Conversations Gone Awry

Detecting Early Signs of Conversational Failure

Justine Zhang, **Jonathan P. Chang**, Cristian Danescu-Niculescu-Mizil,
Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain

To be presented at ACL 2018 (July 15-20, Melbourne, Australia)

Paper, code, and data available at http://www.cs.cornell.edu/~cristian/Conversations_gone_awry.html

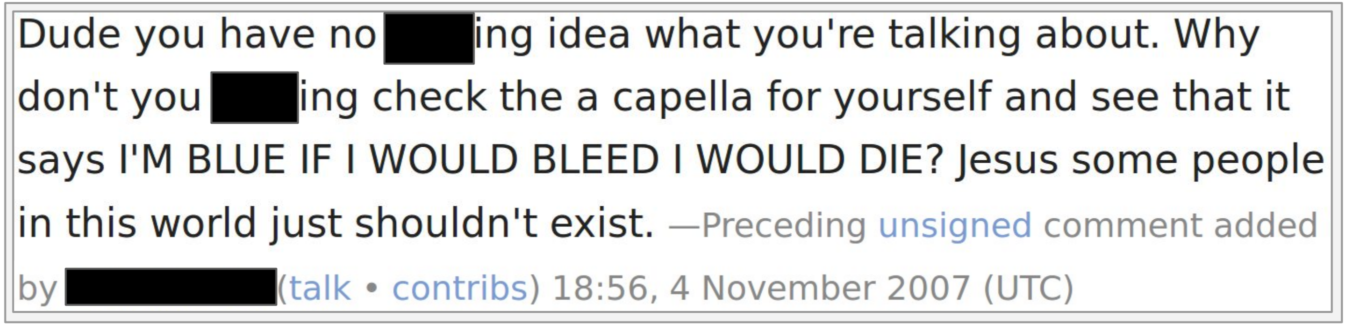
Motivation

1999: “The Internet is becoming the town square for the global village of tomorrow” - Bill Gates

Motivation

1999: “The Internet is becoming the town square for the global village of tomorrow” - Bill Gates

Present Day:

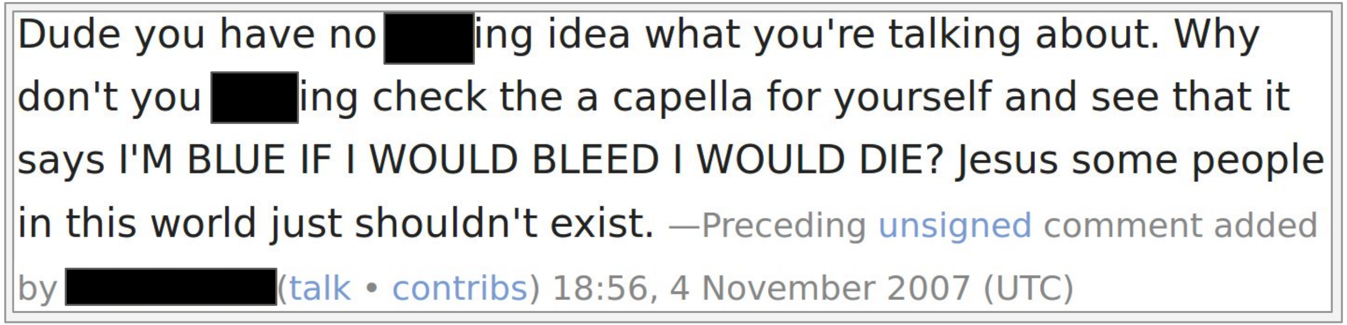


Dude you have no [redacted]ing idea what you're talking about. Why don't you [redacted]ing check the a capella for yourself and see that it says I'M BLUE IF I WOULD BLEED I WOULD DIE? Jesus some people in this world just shouldn't exist. —Preceding unsigned comment added by [redacted](talk • contribs) 18:56, 4 November 2007 (UTC)

Motivation

1999: “The Internet is becoming the town square for the global village of tomorrow” - Bill Gates

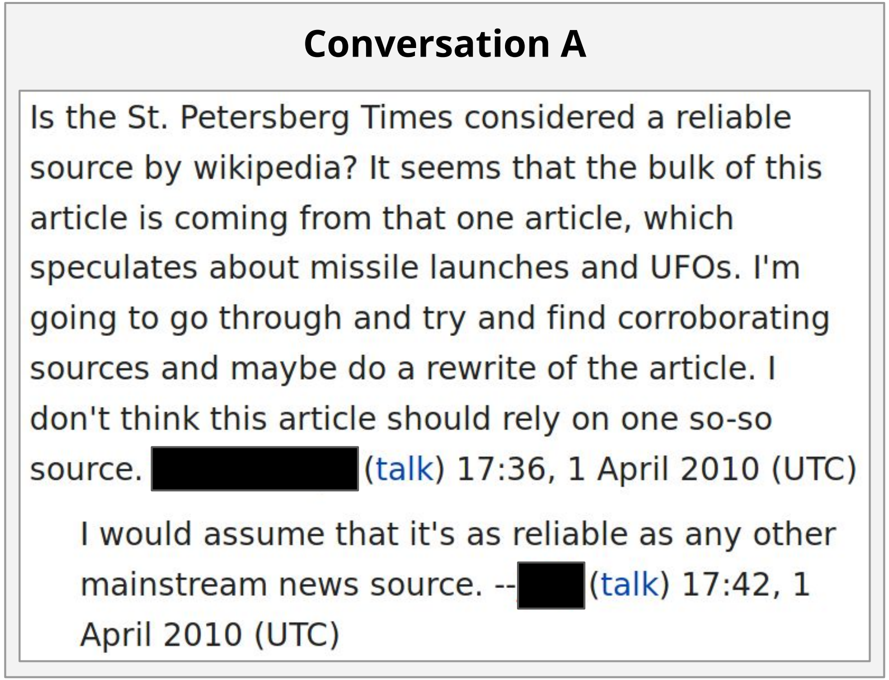
Present Day:



Dude you have no [redacted]ing idea what you're talking about. Why don't you [redacted]ing check the a capella for yourself and see that it says I'M BLUE IF I WOULD BLEED I WOULD DIE? Jesus some people in this world just shouldn't exist. —Preceding unsigned comment added by [redacted](talk • contribs) 18:56, 4 November 2007 (UTC)

What makes civil conversations turn awry?

Conversations Going Awry: An Example



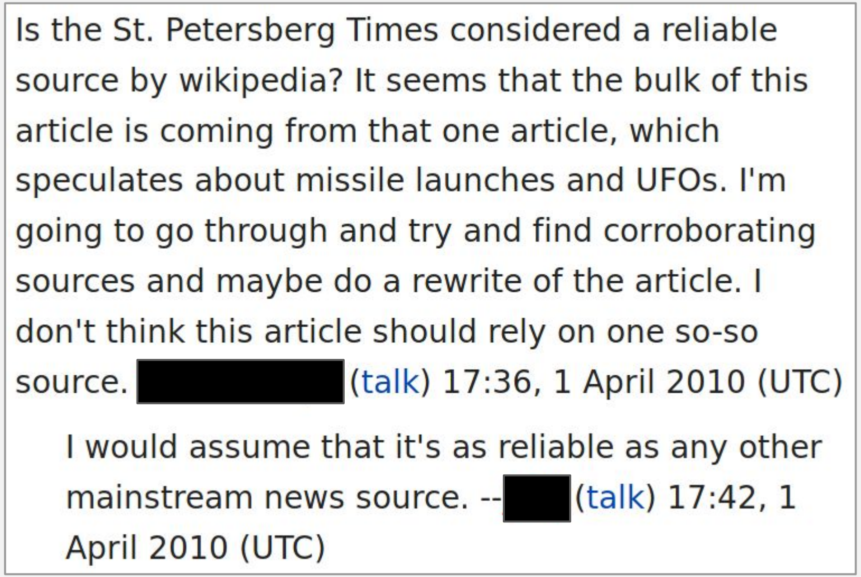
Conversation A

Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversations Going Awry: An Example

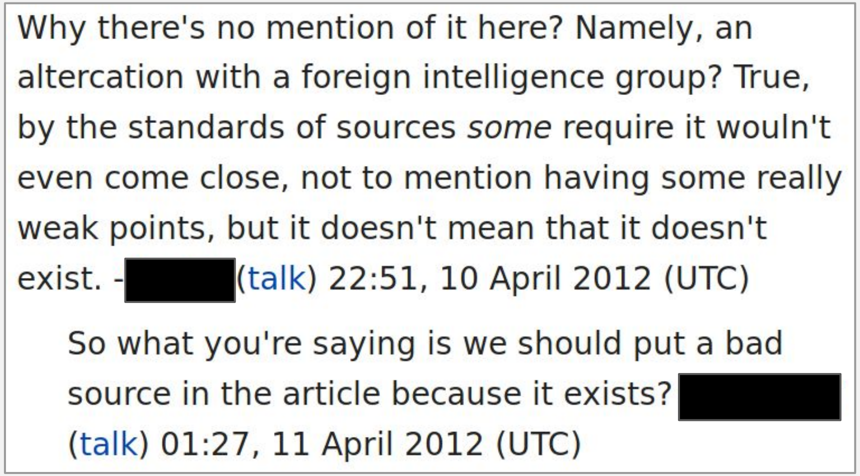
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B

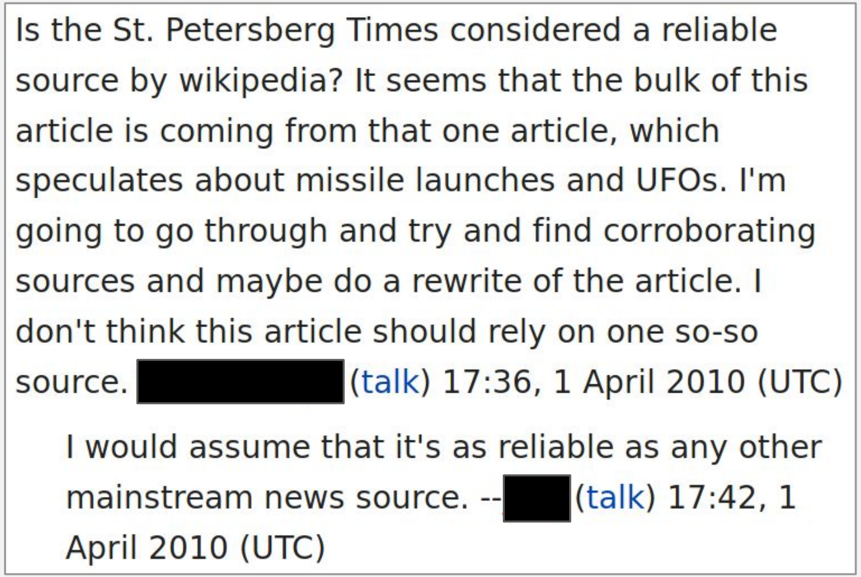


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Conversations Going Awry: An Example

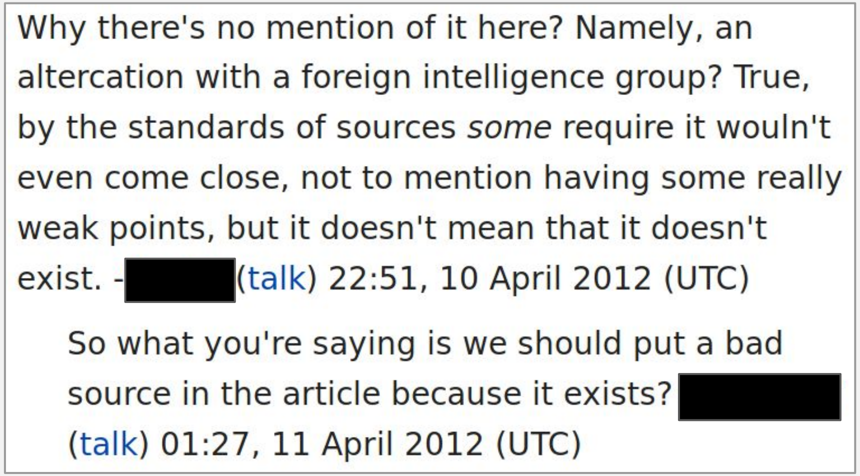
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



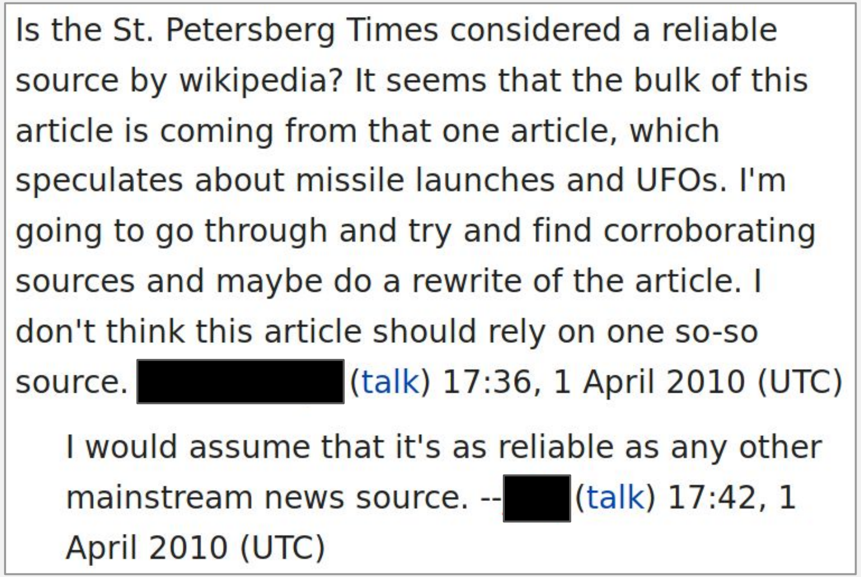
Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Which one leads to: “Wow, you’re coming off as a total d**k...what the hell is wrong with you?”

