I heard he might be guilty, so he probably is: Uncertain evidence statements in iterative reproductions of crime stories

Anonymous CogSci submission

Abstract

Include no author information in the initial submission, to facilitate blind review. The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading "**Abstract**" should be 10 point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header "**Keywords:**" and a list of descriptive keywords separated by semicolons, all in 9 point font, as shown below.

Keywords: iterated narration; transmission chains; crime stories; suspect; guilt

[ek: General notes: make up your mind about generations vs. reproduction; original stories vs. seeds; stories vs. story-type vs. condition,...]

Introduction

One of the central goals in language use is the exchange of information. We obtain new information by reading the newspaper, or listening to the radio or a friend. We can use this newly acquired knowledge and communicate it to other people in our environment. Yet this process of (partially selective) iterated reproduction is not necessarily innocuous: it may distort or alter the original story (Hills, in press). In its simplified linear form, we know this transmission phenomenon as the game of Telephone. The first person whispers a sentence to their neighbor, who in turn has to pass it on to the next person, and so on. After several iterations, the last person in the chain announces the sentence which they ended up with. To everyone's amusement, we often find that this final sentence differs remarkably from the initial one. This simple game nicely exemplifies the information loss and distortion that is associated with repeated exposure and reproduction of information.

Bartlett first introduced this methodology of transmission chains, i.e., chains of reproductions, as a scientific method. In his book "Remembering" Bartlett (1932), he presents a series of transmission chain studies, using stories such as Native American tales or sport reports for reproduction. Bartlett observes a significant information loss of the stories over generations of reproductions and that the content of the stories changes [ek: en par] with the reproducer's prior knowledge.

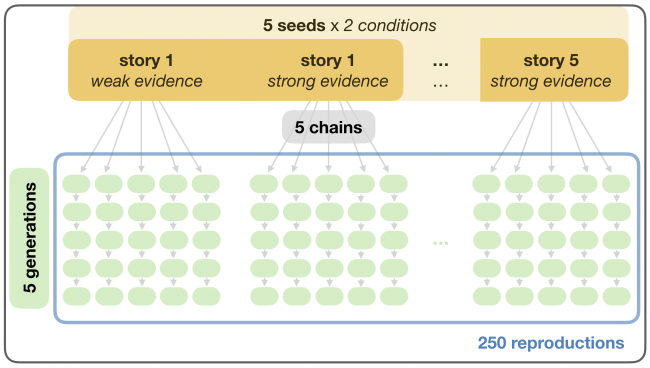

Figure 1: Overview of corpus of stories collected in Exp. 1.

Bartlett used these observations as a foundation for his theory that memory retrieval involves a process of reconstruction.

In recent years, the transmission chain method received a revival in the scientific community. Mesoudi and Whiten (2004) extend Bartlett's generalization hypothesis by using script theory to show that with each iteration, the described events become increasingly abstract. Further research showed that [ek: gender stereotypes: Bangerter 2000, Kashima 2000; cognitive biases: Kalish 2007, Griffiths 2007/2008; Stubbersfield 2015/2017; Hills/Jagiello 2018]

In summary, we know that we use language and communication to exchange information, but we also know that the process of passing on information is flawed in very particular ways. Given their political relevance, we look at how crime stories change in a transmission chain and how this is influenced by seemingly weak and strong of evidence.

To investigate how crime stories evolve over iterations, we conducted two experiments. First we collected a corpus of reproductions for five crime stories, each addressing a different type of crime (e.g., animal smuggling, arson or sexual assault). Each story existed in a weak and a strong evidence condition. This manipulation has successfully been used by (Van Prooijen, 2006) to uncover in- and out-group effects in guilt judgments. Similar to his study, the different conditions were achieved by changing the last sentence in the story which then either suggested strong or weak evidence.

We want to investigate how these stories develop in a transmission chain paradigm (as displayed in [ek: figure ref]).

To evaluate the stories' development, we conducted a second experiment which asked participants to answer questions about the suspect's guilt, the likelihood of conviction and other suspect, author and reader related questions.

