Truth Goggles: Automatic Incorporation of Context and Primary Source for a Critical Media Experience

by

Daniel Schultz

Bachelor of Science in Information Systems, Carnegie Mellon University (2009)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author ______________________________
Program in Media Arts and Sciences
May 18, 2012

Certified by ______________________________
Henry Holtzman
Research Scientist
Media Lab
Thesis Supervisor

Accepted by ______________________________
Mitchel Resnick
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Truth Goggles: Automatic Incorporation of Context and Primary Source for a Critical Media Experience

by

Daniel Schultz

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 18, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Falsehoods come in many forms. Politicians and advertisers make false claims, newsrooms and bloggers make mistakes, and the ease of publication and sharing makes it easy for even the most incredible theories to quickly spread. The risk of interpreting fiction as truth is significant and it is only increasing. How can technology protect consumers from falling into the traps set by dubious information?

This thesis is an exploration of the challenges and possibilities surrounding the implementation of interfaces that mitigate the effects of misinformation. It contains an analysis of the pitfalls, considerations, and opportunities in the space of digital credibility. Those lessons are then applied to Truth Goggles, a technology prototype that attempts to trigger critical thinking in digital media consumers and codify the process of fair analysis.

Thesis Supervisor: Henry Holtzman
Title: Research Scientist, Media Lab

Truth Goggles: Automatic Incorporation of Context and Primary Source for a Critical Media Experience

by

Daniel Schultz

The following people served as readers for this thesis:

Thesis Reader____

Henry Holtzman
Research Scientist
Media Lab

Thesis Reader____

Catherine Havasi
Research Scientist
Media Lab

Thesis Reader____

Ethan Zuckerman
Principal Research Scientist
Center for Civic Media

0.1 Acknowledgements

This thesis is dedicated to the memory of my grandmother, Barbara Burwell, who passed away when it was just getting started. She taught me how to have an open mind and embrace the best in everyone. I can only hope I'll be as good a teacher as she was.

Second billing goes to my most noble and respectable benefactors: The Knight Foundation. Without you I would probably be inadvertently working to promote the forces of evil.

Thanks to my parents for giving me a personality and every opportunity in the world. Also my little brother is pretty cool too.

Special thanks go to my advisor Henry Holtzman and the world famous Information Ecology group, who have patiently helped me learn, troll, and discover outlets to my passions despite my obvious shortcomings.

Thank you also to my wonderful readers whose advice and guidance saved this thesis from completely collapsing.

Infinite <3's go to my beautiful wife, Lyla, who by now has one of the most well trained bullshit detectors in the universe. I'm pretty sure she loves me anyway.

No thanks go out to awesome, a terrible mailing list that I can only hope I have stolen more collective time from than it has stolen from me.

Finally, of course, I have to thank the Internet. Without you this entire thesis would be completely irrelevant.

P.S. sorry for the spam

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "One of the most salient features of our culture is that there is so much bullshit. Everyone knows this. Each of us contributes his fair share. But we tend to take the situation for granted." - Harry G. Frankfurt, On Bullshit

In 1938 the Columbia Broadcasting System aired Orson Welles's radio adaptation of War of the Worlds, a story of a Martian attack on Earth. The program was misinterpreted as an actual report of an alien invasion by thousands of listeners who tuned in midway through. The next morning, when it became clear that Earth was not being overrun, the public felt duped; their ability to discern fact from fiction had failed them. This infamous event has since been used to inspire much deeper conversations about the power of mass media, the power of trust, the ethical responsibilities of content creators, and the importance of media literacy.

Technologies that make it easier to spread and create messages often necessarily make it easier to receive them as well. Those messages will sometimes contain misinformation, and communication technologies do not need to protect media consumers from that misinformation in order to function. The logical conclusion is that media consumers face increased exposure to misinformation without increased protection, a trend often cited by champions of legacy media as a critique of modern methods of information distribution.