Conversations Going Awry: An Example

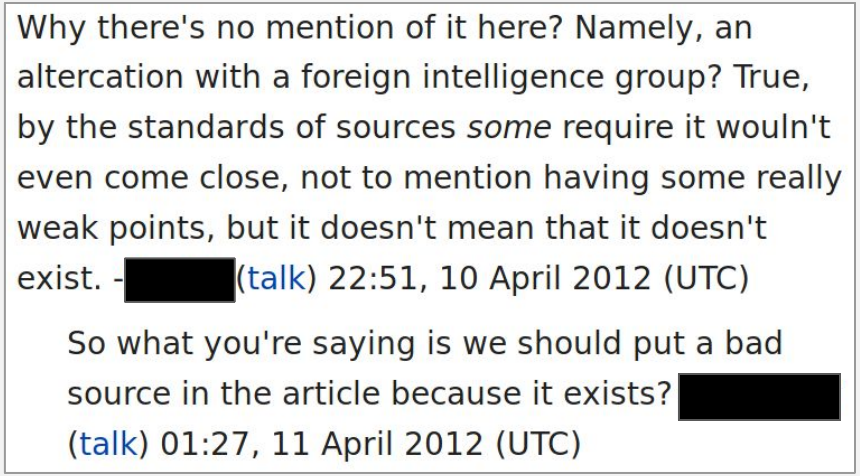
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Which one leads to: “Wow, you’re coming off as a total d**k...what the hell is wrong with you?”

More examples (quiz): http://awry.infosci.cornell.edu/

Capturing Human Intuition

We *seem* to have some intuition for when things are going bad

- Human accuracy is 72% - more on this later

We would like to reconstruct some of this intuition

- Contrast with prior work: *predict* toxicity rather than *detecting* it after the fact (Cheng et al., 2017; Wulczyn et al., 2017)

Two high level challenges:

1. Find cases of conversations “going awry”
2. Encode intuitive signs in some concrete way

Pitfalls to Avoid

Confounding toxicity with disagreement

- Civil disagreement is healthy! (Coser, 1956; De Dreu and Weingart, 2003)

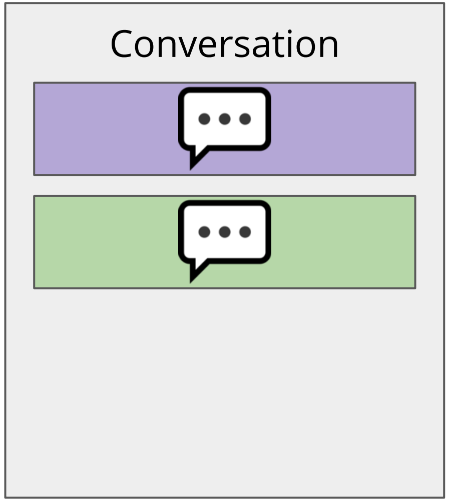
Getting too topic-specific

- Political conversations are more likely to turn toxic – but this doesn't tell us anything about the nature of conversation
- Definitely *don't* want to end up only flagging sensitive topics!

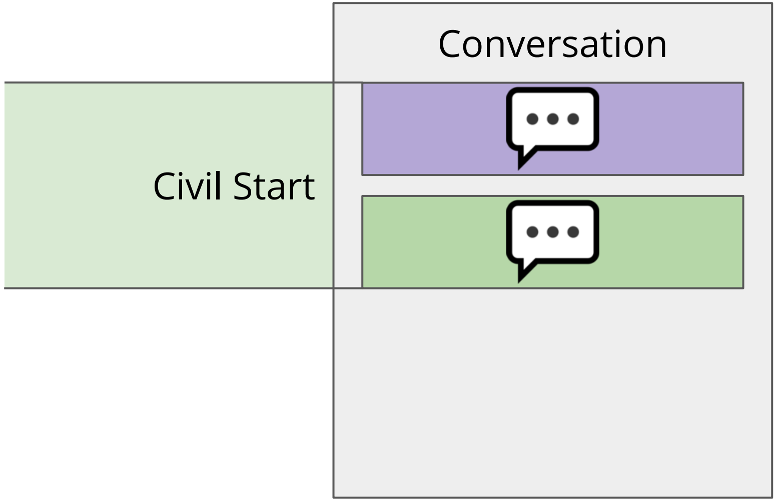
Finding Conversations Gone Awry

What Are We Looking For?

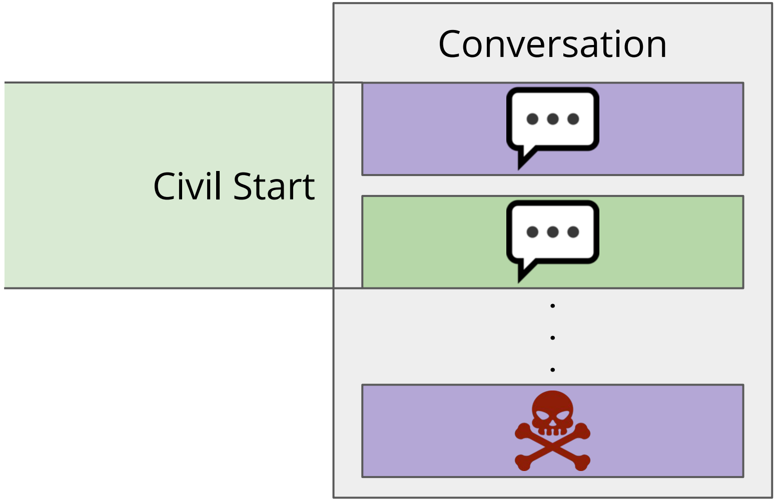
What Are We Looking For?



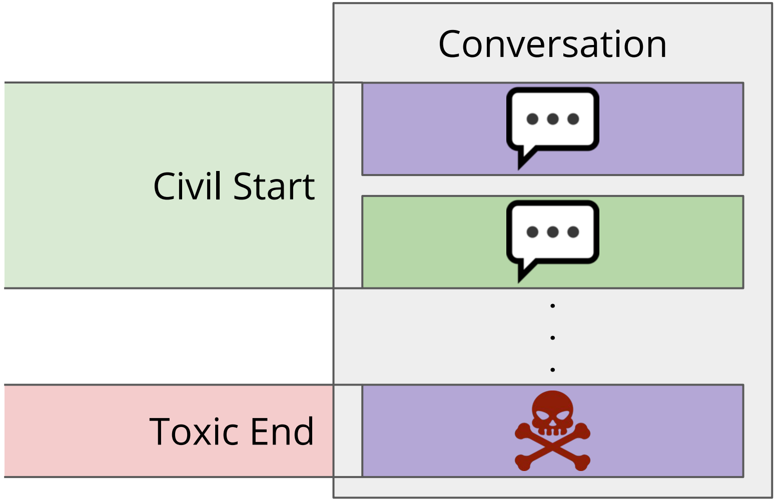
What Are We Looking For?



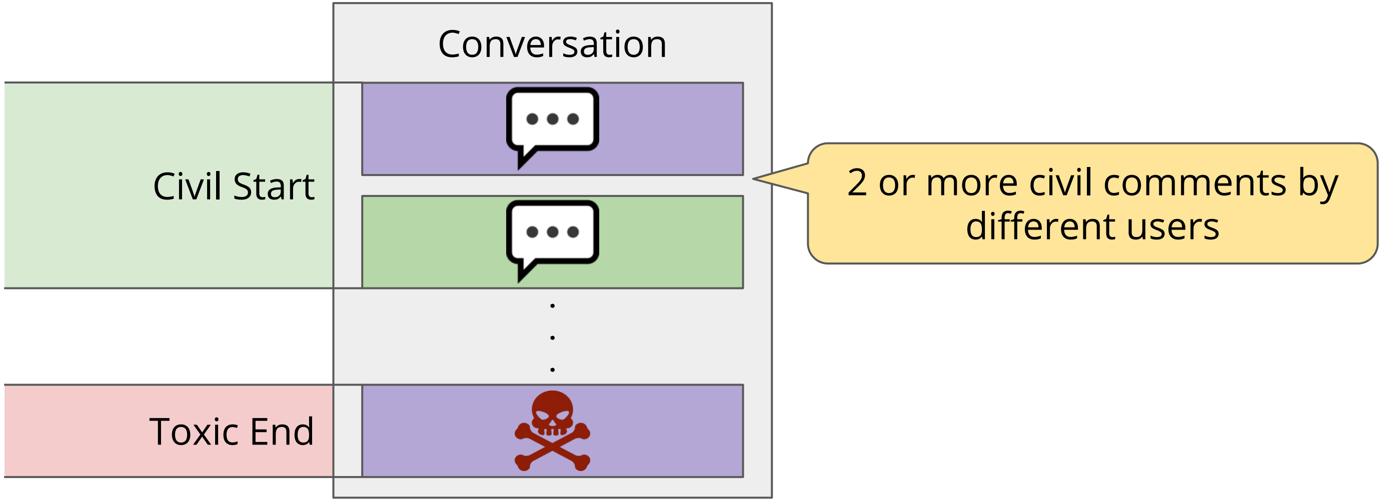
What Are We Looking For?



What Are We Looking For?

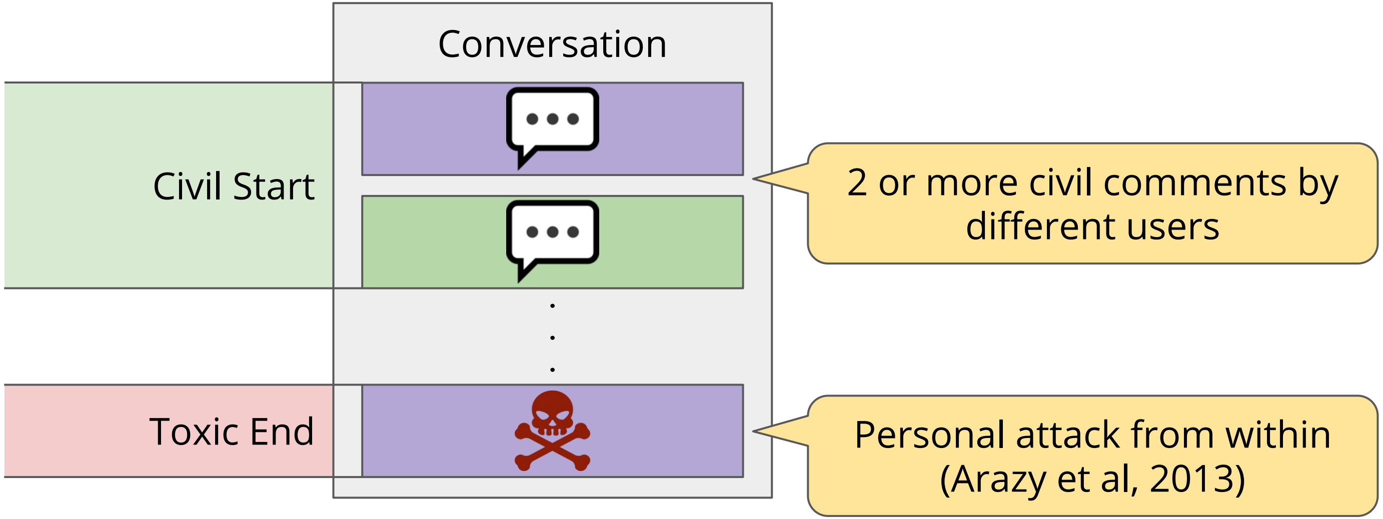


What Are We Looking For?



2 or more civil comments by different users

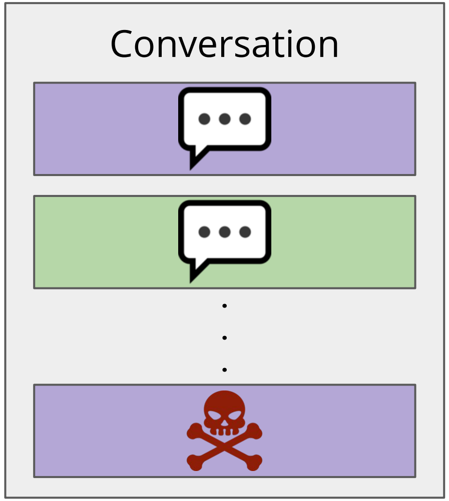
What Are We Looking For?



2 or more civil comments by different users

Personal attack from within (Arazy et al, 2013)

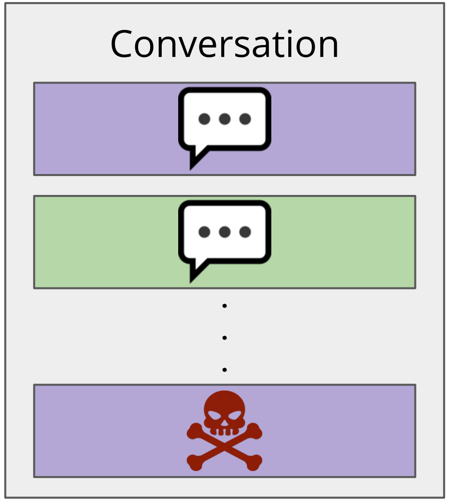
What Are We Looking For?



~ 50 million conversations

Raw data

What Are We Looking For?