Experiment 1: corpus collection

[ek: ...]

Methods

74 Stanford students participated in this online study for course credit. We constructed five stories (*seeds*) that marked the beginning of each reproduction chain. Stories were written in the style of short news articles and followed a similar structure. They reported a crime or moral rule violation that occurred, the authorities' determination of and search for the perpetrator(s), and the possible punishment the suspect(s) would face if found guilty. Furthermore, each of these five seed stories occurred in one of two conditions: a *weak evidence* and a *strong evidence* condition. Evidence strength was manipulated in the final sentence of the story (see example seed in Table 1).

Each participant read and reproduced five stories (either only seed stories, a mix of seeds and reproductions from previous participants, or only reproductions). The assignment of the condition for each story was random. [mf: not sure if I understand the procedure based on this description; did it even matter which evidence condition a later reproduction was from for the allocation of participants to input stories?] On each trial, participants first read a story. They were told to click the 'Continue' button when they were confident that they had internalized the story. Once they clicked the button, the story disappeared and they were asked to reproduce it freely in a text field. Order of stories was randomized.

Results

Participants produced 370 stories. For each seed, we defined a complete chain as one that has 5 reproductions/generations. For subsequent analysis, we randomly selected 50 complete chains, evenly distributed across stories and conditions. This yielded a corpus of 250 reproductions (5 seeds in 2 conditions with 5 complete chains each, see Figure 1). [mf: Can we further motivate exactly this selection? Was this somehow a maximal set that yielded the same number of obseravations for each cell in the design? This is, after all only 2 thirds of all data points.] This corpus is a rich source of linguistic information which merits detailed investigation. Yet, with an eye to clear operationalizability, we focus here on a few general features, which we will subsequently use as predictors in the analyses of Exp. 2 below.

Story length. As shown in Figure 2, the number of words of a reproduction decreased across generations ($\beta = -17.12$, $SE = 1.02$, $t = -16.79$, $p < 0.0001$), replicating a well-known phenomenon in reproduction studies (Bartlett, 1932). While the original generation 0 seeds consisted on average of 159 words, that number dropped to 25 by generation 5. Examples of reproductions of the seed in Table 1 from generation 1 and 5 are shown in (1) and (2) below.

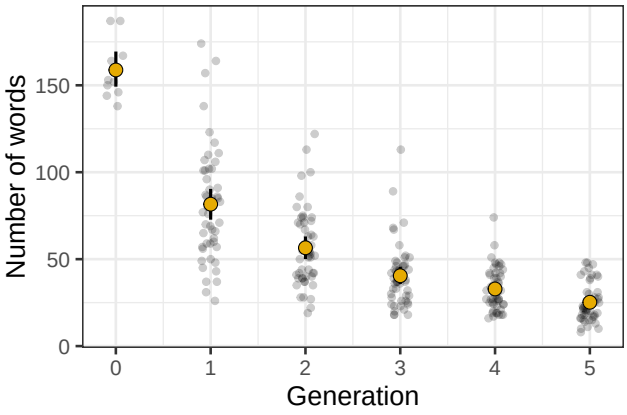

Figure 2: Mean story length in number of words by generation. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories. [mf: would it not be more systematic (and interesting?) to have this also seperately for strong and weak evidence conditions?]

(1) In late December 2017, a couple in Iowa went to check on their beehives. They found a tragic scene: their hives had been overturned and their equipment and facilities had been ransacked. A few weeks later, the police arrested a 12-y.o. and 13-y.o. for the crime. They are charged with multiple offenses, with fines up to $100,000 and up to 10 years in prison, yet will be tried as minors. The trial hasn't happened yet, but they seem guilty.

(2) A 12 and 13 year old were arrested for destroying a beehive, and face up to 10 years of jail time.