This thesis is an exploration of the challenges and possibilities surrounding the implementation of credibility layers—interfaces that mitigate the effects of misinformation. I begin by discussing the nature of misinformation, describing several existing efforts to combat it, and explaining the practical and psychological constraints surrounding the act of correction. I continue with a deep dive into the decisions I made when creating and testing a new credibility layer called Truth Goggles.

Truth Goggles consists of a database and API for fact checked claims, a set of processes to identify instances of those claims, and a browser-based interface guides users through the process of critical consumption.

1.1 Motivation

1.1.1 Critical Ability

Falsehoods come in many forms. Politicians and advertisers make false claims, newsrooms and bloggers make mistakes, and the ease of publication and sharing makes it easy for even the most incredible theories to quickly spread. The risk of interpreting fiction as truth is significant, and even professional reporting institutions report and create misinformation on a regular basis despite a structure and business model built around credibility [66].

When Osama Bin Laden was assassinated, a very convincing image of a bloodied Bin Laden head was picked up and shared by professional and social channels alike. The photo was altered via Photoshop, as many denizens of the internet could tell from having seen quite a few manipulated—or "shopped"—images in their time [31], but after the image was revealed as a fake it continued to be presented across news sources and social networks as legitimate. Such occurrences are all too common in the modern mainstream press, where even professionals have been caught doctoring their own work in the name of creating a more dramatic story and at the expense of accuracy [35].

Content consumers have a line of defense against misinformation: critical ability, a concept introduced by Hadley Cantril in his writing about the panic surrounding Welles's 1938 War of the Worlds broadcast [17]. It refers to a persons natural instinct and capacity to critically inspect information before incorporating it into their world view. For instance, in the context of the CBS broadcast, a person with higher levels of critical ability would have been more likely to question the alarming information they heard on the radio and understand that it was a fake before responding with panic.

Cantril could not have been thinking about the Internet when he codified critical ability—during his time radio audiences were "the most modern type of social group" [17]—but today this psychological trait is arguably even more important. We spend more time consuming information today [12], our information channels are more diverse, and across both traditional and new media formats there seem to be fewer consequences to misinformation.

To put it simply we have more chances of being exposed to false information.

Technology can help consumers protect themselves from misinformation by helping them hone their critical abilities. Evidence about a situation or claim is usually just a few minutes of research away, and we can create credibility layers that leverage this information to connect dots, identify dubious claims, and remind people to think twice before accepting new information into their world views. For situations where the information is not available or is not clearly associated, credibility layers can leverage human networks to help uncover or associate new knowledge [10]. Between data mining, traditional journalism, and crowd sourcing, it is entirely possible to construct systems of credibility that encourage vigilance when it is needed most.

1.1.2 Common Respect for Information

If skepticism of information sources can trigger critical ability then we should be witnessing a critical golden age. In 2010 there were more people with negative opinions of professional media than there had been for over two decades [18]. Two-thirds of Americans believed that professional news outlets were "often inaccurate," over three-quarters believed that the content was biased, and four-fifths believed that journalists were being influenced by powerful people and organizations [18].

Unfortunately these statistics do not reflect the healthy skepticism that consumers should have of all media. Consumers have negative opinions towards the broad concept of "media" while maintaining higher levels of respect for their personal sources. 62% of respondents from the same survey believed that the news sources they personally use the most "generally get the facts straight." Only 25% felt the same way about news organizations in general [18].

This type of skepticism is polarizing —it represents a social landscape that reeks of selection bias (choosing to view information that aligns with your world view) and disconfirmation bias (arguing more strongly against information that does not align with your world view). Media fragmentation has allowed voting citizens with divergent points of view to consume very different pieces of information in isolation from one another [45]. Audiences cannot even have a productive discussion afterwards because they are also less likely to respect the other person's sources [69]. It is not easy to reach a common ground with someone who does not trust your information and has never been exposed to your perspective.

One hope for Truth Goggles is that it will be able to provide some common ground without forcing a change in consumption patterns. It can be applied to the Boston Globe just as easily as it can be applied to CNN, MSNBC, The New York Times, or Fox News and every user is exposed to the same pool of information. As the set of fact-checking services expands, users will also be exposed to a wider variety of perspectives. These people are then able to exhibit a more balanced form of skepticism by placing their own sources under productive scrutiny and subsequently become more informed together.