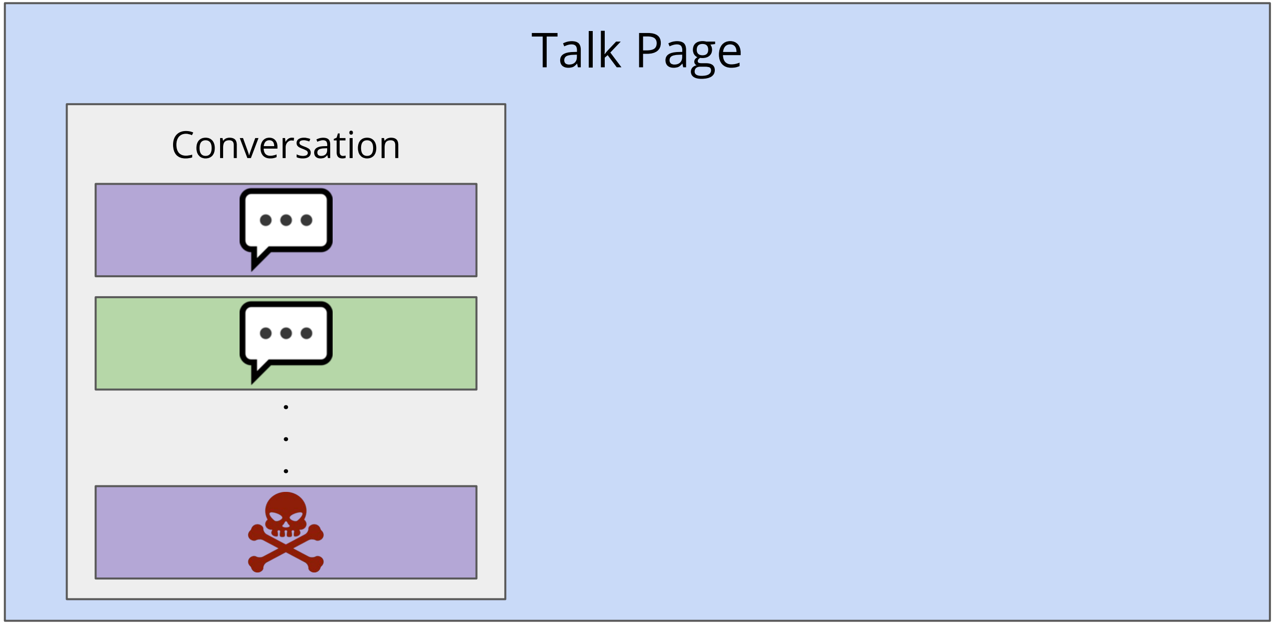


~ 50 million conversations
Raw data

→

~3,000 toxic candidates
Automated pre-filtering

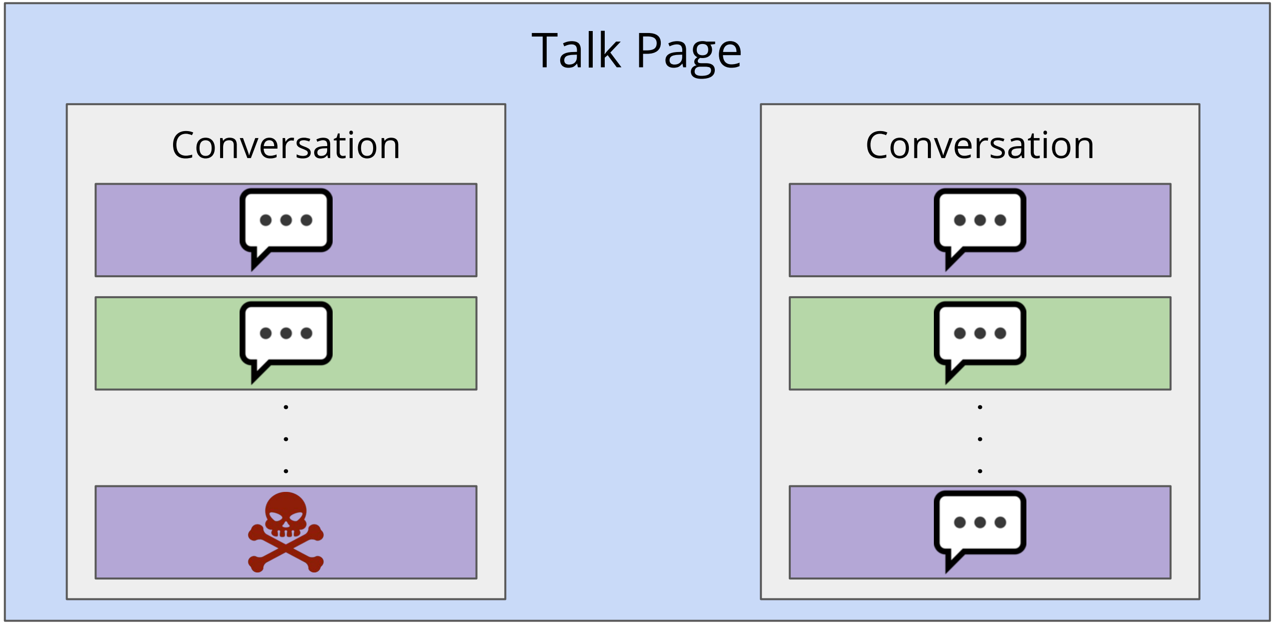
What Are We Looking For?



~ 50 million conversations → ~3,000 toxic candidates

Raw data → Automated pre-filtering

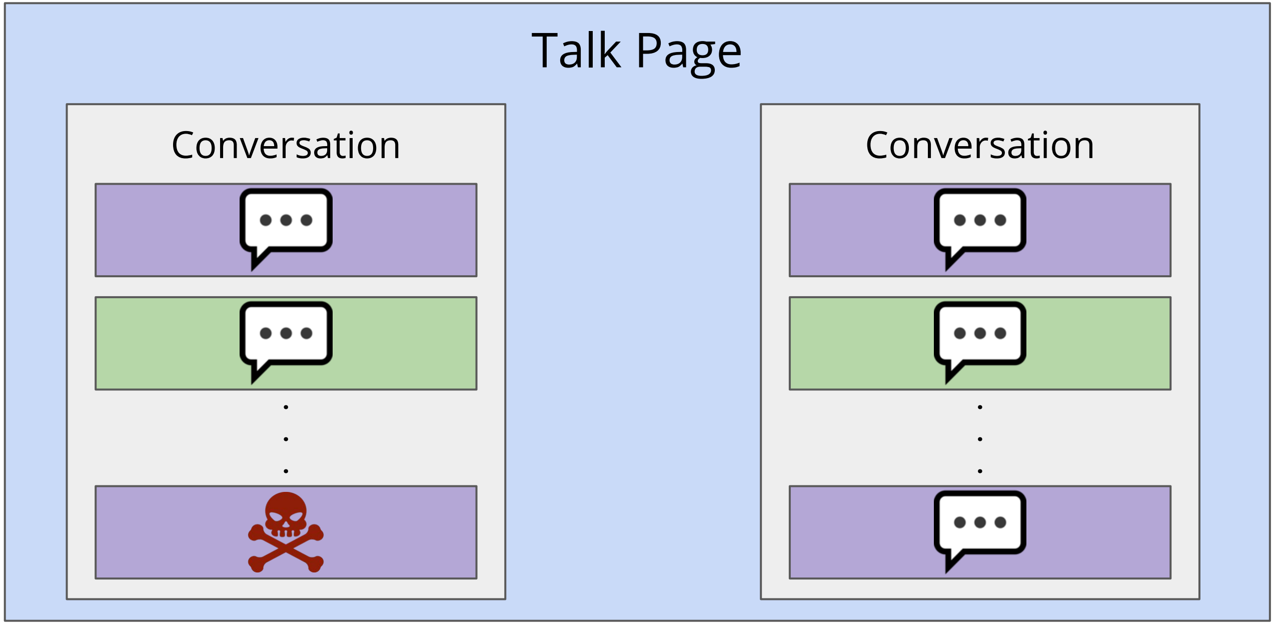
What Are We Looking For?



~ 50 million conversations → ~3,000 toxic candidates

Raw data → Automated pre-filtering

What Are We Looking For?



~ 50 million conversations
Raw data

→

~3,000 toxic candidates
Automated pre-filtering

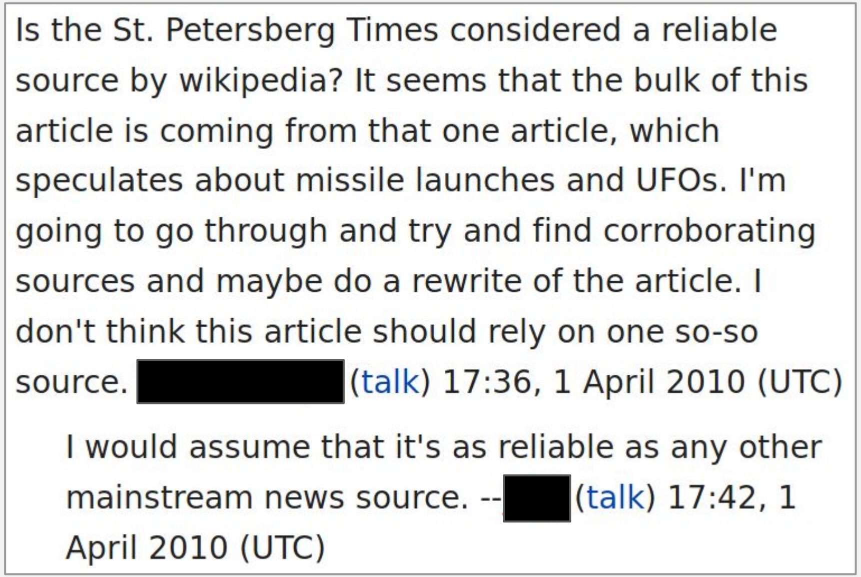
→

635 pairs
Human-validated set

Recovering Human Intuition

Back to our example...

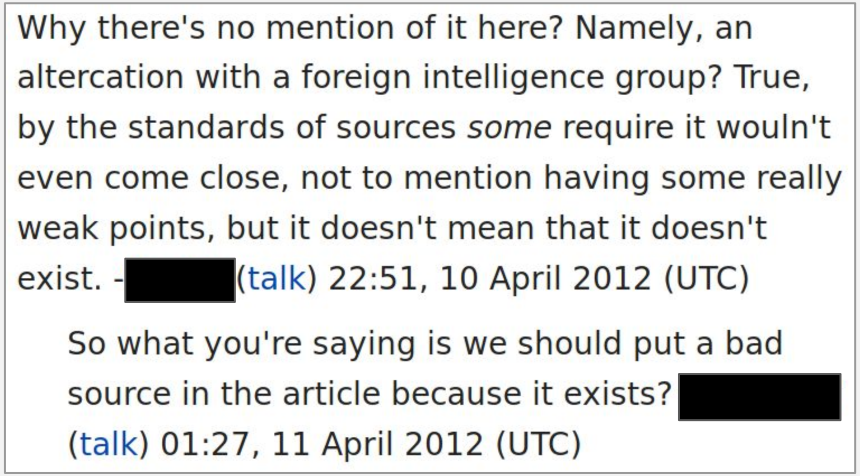
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B

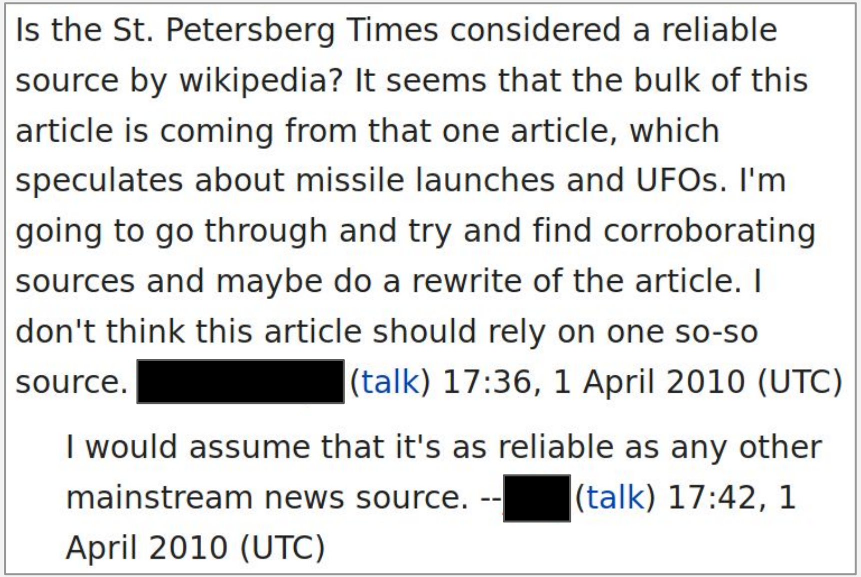


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Back to our example...

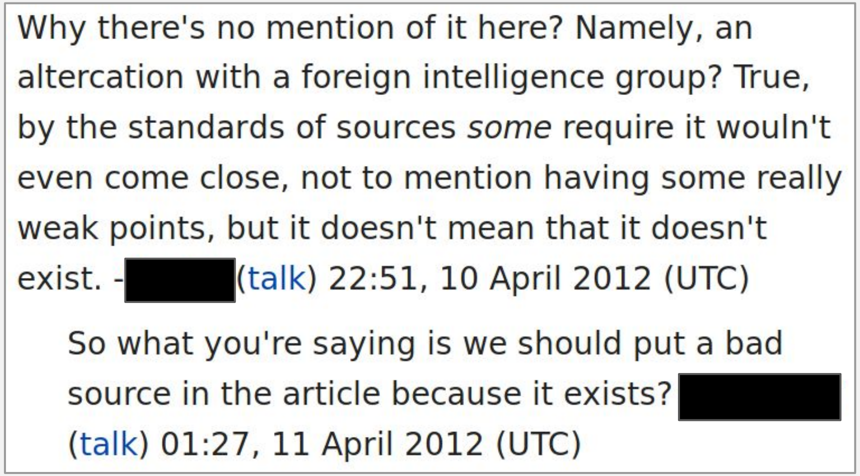
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



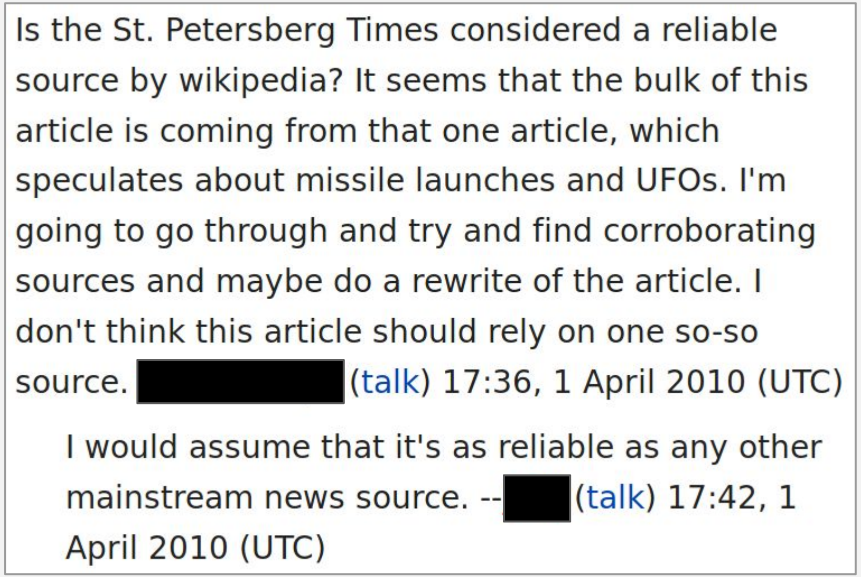
Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

How did we decide?

Back to our example...

