.; Cairns, P.; Colton, S.; Miller, P.; and Baumgarten, R. 2010. Capturing player experience with post-game commentaries. In *Proc. 3rd Int. Conf. on Computer Games, Multimedia & Allied Technologies*.

Kockelman, P. 2012. The ground, the ground, the ground: Or, why archeology is so 'hard'. *The Yearbook of Comparative Literature* 58:176–184.

Llano, M. T.; Colton, S.; Hepworth, R.; and Gow, J. 2016. Automated fictional ideation via knowledge base manipulation. *Cognitive Computation* 1–22.

Lombardi, D.; Nussbaum, E. M.; and Sinatra, G. M. 2015. Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist* 1–22.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2015. Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL].

Rescher, N. 1976. *Plausible reasoning: An introduction to the theory and practice of plausibilistic inference*. K.Van Gorcum & Co.

Rothbauer, P. M. 2008. Triangulation. In Given, L. M., ed., *Encyclopedia of Qualitative Research Methods*. SAGE.

Veale, T. 2012. A context-sensitive, multi-faceted model of lexico-conceptual affect. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics, Vol. 2: Short Papers*, 75–79. The Association for Computer Linguistics.

Wallraven, C.; Fleming, R.; Cunningham, D.; Rigau, J.; Feixas, M.; and Sbert, M. 2009. Categorizing art: Comparing humans and computers. *Computers & Graphics* 33(4):484–495.