Similarity of seeds and reproductions. To assess the similarity of reproductions and their seed stories quantitatively, we computed the Jaccard distance between each reproduction and its generation 0 seed. Jaccard distance captures the amount of overlap between two stories in the following way:

$$D_J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

where X is the reproduction and Y the respective original seed story [mf: should this not better be "where X is the set of words in one story ..."]. In this case, we took words as the basic unit over which distance was computed. Figure 3 shows that D_J increased across generations ($\beta = 0.05$, $SE = 0.00$, $t = 14.17$, $p < 0.0001$). [mf: what about an effect of condition? did we test this? if it doesn't come up significant, we should still report that.] This is not surprising given that as story length decreases, D_J between seed and any of its reproductions necessarily increases. However, D_J increased more

Table 1: Example of a seed story used in Exp. 1.

In late December 2017, a couple in Iowa was checking on their 50 beehives when they discovered a tragic scene. The hives had been overturned and hacked apart, and the equipment had been thrown out of the shed and smashed. This destruction caused the death of about half a million bees and approximately $60,000 in property damage. Nearly three weeks later, police arrested two boys (12 and 13 years old) who, allegedly, were responsible for the damage. The charges against them include criminal mischief, burglary, and offenses to an agricultural animal facility. Since they are still minors, they will be charged in juvenile court where they face up to 10 years in prison and fines of up to $10,000 if convicted.	
(*strong evidence condition*) Police officials explained that the investigation is still in progress, but the evidence so far overwhelmingly speaks to the guilt of the suspects.	(*weak evidence condition*) Police officials explained that the investigation is still in progress, and the evidence so far doesn't warrant rushed conclusions about the guilt of the suspects.

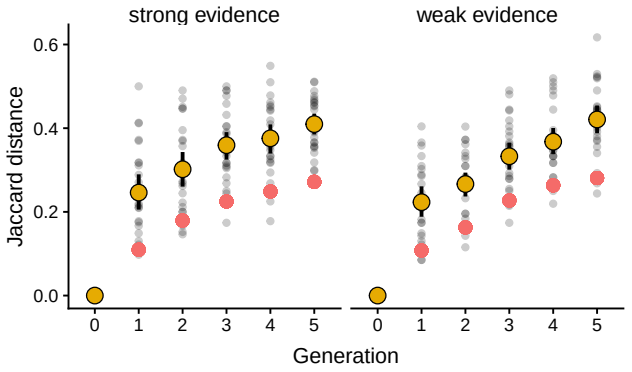

Figure 3: Mean Jaccard distance between seed and reproductions by generation in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories, red dots indicate the lowest possible distance given the mean length of the stories.

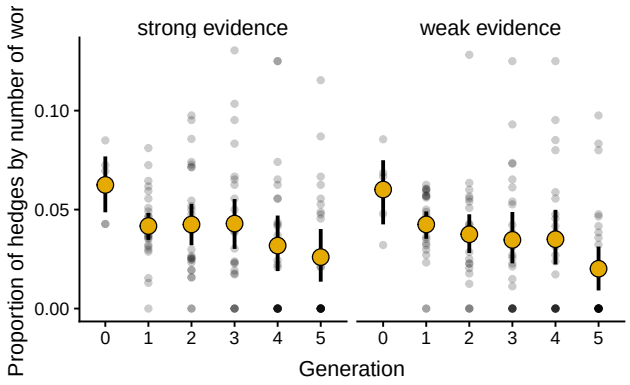

Figure 4: Mean proportion of hedges (by number of words) in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories.

strongly than expected if the difference between stories was only due to the decrease in length [ek: is this true?] [mf: this is tricky. I wouldn't know how to *quickly* check this; first thing that comes to mind is, unsurprisingly, simulation; we could sample random subsets of words corresponding to the mean story lengths and check JD for that; but that's time consuming and I don't think that this is worthwhile currently; I'd rather suggest we formulate this differently, weaker, less prone to criticism if we do not back it up], suggesting that information was lost across generations. This can also be observed qualitatively in the comparison of the representative examples (1) and (2) above.