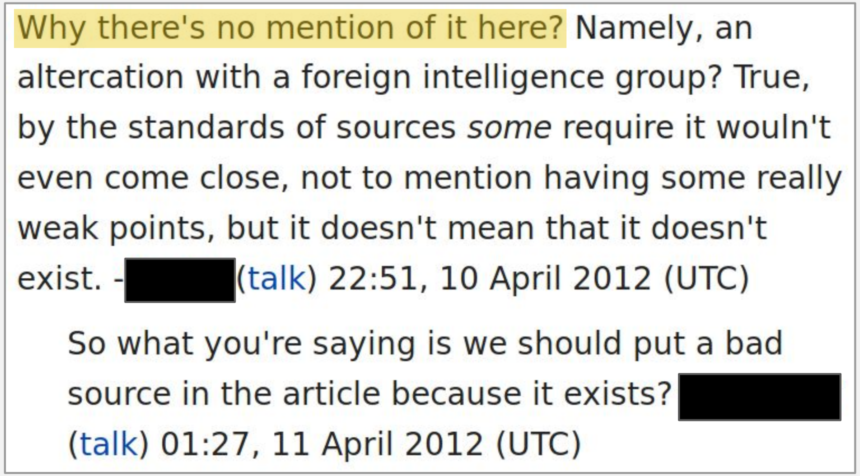
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B

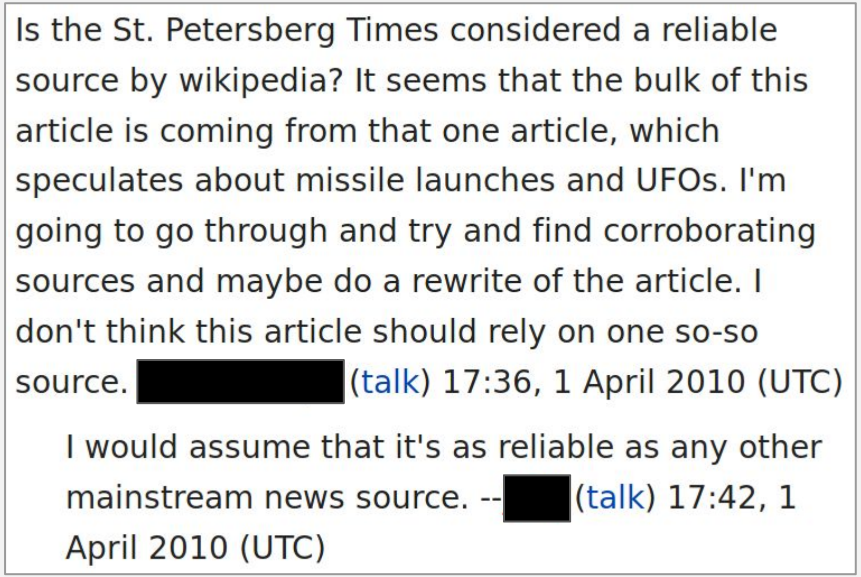


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Back to our example...

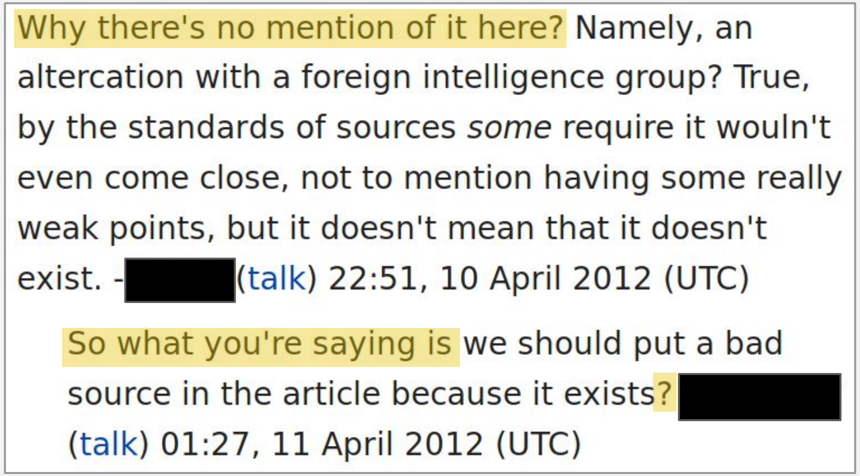
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B

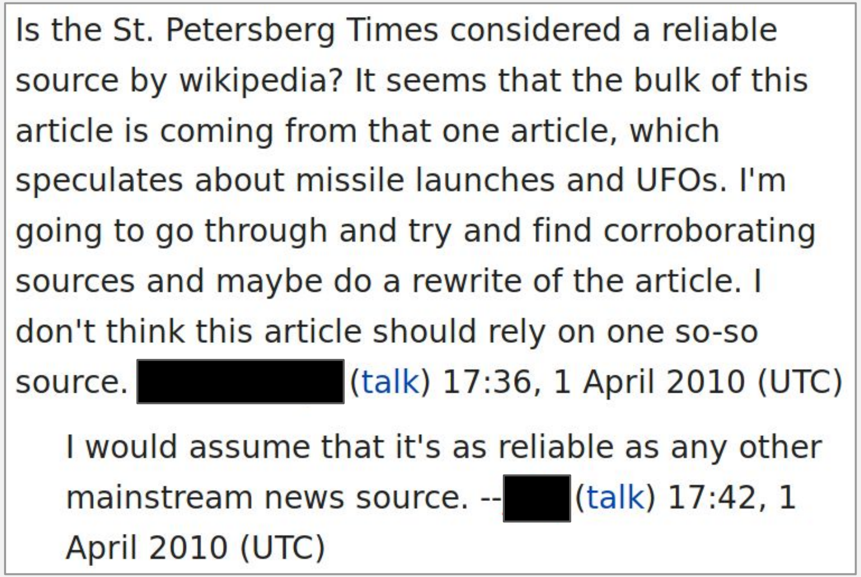


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Back to our example...

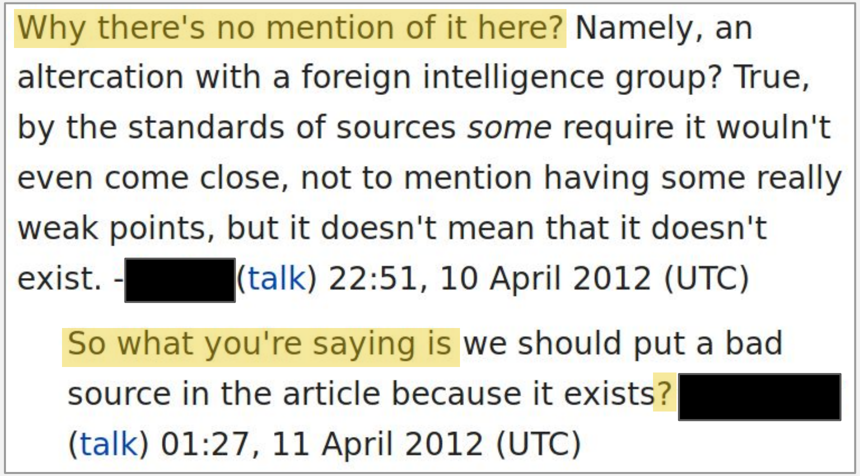
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



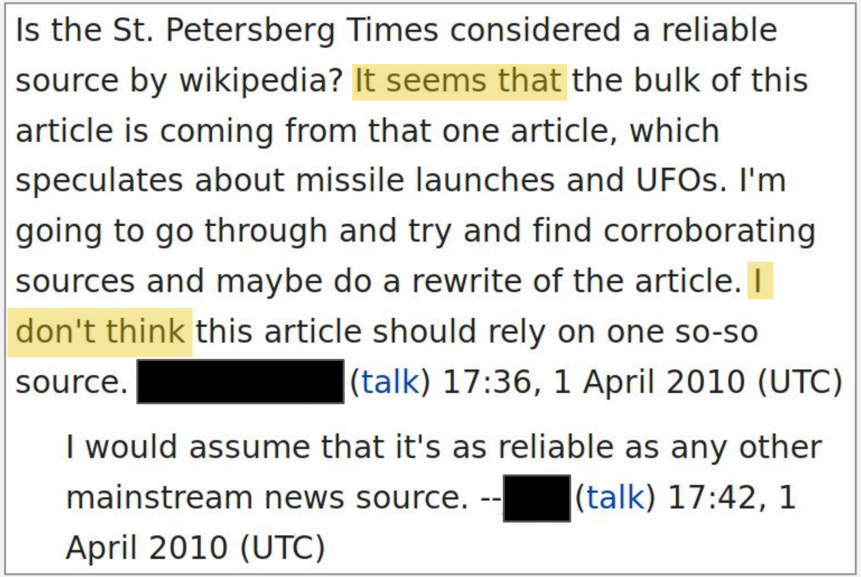
Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Direct questioning

Back to our example...

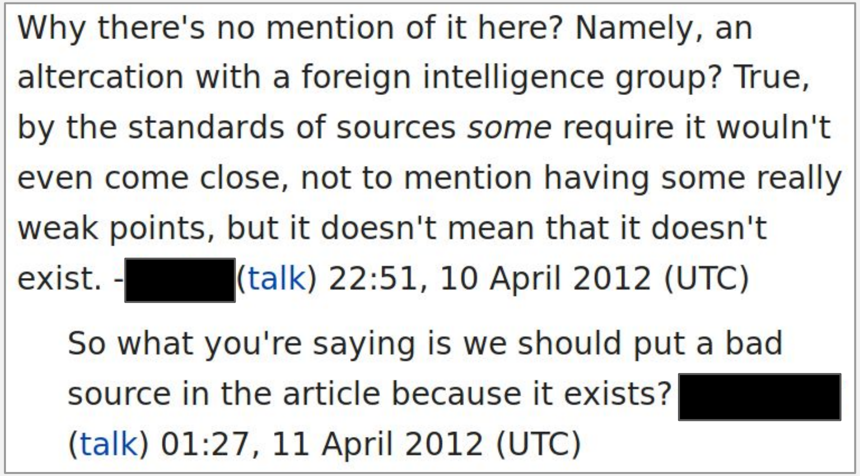
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



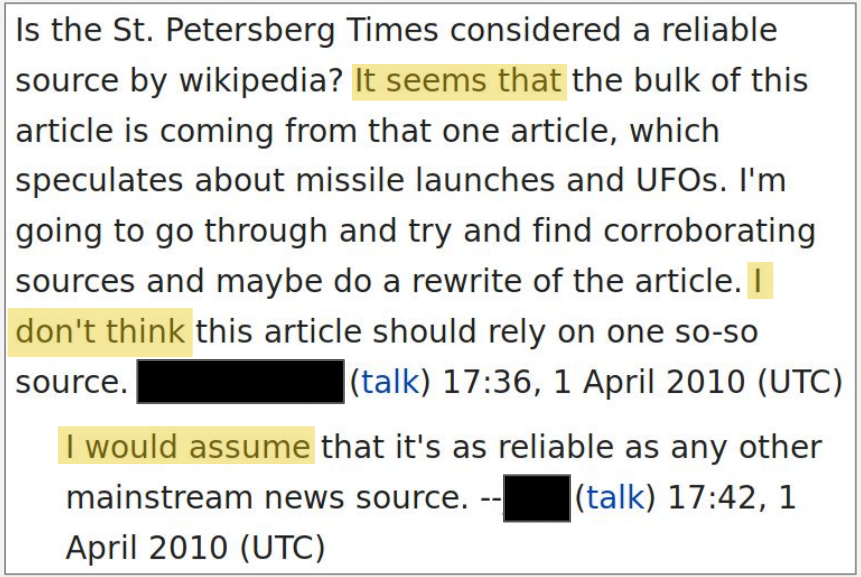
Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Direct questioning

Back to our example...

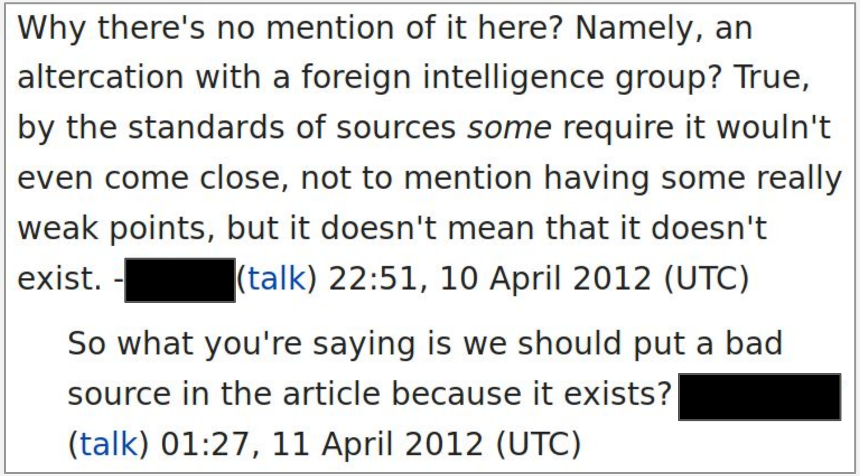
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



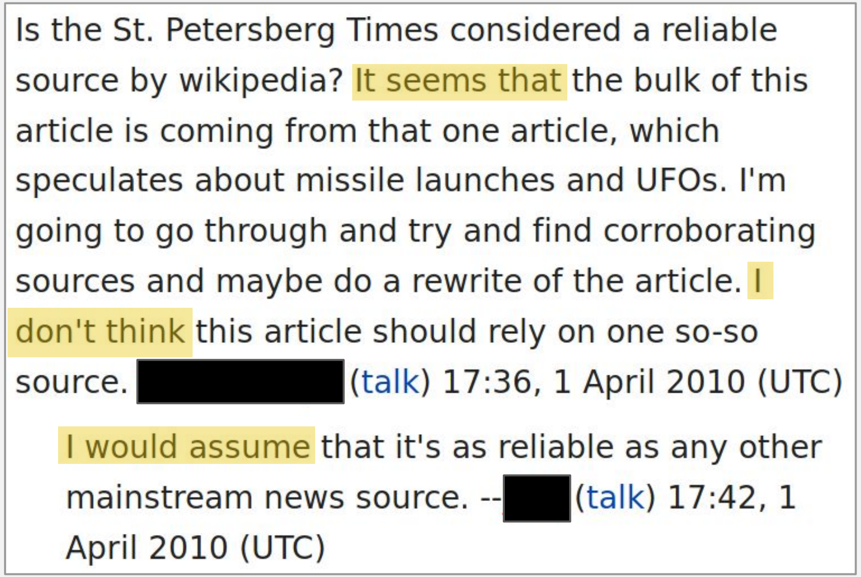
Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Direct questioning

Back to our example...

