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

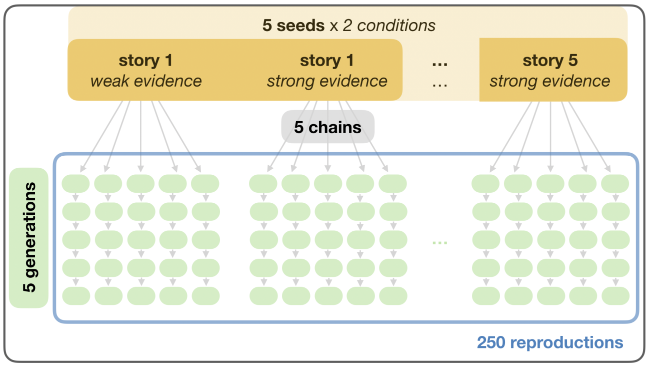

Figure 1: Overview of corpus of stories collected in Exp. 1.

In recent years, the transmission chain method received a revival in the scientific community. (Mesoudi & 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: transmission chain method]

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 violation that occurred, the authorities' determination of and search for the perpetrator(s), and the possible punishment the suspects 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. On each trial, participants first read a story. They were told to click the 'Continue' button when they were confident 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 each with 5 complete chains each, see Figure 1).

While the linguistic changes across generations as a function of the original evidence condition merit their own detailed analysis, we focus here on reporting only a few general features of the collected corpus, which we will subsequently use as predictors in the analyses of Exp. 2 below.

Story length. As shown in Figure 2, the length of the stories 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.

(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.

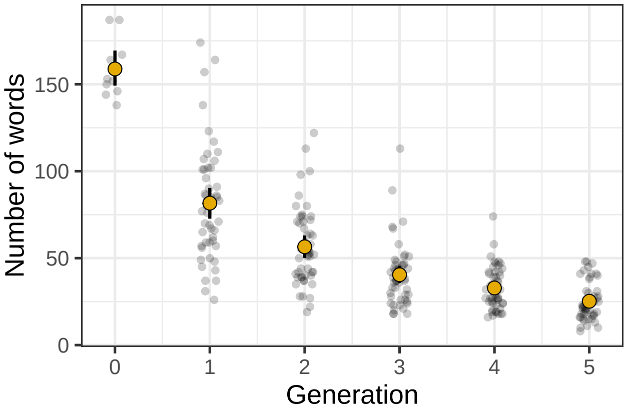

Figure 2: Mean story length in number of words by generation. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories.

Similarity of seeds and reproductions. Of interest is the extent to which stories retain the gist or deviate from it. 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. In this case, we took words as the basic unit over which distance was computed. Figure 3 shows that D_J increased across generations ($\beta = 0.05$, $SE = 0.00$, $t = 14.17$, $p < 0.0001$). 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 strongly than expected if the difference between stories was only due to the decrease in length, suggesting that information was lost across generations. This can also be observed qualitatively in the comparison of the representative examples (1) and (2) above.

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 crowd-sourced judgments about the suspect's perceived

Table 1: Example seed story in Exp. 1. [ek: I don’t know how to wrap this.]

<table>
<tr><td colspan="2">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.</td></tr>
<tr><td>Police officials explained that the investigation is still in progress, but the evidence so far overwhelmingly speaks to the guilt of the suspects. (<i>strong evidence condition</i>)</td><td>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. (<i>weak evidence condition</i>)</td></tr>
</table>

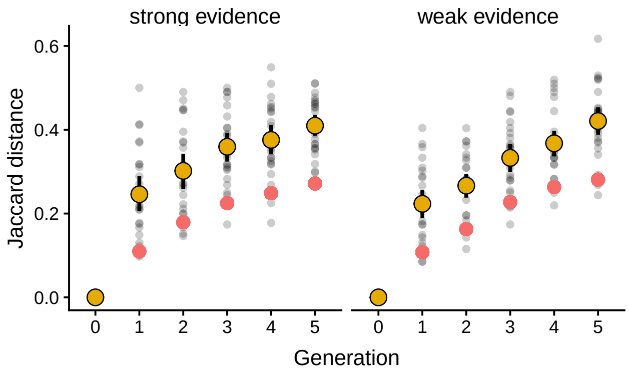

Figure 3: Mean Jaccard distance between seed and reproductions by generation in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories, red dots indicate the lowest possible distance given the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

guilt, the evidence for the crime, the author, and the reader themselves.

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 (4). Questions XX - XX assessed the extent to which the reader believes the suspect(s) is/are guilty of the alleged crime. Questions XX - XX 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. [jd: describe the attention checks?]

[jd: fill out the list of questions]

(3) bla

...

(4) bla

Results

[ek: ...]
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 ([jd: provide range?]).