Proportion of Hedges. As a proxy for vagueness, we extract the number of hedges per story relative to its length. The seed stories were designed to contain various hedges, such as "nearly", "about", "up to" or "allegedly". [ek: Figure ..] shows the proportion of hedges that remain throughout the generations. We can see that this proportion decreases over generations ($\beta = -0.01$, $SE = 0.00$, $t = -4.16$, $p < 0.0001$). The reproductions therefore appear to get less mitigated over iterations.

In summary, our investigation of the corpus so far revealed that the length of the stories, the similarity to the seed story and the proportion of hedges decreases over generations.

Experiment 2: story ratings

In order to assess the extent to which, as a function of the originally provided evidence, the generation of reproduction affects readers' perception of various aspects of the stories we collected judgments from a second group of independent readers of the reproductions. We are particularly interested in aspects related to the uncertainty of presented evidence, and so collected intuitive judgments related to the suspect's perceived guilt and the strength of evidence but also concerning the readers' general attitude towards the author and the story.

Methods

5392 participants were recruited over Amazon Mechanical Turk. Each participant read one story from the 250 story corpus reported in the previous section, and answered twelve

questions about the story (including four attention checks). They indicated their response by moving a slider on a continuous scale. Each question was shown in isolation in a randomized order. Participants spent on average two to three minutes on this experiment and were paid $0.60 ($12-$18 per hour). The story was visible throughout the experiment.

The list of questions asked is provided in (3) to (10). Questions (3) - (7) assessed the extent to which the reader believes the suspect(s) is/are guilty of the alleged crime. Questions (8) - (10) assessed the reader's trust in the author, the extent to which they considered the story to be objectively written, and the extent to which they felt emotionally connected to the story. Overall, participants were asked eight questions of interest and four attention check questions designed to filter out participants who were just clicking through the experiment. The attention checks were counter balanced and asked about the likelihood that the passage is a Greek fairy tale, a Bible quote, contains more than five words, and that the story involves *X*, where X was replaced by a topic which was likely to occur in the story (e.g., *bees* for the story in Table 1)

(3) How strong is the evidence for the suspect's / suspects' guilt?

(4) How likely is it that the suspect is / the suspects in the crime are guilty?

(5) How likely is a conviction of the suspect(s) in the crime?

(6) How justified would a conviction of the suspect(s) in the crime be?

(7) How much does the author believe that the suspect is guilty?

(8) How much do you trust the author?

(9) How objectively / subjectively written is the story?

(10) How affected do you feel by the story?

Results

We excluded 12 participants because they completed the study multiple times and another 535 because they failed at least two of the attention check questions. This leaves us with 4573 participants (84.8% of the original set). After exclusions, each reproduction received on average 17 ratings, ranging from 9 to 22 and two outliers with 27 and 38 ratings[1]. The original seed stories received between 25 and 31 ratings.

Mean slider ratings are shown in Figure 5. Qualitatively, there is a strong effect of condition in the guilt related measures, i.e., *strength of evidence*, *suspect guilt*, *suspect conviction*, *suspect conviction justified* and *author's belief in guilt*. In the responses to these measures, the ratings for the strong evidence condition are declining over generations, but consistently higher than for the weak evidence condition. In other words, the different guilt measures retain higher guilt judgments than the weak condition over all generations. However, the judgments of the weak conditions pattern differently. In contrast to a decline in the ratings over generations, the judgments in the weak condition stay constant or increase. This narrows the distance between the judgments in the strong and weak conditions. [ek: In sum,] the stories for the two different conditions become more similar over generations with respect to the guilt measures.

The difference between the conditions is far smaller (if existent at all) in the responses to the reader's *trust in the author*, the *subjectivity of the story* and the *reader's emotional engagement*. *Trust in the author* and the *reader's emotional engagement* decline over generations of reproductions, independent of the condition. Interestingly, the measure on how subjectively the story is written does not change over generations and remains on the "objective" side of the scale.

[ek: maybe make a note that there is no sign of convergence to .5]

The judgments were analyzed using linear mixed effects models. For each question, slider rating was predicted from fixed effects of generation (reference level: 0), condition (reference level: strong), and their interaction. The model also included random by-story intercepts. An overview of the results is shown in Table 2. [mf: the reader needs help reading this table; what is where how?]