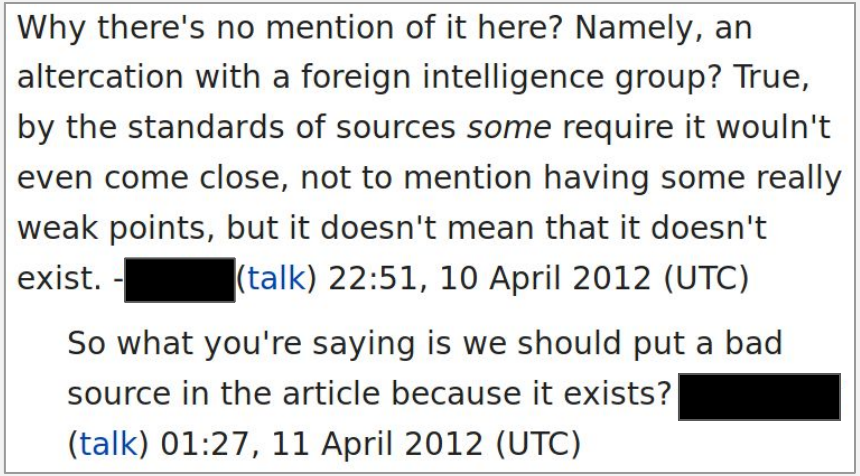
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

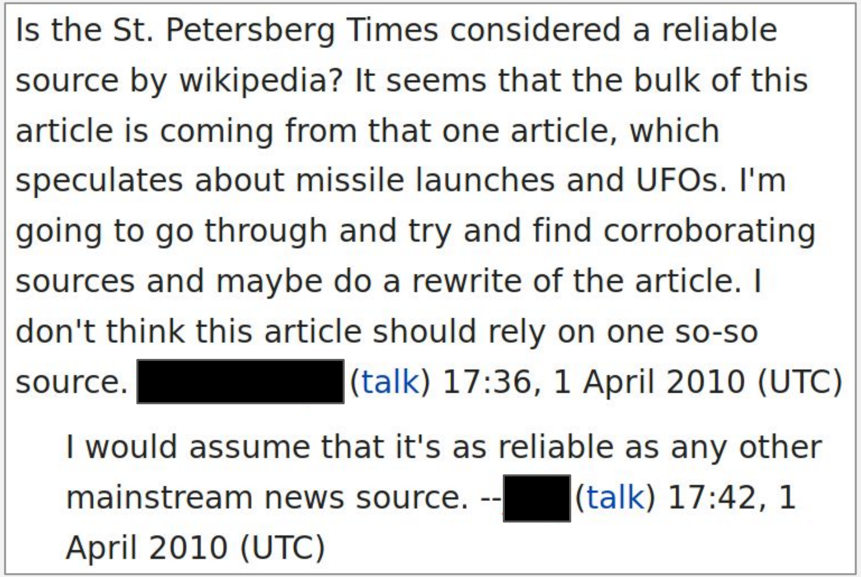
> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

Hedging

Direct questioning

Back to our example...

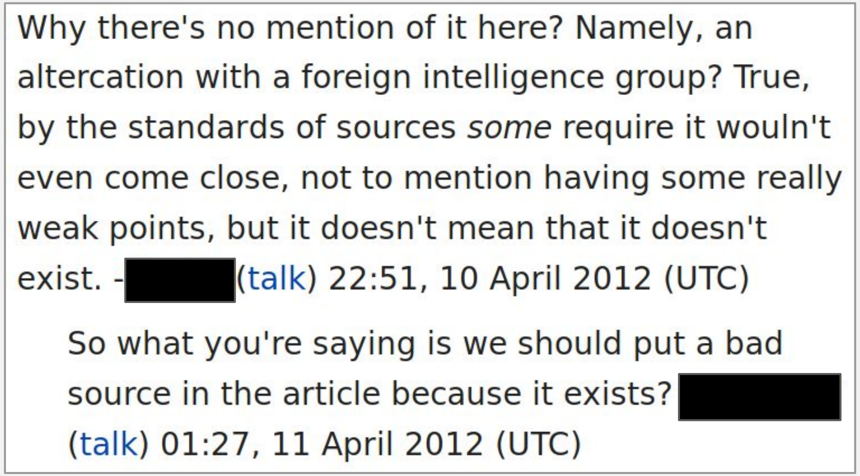
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

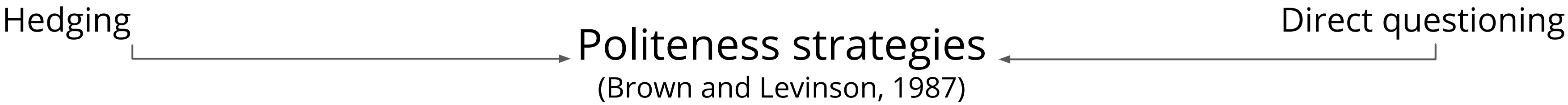
> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)



Hedging

Politeness strategies

(Brown and Levinson, 1987)

Direct questioning

The Role of Politeness

Theory suggests role of politeness in determining conversation trajectory

- Fraser, 1980: Politeness softens the perceived force of a message
- Brown and Levinson, 1987: Politeness acts as a buffer between speakers' conflicting goals
- Goffman, 1955: Politeness is a face-saving tool

But, little empirical investigation so far

Measuring Politeness

How can we detect uses of politeness strategies?

Measuring Politeness

How can we detect uses of politeness strategies?

Danescu-Niculescu-Mizil et al., 2013: pattern match on parsed sentences

- Think regular expressions, but at level of sentence structure

  I [think|feel|believe] that ...

- Try it out: http://politeness.cornell.edu/

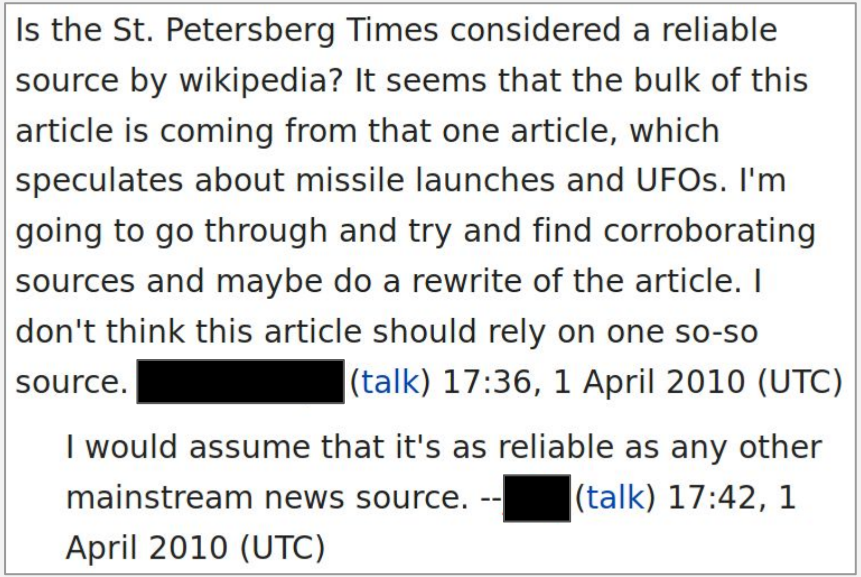
Beyond Politeness: Other Rhetorical Devices

Politeness is a promising feature – but it’s very general

How do we account for domain-specific behavior patterns?

The Example, Once Again

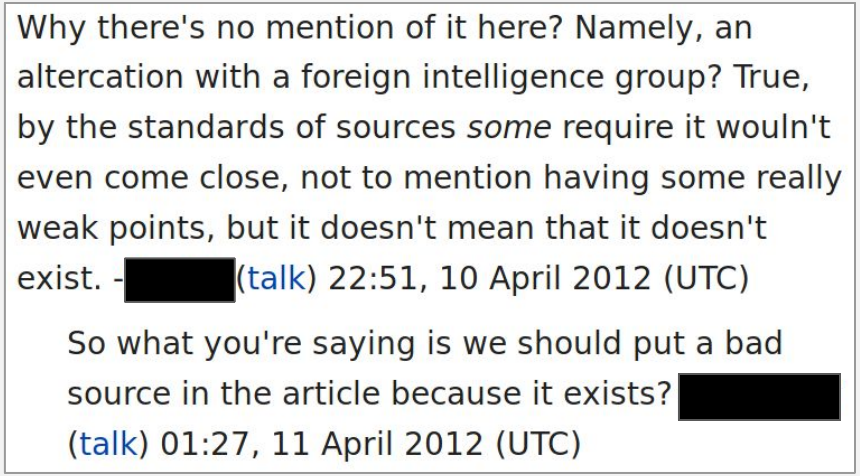
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted](talk) 17:42, 1 April 2010 (UTC)

Conversation B

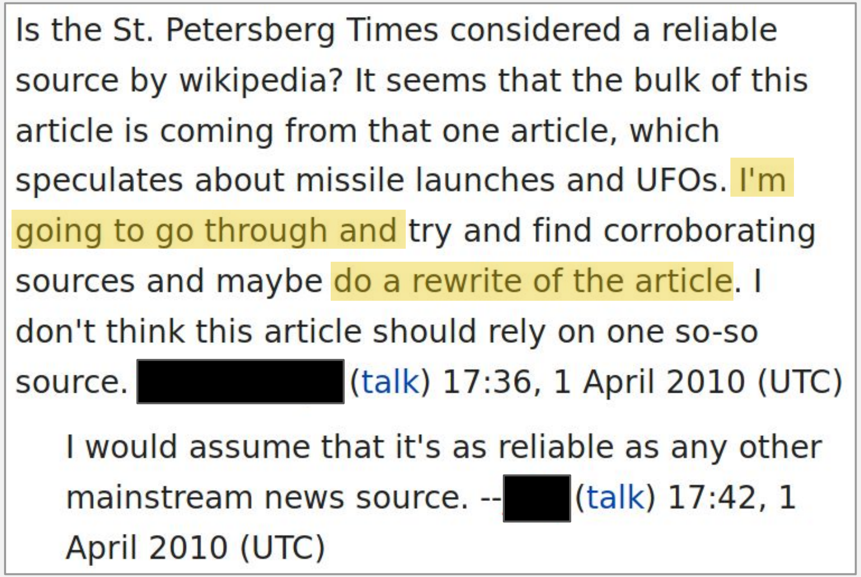


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted](talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

The Example, Once Again

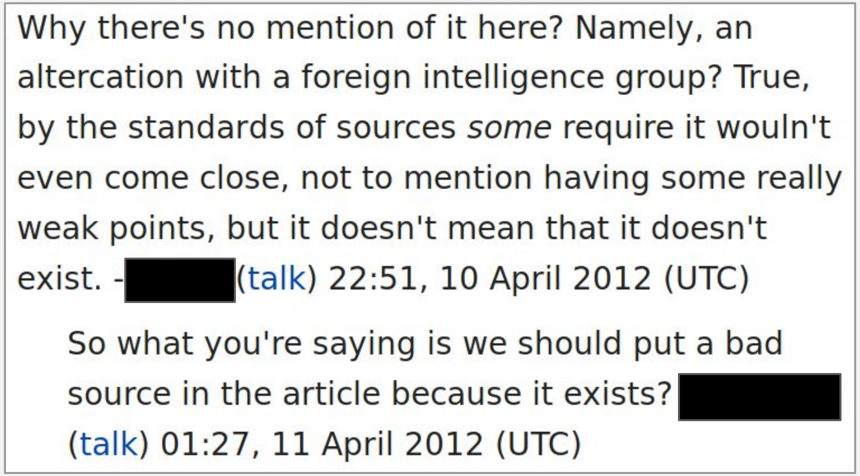
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

> I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B

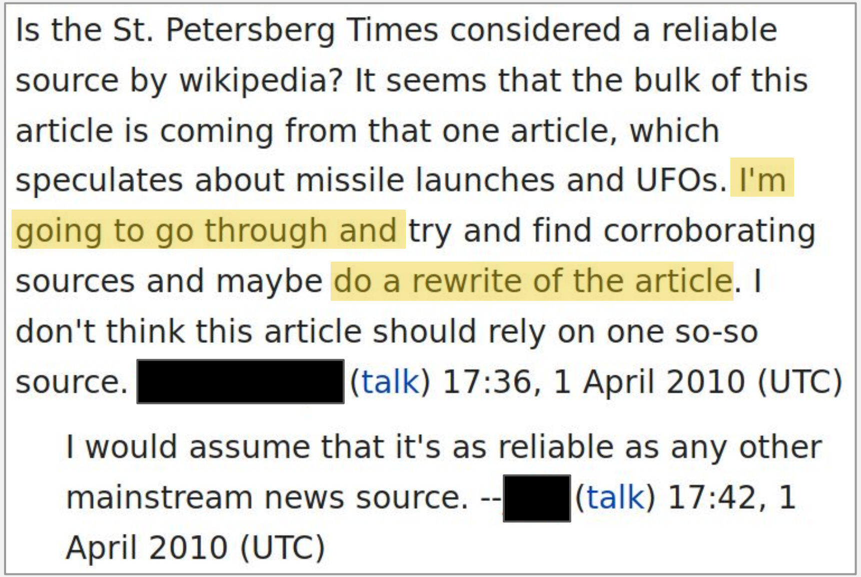


Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

> So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

The Example, Once Again

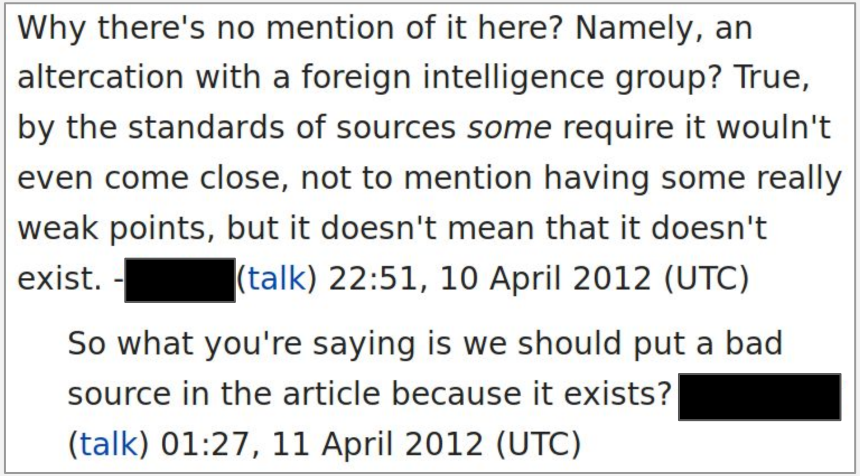
Conversation A



Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source. [redacted] (talk) 17:36, 1 April 2010 (UTC)

I would assume that it's as reliable as any other mainstream news source. --[redacted] (talk) 17:42, 1 April 2010 (UTC)

Conversation B



Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources *some* require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist. -[redacted] (talk) 22:51, 10 April 2012 (UTC)

So what you're saying is we should put a bad source in the article because it exists? [redacted] (talk) 01:27, 11 April 2012 (UTC)

“Plan (to)…”, “like (to)…”, “help…”, etc. - *coordination*

Conversational Prompt Types

A “template” used to initiate conversations

Conversational Prompt Types

A “template” used to initiate conversations

Want to *discover* these automatically - no supervision

Conversational Prompt Types

A “template” used to initiate conversations

Want to *discover* these automatically - no supervision

Solution: extend methodology for finding *question types* (Zhang et al., 2017)

- Original intuition: similar questions trigger similar answers
- Our extension: similar *prompts* trigger similar *replies*

Conversational Prompt Types on Wikipedia

Prompt Type (names manually assigned)	Example
Factual Check	The census **is not talking about** families here.
Moderation	He’s **accused** me **of** being a troll.
Coordination	I **could do with** your **help**.
Casual Remark	**What’s with** this flag image?
Action Statement	The page **was deleted as** self-promotion.
Opinion	**I think** it should be the other way around.