Mean ratings are shown in Figure 4. [jd: describe qualitatively. first the guilt measures; then the author/story/reader related measures]

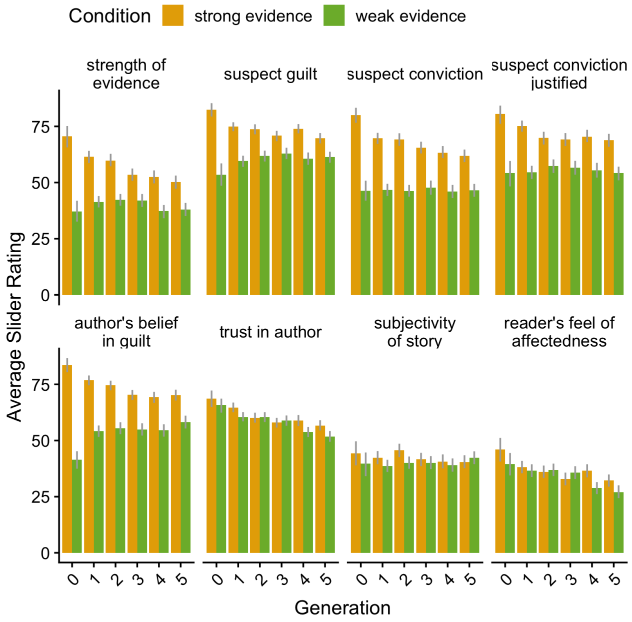

Figure 4: Mean ratings in strong (orange) and weak (green) evidence condition for each dimension (facets). [jd: elisa, put these all in one row, make the x-axis labels not slanted, and make it a two-column figure. again, make it pdf and make sure in axis titles only the first word is capitalized][ek: judith: I don't know what you mean here with all in one row but only two columns?]

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.

Conclusion

Discussion

[ek: discuss differences between stories iwth in- and out-group effects for smuggler and professor]

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.

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.

		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]

		condition			distance			condition*distance	
	β	SE	p	β	SE	p	β	SE	p
evidence	-24.90	5.24	<0.0001***	-36.49	10.92	<0.001***	27.75	15.41	<0.08
suspect committedCrime	-20.12	4.32	<0.0001***	-14.94	9.00	<0.10	26.15	12.71	<0.05*
suspect conviction	-31.87	5.26	<0.0001***	-36.83	10.96	<0.001***	39.48	15.47	<0.05*
suspect convictionJustified	-21.181	5.54	<0.001***	-21.42	11.55	<0.07	19.35	16.30	<0.24
author belief	-29.90	4.74	<0.0001***	-9.01	9.87	<0.37	39.02	13.94	<0.01**
author trust	-1.19	2.83	<0.68	-24.73	5.91	<0.001***	-4.93	8.36	<0.56
story subjectivity	-6.12	2.77	<0.05*	-5.05	5.79	<0.39	12.77	8.22	<0.13
reader emotion	0.54	3.70	<0.89	-25.34	7.72	<0.01**	-10.44	10.93	<0.35

Table 3: lmer(suspectconvictionJustified sim * condition + (1—storyreproduction), data=dfmodel); high correlation of fixed effects

	condition			hedgesprop			condition*hedgesprop		
	β	SE	p	β	SE	p	β	SE	p
evidence	-15.74	1.95	<0.0001***	101.20	56.55	<0.08	-119.12	82.37	<0.15
suspect committedCrime	-11.93	1.59	<0.0001***	43.01	45.98	<0.36	-118.58	66.97	<0.08
suspect conviction	-19.10	1.96	<0.0001***	102.66	56.65	<0.08	-132.10	82.50	<0.12
suspect convictionJustified	-14.94	2.04	<0.0001***	30.91	59.10	<0.7	-70.91	86.08	<0.42
author belief	-17.91	1.75	<0.0001***	54.69	50.54	<0.29	-188.17	73.61	<0.05*
author trust	-2.16	1.13	<0.06	46.60	32.70	<0.16	27.80	47.74	<0.57
story subjectivity	-2.22	1.06	<0.05*	6.10	30.61	<0.85	-45.08	44.85	<0.32
reader emotion	-2.25	1.46	<0.13	-7.18	42.26	<0.87	49.49	61.67	<0.43

Table 4: lmer(suspectconvictionJustified hedgesprop * condition + (1—storyreproduction), data=dfmodel); hedges is centered; hedges = c(”allegedly”, ”possibly”, ”maybe”, ”probably”, ”if”, ”around”, ”over”, ”nearly”, ”almost”, ”approximately”, ”vaguely”, ”up to”, ”roughly”, ”mainly”, ”kind of”, ”sort of”, ”kinda”, ”sorta”, ”about”, ”supposedly”, ”seem”, ”tend”, ”look like”, ”looks like”, ”appear to be”, ”think”, ”believe”, ”doubt”, ”be sure”, ”indicate”, ”suggest”, ”assume”, ”might”, ”perhaps”, ”possibility”)