In the qualitative analysis, we have observed a strong condition effect in the guilt measures. This is reflected in the model results. In the non-guilt measures, story subjectivity shows a significant effect on condition. The trust in the author and the reader's emotional engagement appear to be independent of the condition and therefore cannot explain the difference that we see in the guilt related measures. Even though the subjectivity shows an effect of condition, it is too small [ek: rephrase this...] to explain these effects in the guilt measures either. [ek: We also know that the length and Jaccard distance (and maybe hedges?) doesn't differ in the two conditions, so...] This suggests that the strong effects of condition in the guilt measures [ek: are about some more subtle changes in the actual language and content of the stories].

With respect to the guilt measures, the quantitative analysis suggests clear effects of generation in the strength of the evidence, the likelihood of conviction, and whether the author believes in the suspect's guilt. Those three guilt measures show a very similar pattern in their main effect structure. A simple effects analysis further supports this grouping intuition, since none of them show an effect in the interaction between *weak* evidence and generation, but all of them show an effect in the interaction between *strong* evidence and generation. This group stands in contrast to the other two guilt measures, asking about the suspect's guilt and whether a conviction would be justified. These two do not show a main effect in generation and only a marginal effect in the simple effects analysis for the interaction between the strong evidence condition and generation.

[1]The outliers are due to a mistake in the recruitment process.

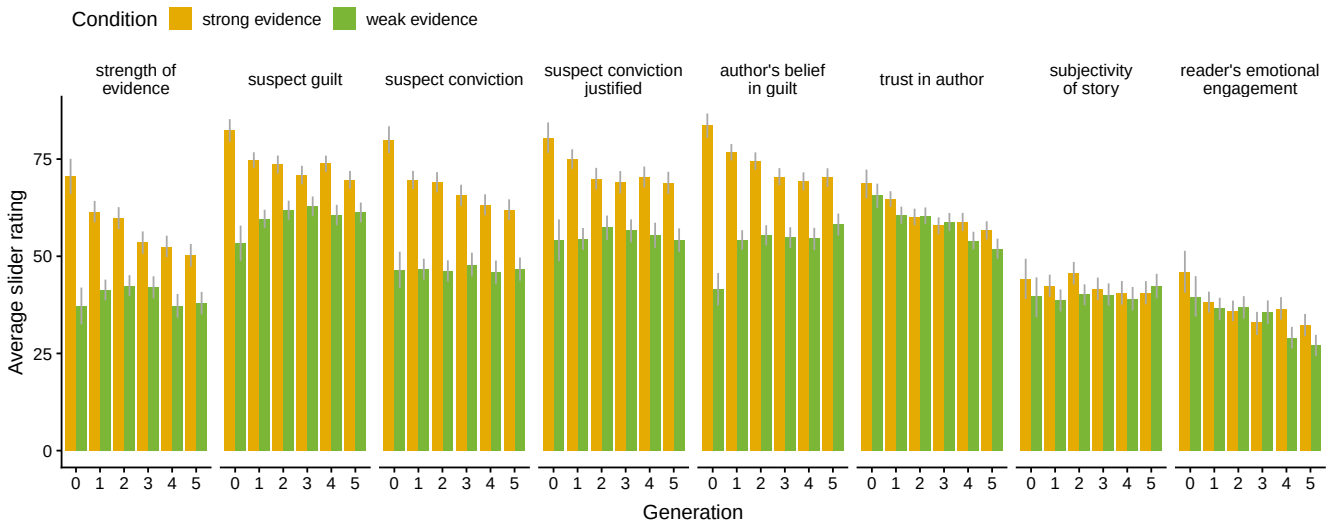

Figure 5: Mean ratings in strong (orange) and weak (green) evidence condition for each dimension (facets).

We want to emphasize here that even though we intuitively expect the strength of evidence to inform the suspect's guilt judgment, we do not see the same effect structure in these two cases. Qualitatively, the development in the strong evidence condition suggests that the *strength of evidence* deteriorates more than the *suspect* guilt judgments. In the weak evidence condition, this pattern appears similar, in that the judgments of guilt seem to increase more than those of evidence.