Analysis

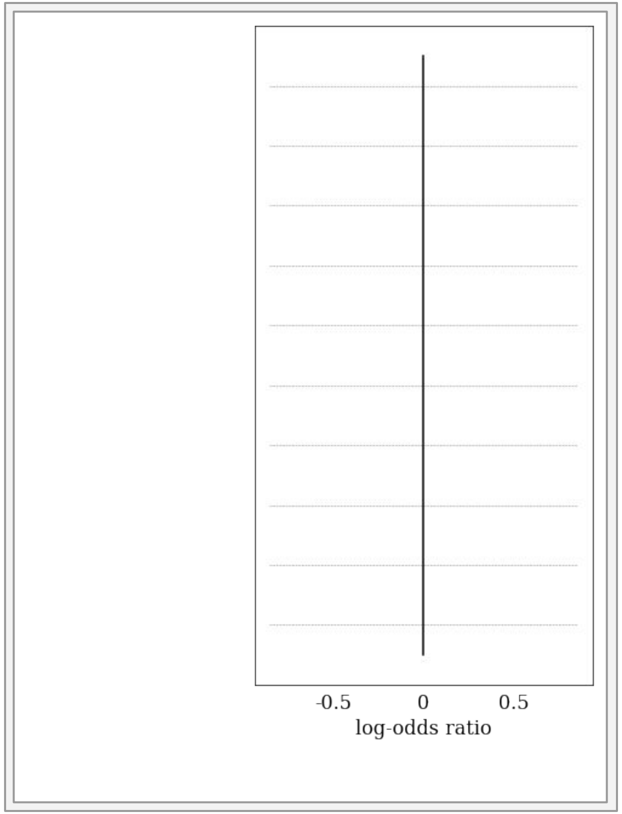
Question of Interest

How well do the prompt types and politeness strategies features actually capture human intuition?

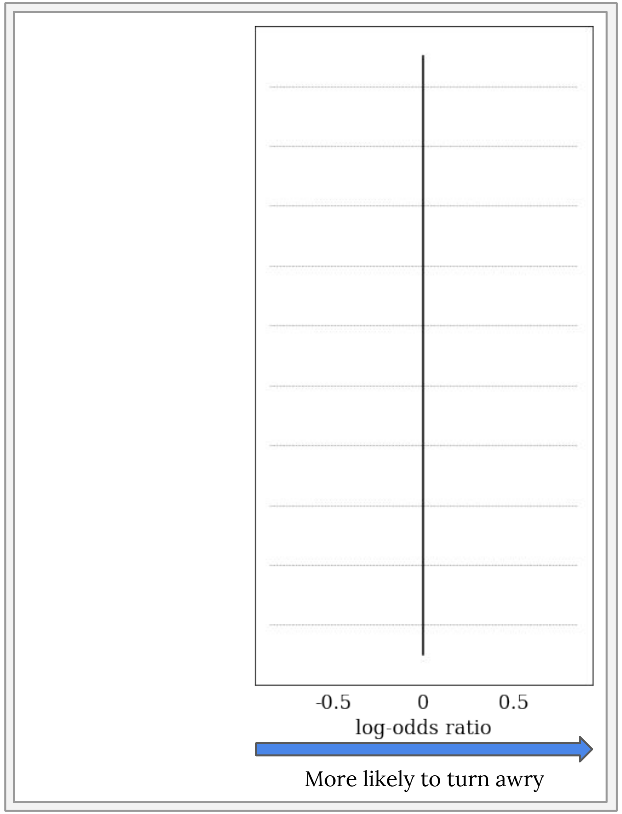
Two ways to answer this question:

1. See if any features are significantly more likely to show up in awry-turning conversations
2. Use the features to create a machine learning classifier that plays the “guessing game” (like the example) and compare to human performance

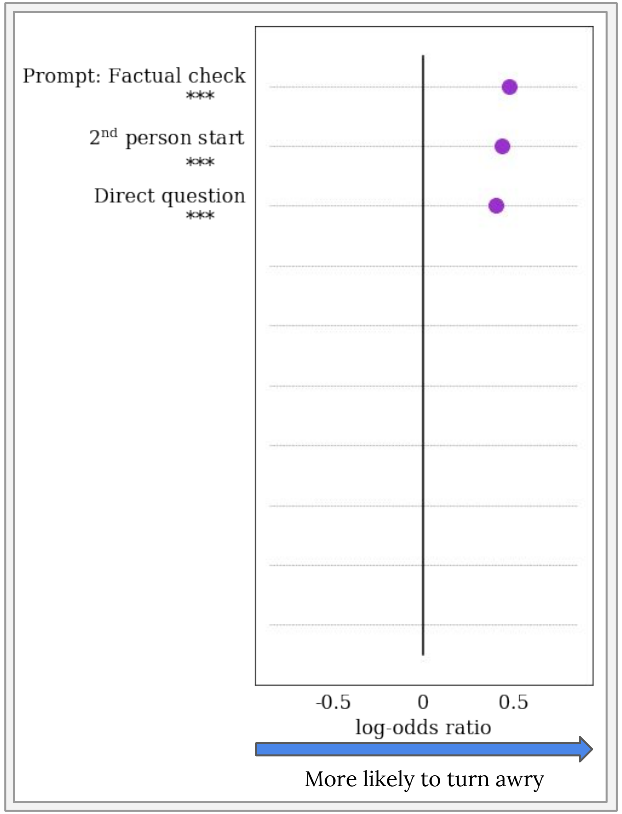
Feature Comparisons (First Comment Only)



Feature Comparisons (First Comment Only)

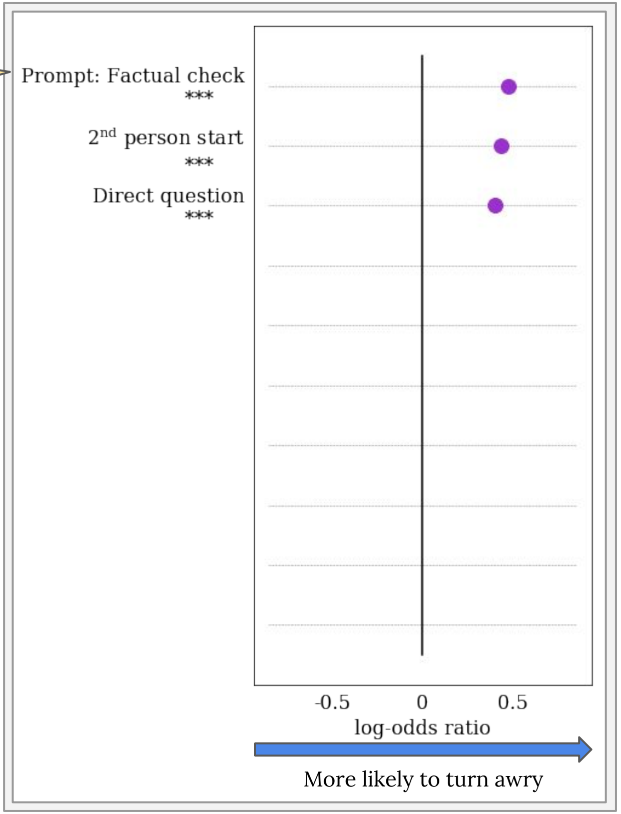


Feature Comparisons (First Comment Only)

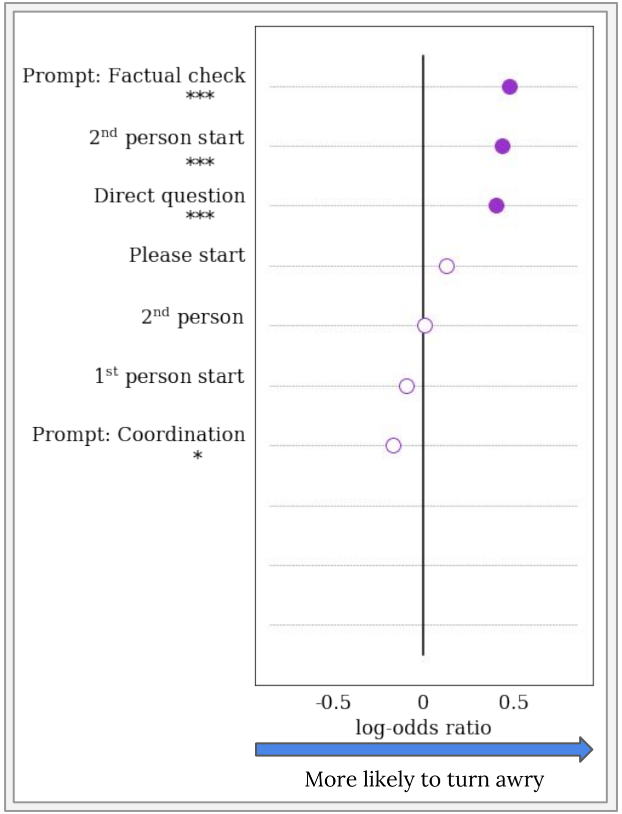


Feature Comparisons (First Comment Only)

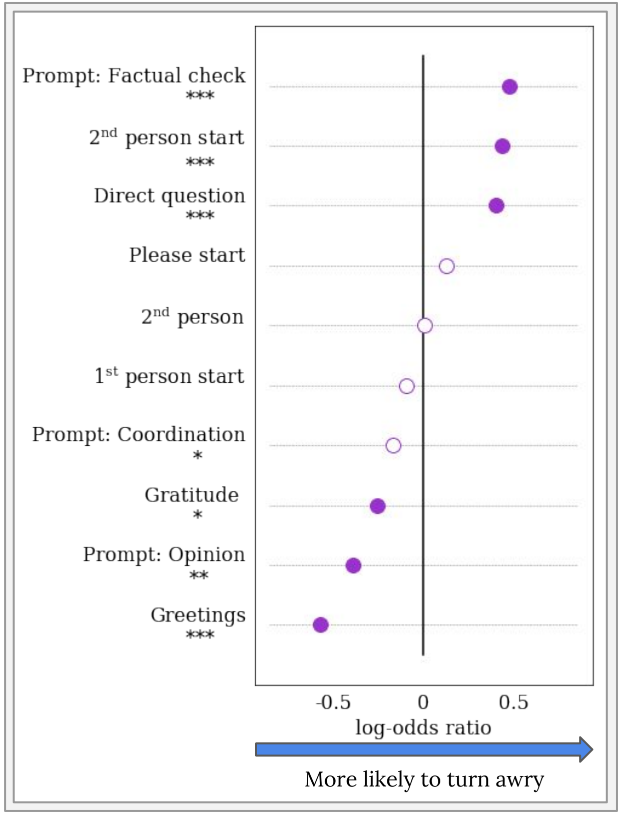
The census **is not talking about** families here.



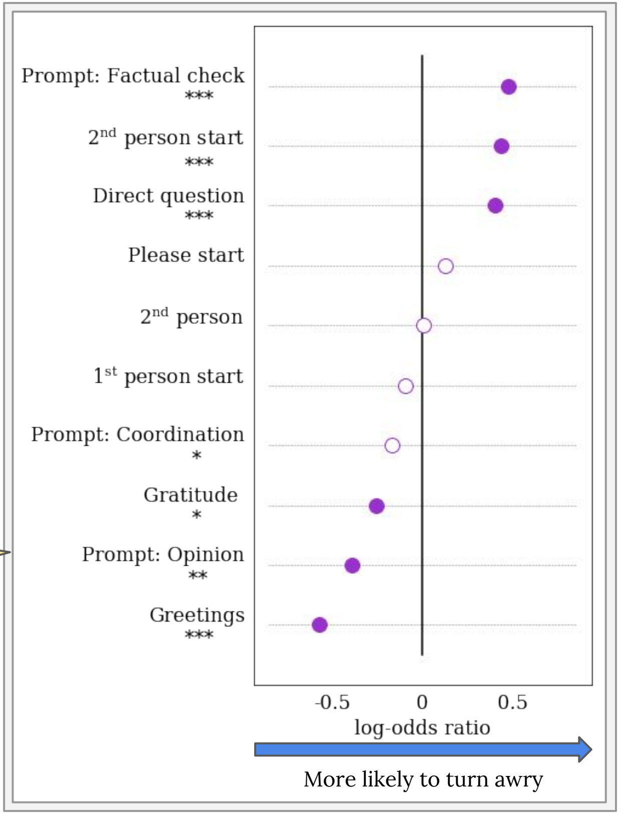
Feature Comparisons (First Comment Only)



Feature Comparisons (First Comment Only)

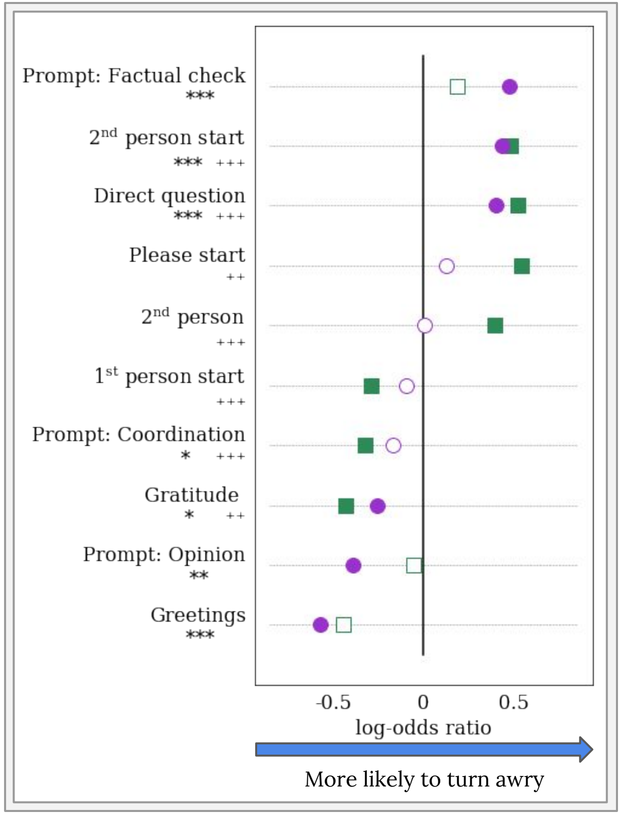


Feature Comparisons (First Comment Only)



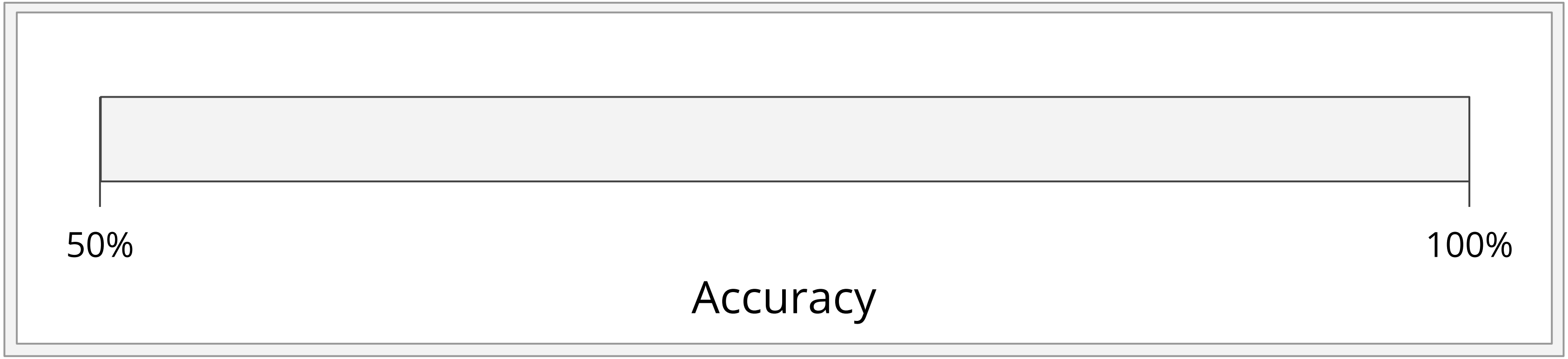
I think it should be the other way around.

Feature Comparisons (First Comment + Reply)

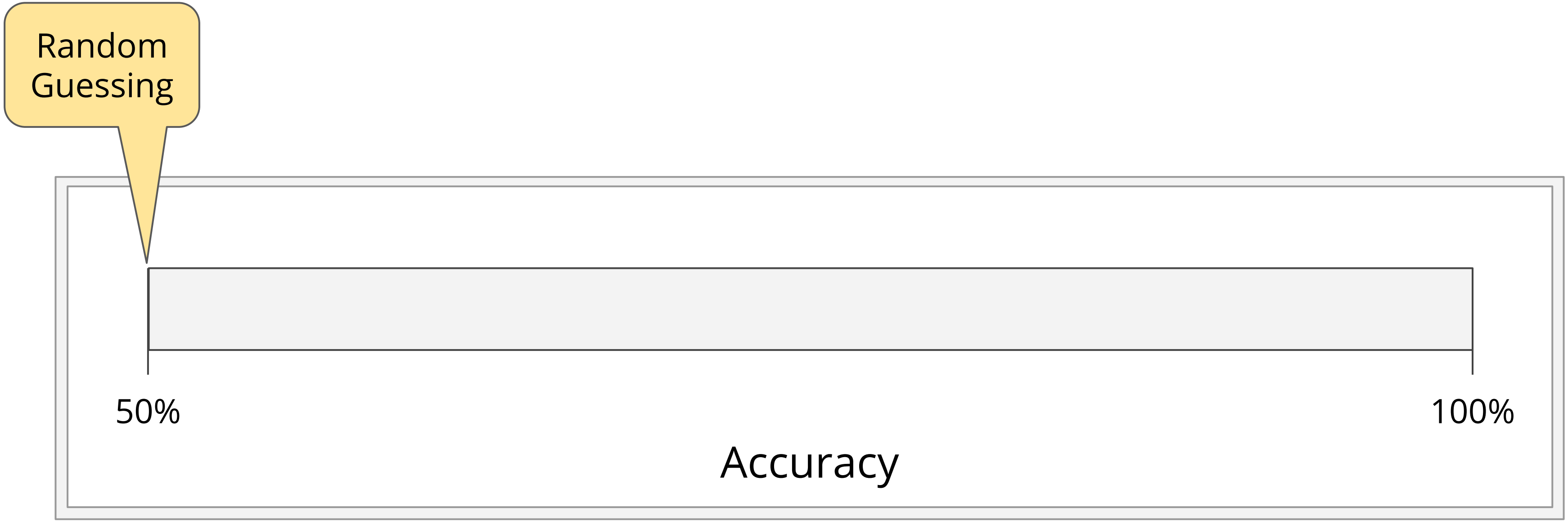


“Guessing Game” Performance

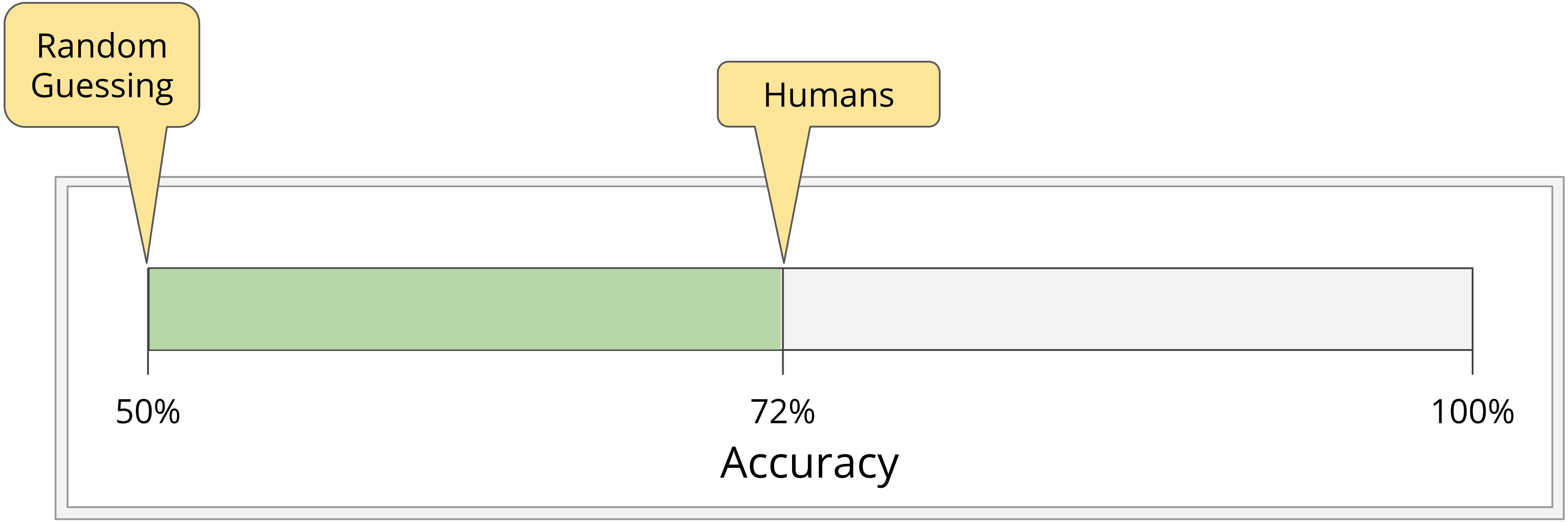
“Guessing Game” Performance



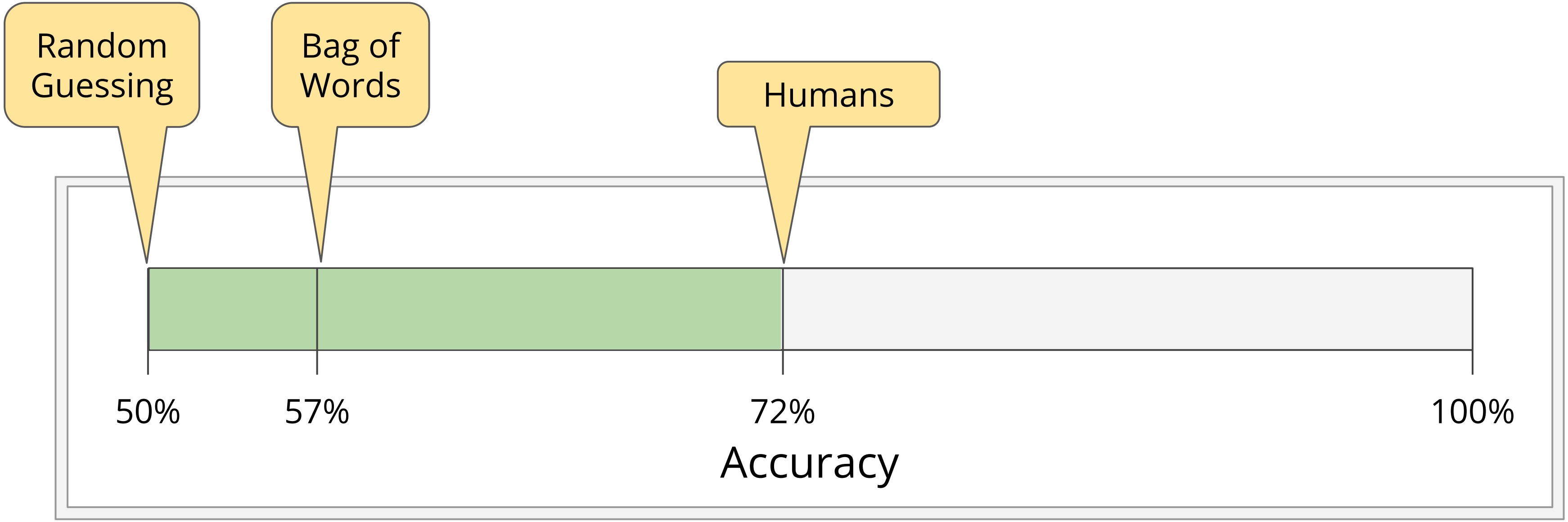
“Guessing Game” Performance



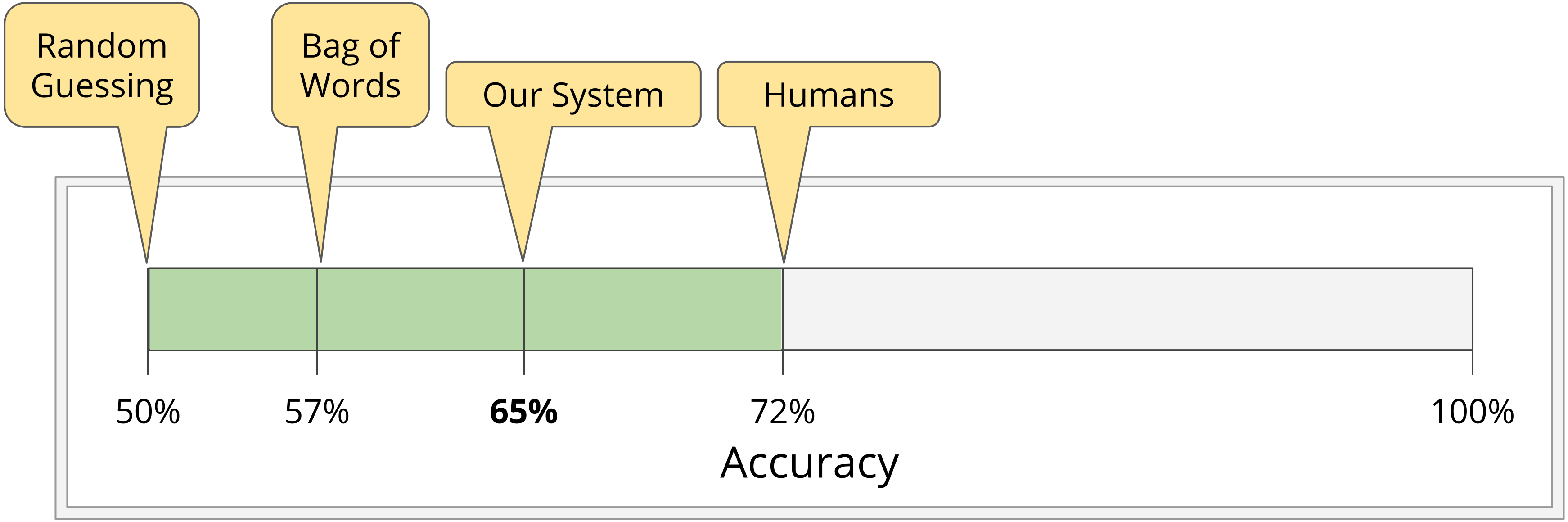
“Guessing Game” Performance



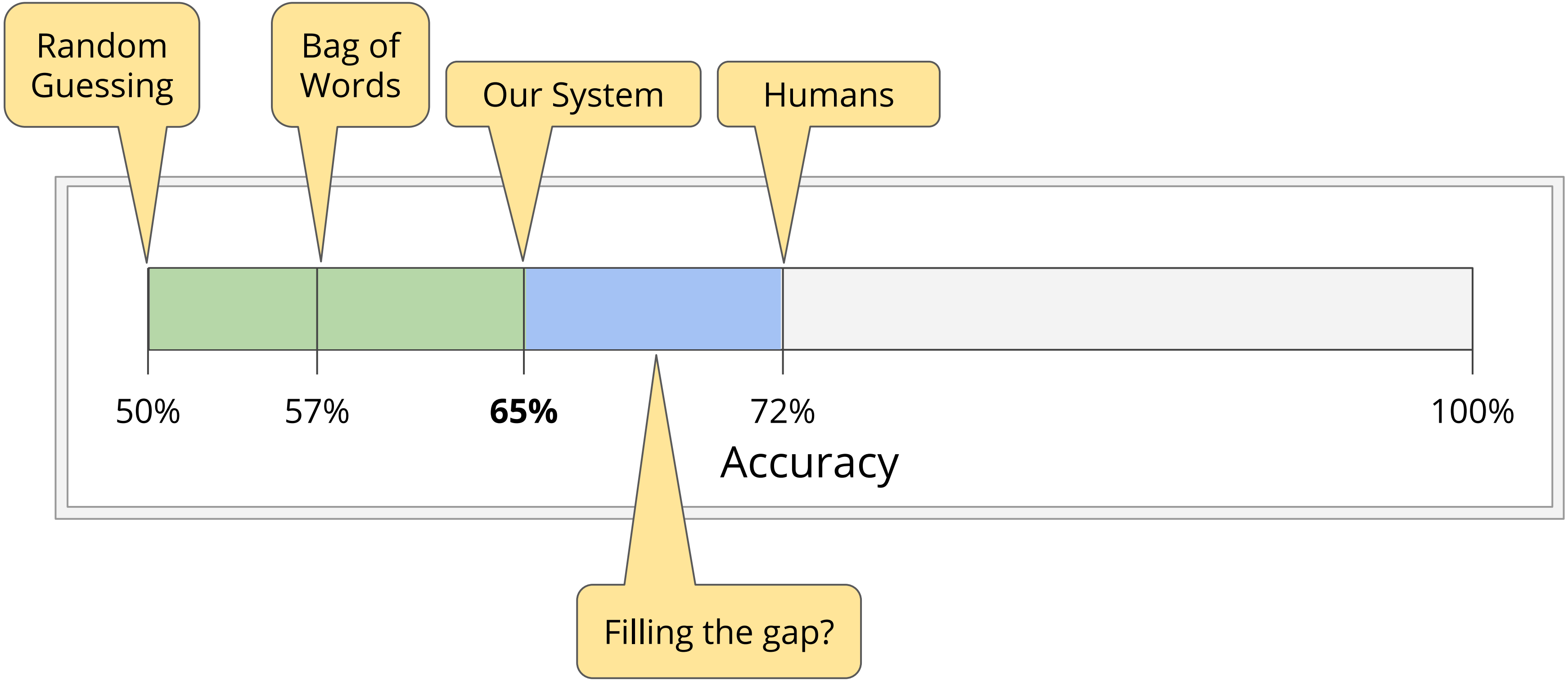
“Guessing Game” Performance



“Guessing Game” Performance



“Guessing Game” Performance



Future Work: Closing the Gap

What parts of human intuition are missing from model? How do we find out?