Trust in author and the *reader's emotional engagement* did not show a main effect of condition, but of generation. The further it gets from the original seed, the less trust is put in the author and the less emotional engagement the readers feel.

Overall, we find a strong effect of condition on the guilt measures that is not reflected in this extent by the trust in the author, subjectivity of the story or the reader's emotional engagement.

Conclusion

[ek: insert intro sentence]

First we constructed a corpus which consists of 10 original seed stories and 250 reproductions. Half of the seed stories suggested strong and the other half weak evidence that speaks for the suspect's guilt. Our corpus collection replicates previously found effects of a decrement in length over generations of reproductions ((Bartlett, 1932)). Furthermore, the stories become less similar to the original seed story and the proportional number of hedges decreases.

In a second study, we obtained subjective ratings for each story regarding the question of guilt, conviction, the author of the story, the subjectivity of the story and the reader's emotional engagement. Our results suggest **that, for one, the subtle manipulation of varying evidential strength in the original seed stories did have a lasting effect on reproductions lasting across several generations.** We also find effects of generation. It is plausible that trust in the author and the reader's emotional engagement decrease because the reproductions become shorter with each iteration. It is less likely that this explanation holds for the generation effects in *strength of evidence*, likelihood of *suspect conviction* and the *author's belief in the suspect's guilt*. In the ratings for these questions we see a stronger difference between conditions that the changes in length, similarity and the proportion of hedges could not account for. [ek: Therefore, it must be something about the content of the stories itself that changes.]

Most interestingly, it seems that readers became, over generations, more inclined to endorse the guilt of the suspect(s) in the weak evidence condition, despite no similar increase in ratings of the strength of evidence. This may be due to a baseline or prior level of conviction that a suspect is guilty. And it would suggest that, indeed, subtle evidential information and uncertainty about evidence is washed out over several linguistic transmissions.

[mf: I have tried to give this some top-level interpretation, but I have to admit that I'm pretty lost when it comes to interpreting the results.]

Discussion

[ek: discuss differences between stories with in- and outgroup effects for smuggler and professor] [ek: next steps?]

[mf: it is not common to have a discussion section after the conclusions. let's put it all in one section]

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Hills, T. T. (in press). *The dark side of information proliferation.* (Journal: Perspectives on Psychological Science)

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture*, *4*(1), 1–24.

	condition			generation			condition*generation			simple effects		
	β	SE	p	β	SE	p	β	SE	p	weak	str*gen	we*gen
evidence	-23.25	4.09	<0.0001***	-3.42	0.89	<0.001***	2.59	1.26	<0.05*	***	***	
suspect guilt	-17.28	3.40	<0.0001***	-1.34	0.74	<0.08	1.90	1.05	<0.08	***	.	
conviction	-27.01	4.15	<0.0001***	-2.79	0.90	<0.01**	2.74	1.28	<0.05*	***	**	
convicJustified	-19.02	4.35	<0.0001***	-1.69	0.95	<0.08	1.43	1.34	<0.29	***	.	
author belief	-27.53	3.72	<0.0001***	-2.14	0.81	<0.01**	3.42	1.15	<0.01**	***	**	
author trust	-0.82	2.25	<0.72	-1.94	0.49	<0.001***	-0.54	0.70	<0.44		***	***
story subjectivity	-6.12	2.21	<0.01**	-0.86	0.49	<0.08	1.40	0.69	<0.05*	**	.	
reader emotion	0.85	2.99	<0.78	-1.49	0.65	<0.05*	-1.11	0.92	<0.24	*	***	

Table 2: Model output for each fixed effect (condition, generation, and their interaction) for each rated question (rows). [jd: simple effects results should not be reported in this table – this is just here for us, right?][ek: yes]

Van Prooijen, J.-W. (2006). Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality and Social Psychology Bulletin*, *32*(6), 715–726.