Idea: examine cases that humans get right, but model gets wrong

- Model correctly guesses 80% of cases humans got right - what about the other 20%?

Future Work: Beyond Conversation Starters

Currently limited to looking only at start of conversation

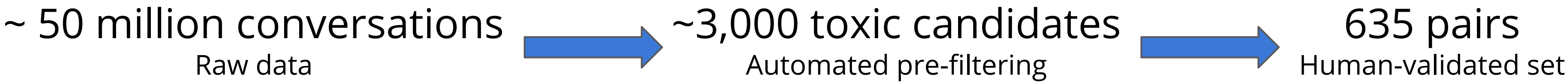
- Ideal model would pick up signal from anywhere in conversation
- Can imagine conversations escalating over time - want to model this

Future Work: Overcoming Biases

What are sources of bias in the current model?

Future Work: Overcoming Biases

What are sources of bias in the current model?



~ 50 million conversations

Raw data

~3,000 toxic candidates

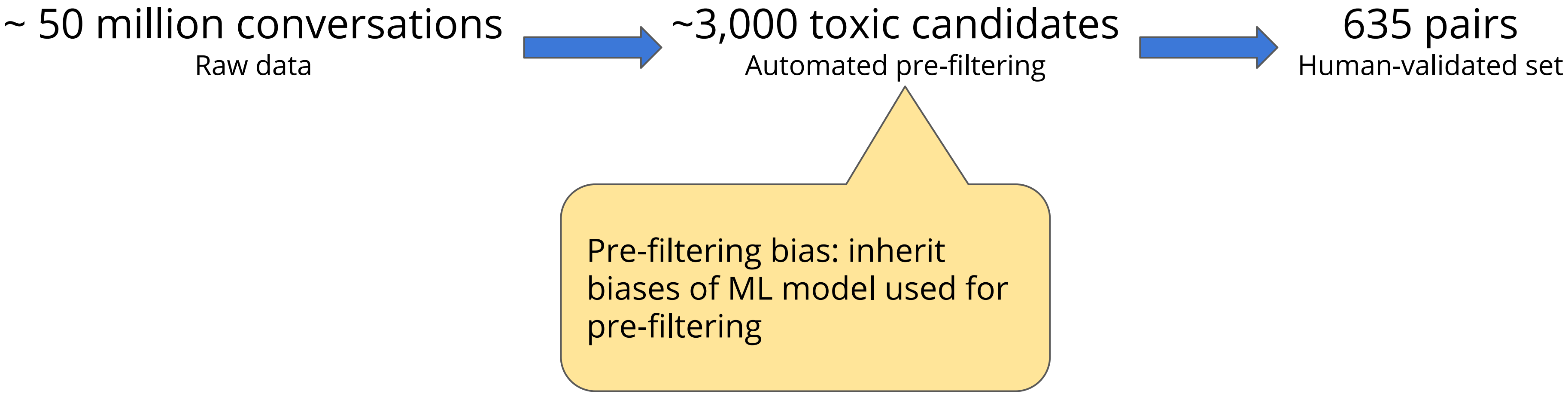
Automated pre-filtering

635 pairs

Human-validated set

Future Work: Overcoming Biases

What are sources of bias in the current model?



~ 50 million conversations

Raw data

~3,000 toxic candidates

Automated pre-filtering

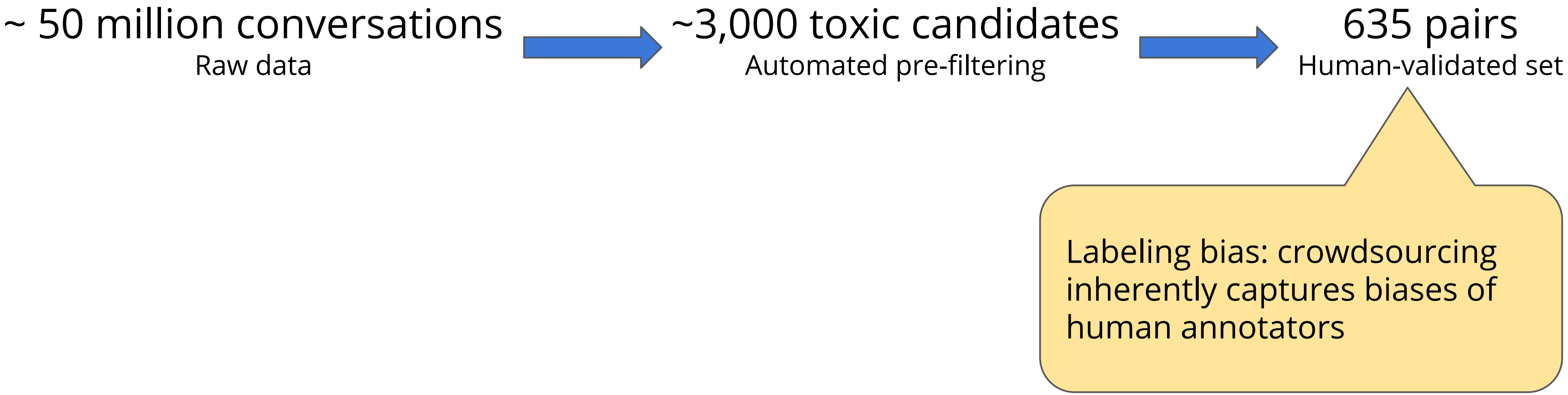
635 pairs

Human-validated set

Pre-filtering bias: inherit biases of ML model used for pre-filtering

Future Work: Overcoming Biases

What are sources of bias in the current model?



~ 50 million conversations
Raw data

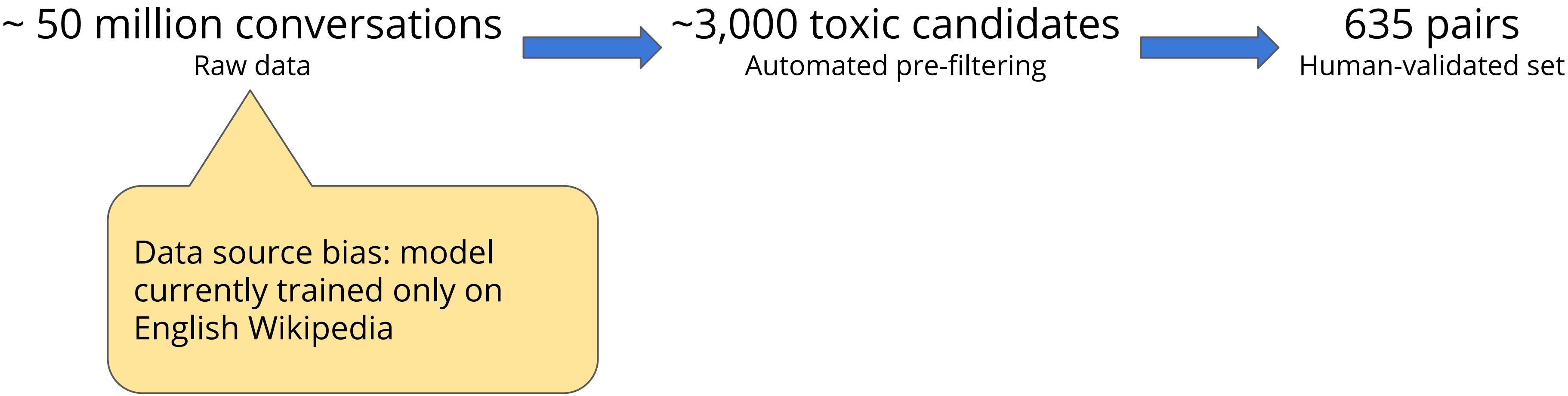
~3,000 toxic candidates
Automated pre-filtering

635 pairs
Human-validated set

Labeling bias: crowdsourcing inherently captures biases of human annotators

Future Work: Overcoming Biases

What are sources of bias in the current model?



~ 50 million conversations
Raw data

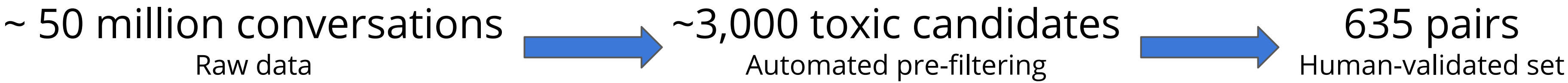
~3,000 toxic candidates
Automated pre-filtering

635 pairs
Human-validated set

Data source bias: model currently trained only on English Wikipedia

Future Work: Overcoming Biases

What are sources of bias in the current model?



~ 50 million conversations
Raw data

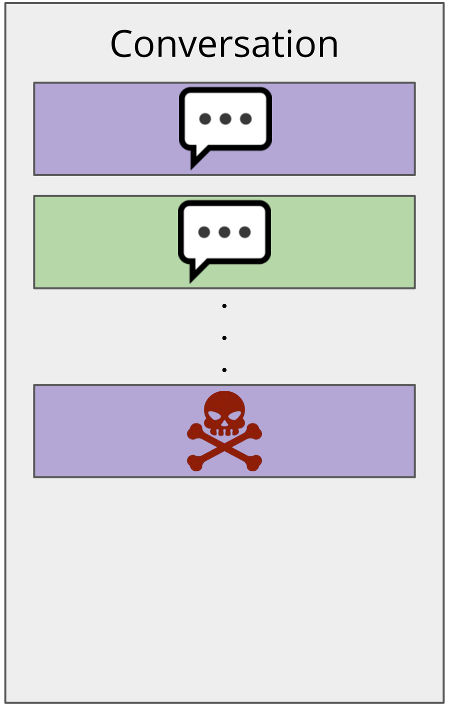
~3,000 toxic candidates
Automated pre-filtering

635 pairs
Human-validated set

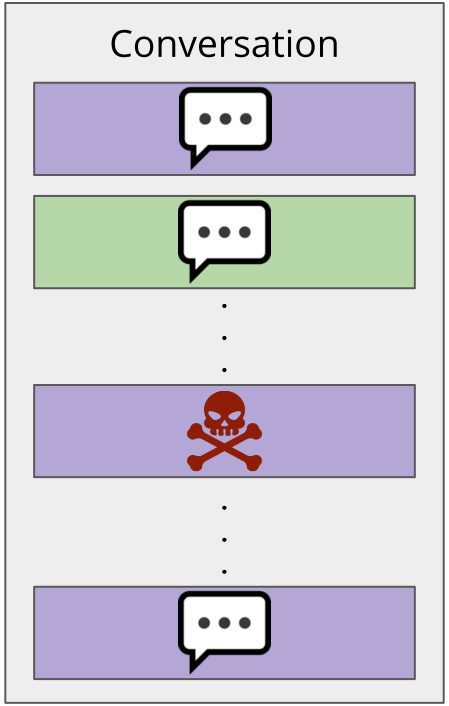
What can we do about it?

- Current direction: explore other ways of pre-filtering and/or labeling

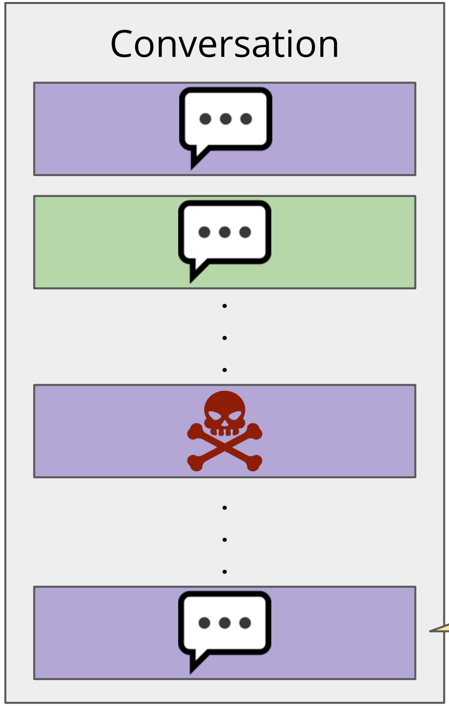
Future Work: Conversation Recovery



Future Work: Conversation Recovery



Future Work: Conversation Recovery



What makes this happen?

Conclusions

Forecasting future attacks in conversations is feasible

Politeness strategies and prompt types capture some human intuition

Experimental verification of politeness theories

Acknowledgements

Everyone who worked on the Wikipedia conversation reconstruction project

The Wikimedia Foundation anti-harassment program

Crowdflower workers who annotated our data

The volunteers who provided annotations for human performance estimate

Questions?

More information at: http://www.cs.cornell.edu/~cristian/Conversations_gone_awry.html
Data and code: http://convokit.infosci.cornell.edu
Online guessing game: http://awry.infosci.cornell.edu/