1.1.3 The Fact Checker from Nowhere

In January 2012 Arthur Brisbane of the New York Times posed a simple question that began a firestorm of opinions: "Should professional journalists call out lies?" [13] [14]. To an individual unfamiliar with the ethics and practices of professional journalists the question might appear quite silly. Isn't it a journalist's job to tell the truth, only the truth, and nothing but the truth? In one sense, of course! But then again, of course not!

There are many different ways to report a lie. It would be an obvious mistake to report that something happened when it didn't, or cite a statistic that isn't accurate. But what happens when the report is simply that someone made a statement which also happened to be inaccurate? If the journalist omitted that statement it would be censorship. If they call out the statement as a lie that would imply judgement, which implies perspective, which implies opinion, which implies potential bias.

In the past this quandary was handled by reporters who would seek out third parties to make counterpoints and provide alternative information, which could then be reported alongside the original statement to help edge the consumer towards a more refined perspective. Today, however, the pace of the news is far too fast and the resources of newsrooms are far too small for this to be a scalable practice any longer; and so Brisbane's question makes a bit more sense.

When blatant mistruth can be uttered and spread in a matter of minutes is this view of journalism still what democratic societies need? Modern fact checking, the type of journalism that is willing to reach a verdict, provides different ways of thinking about what it means to report. While traditional journalism left it to the reader to reach an informed conclusion, fact checkers attempt to take some of that onus off of the informationally overloaded individual at the expense of neutrality.

We need guidance, but we also need reports and raw information to reach our own conclusions. With current consumption interfaces these are conflicting requirements, but technology makes it possible to have it both ways. Credibility layers and tools like Truth Goggles enable new experiences, new ways of thinking about journalism, and new standards for information presentation. We, as readers, can have traditional reports written by traditional reporters while also getting guidance through credibility layers to protect us from being tricked by spin, framing, and misinformation.

1.2 Paper Organization

This thesis describes the plans, designs, and efforts that went into the creation of Truth Goggles. I am hopeful that the concepts discussed can help inspire and guide future developers to bring additional credibility to the Internet. The document is broken into an introduction, five sections, and an appendix.

- *Context* is an exploration of the spaces of fact checking, consumer biases, and existing credibility interfaces.
- *Design* explains how the lessons learned from past efforts have fed into the design of Truth Goggles. It also introduces the use cases and guidelines taken into consideration.
- *Implementation* provides a detailed description of the resulting system, an explanation of key decisions, and an overview of the available functionality.
- *Evaluation* documents the user tests that were performed using a version of Truth Goggles, explores the resulting data, and synthesizes the feedback provided.
- *Conclusion* describes lessons learned, next steps, and high hopes for the future of journalism.

1.3 Contribution

This document describes the APIs, interfaces, and goals behind Truth Goggles. It contains an analysis of the pitfalls, considerations, and opportunities surrounding digital credibility. I contribute a better understanding of the potential effectiveness of consumer-facing credibility layers, an open source architecture for credibility layers, and a proposed standard for a credibility API that future fact-based interfaces can harness.

By implementing Truth Goggles I also offer an interface and first attempt at a new system for media consumers who want to be able to trust their sources.

Chapter 2

Context

In order to defend against misinformation one needs to understand how consumers reach conclusions, how mistakes are manifested and corrected in the media, and what best practices already exists. There is significant work, both completed and ongoing, that attempts to leverage technology in the name of critical thinking. Journalists have also begun to understand what it means to respond to lies and mistakes through new processes such as fact checking and transparency. This section explores current truth-seeking efforts, describes what we know about the biases and effects that can hinder critical ability, and explains a few ways that consumers handle unhappiness with media messages.

2.1 Mistakes and Lies

Where does misinformation come from? A media message can be packed with mistakes long before it is even shown to a consumer. This section describes some of the ways in which misinformation can be created and perpetuated. It is by no means an exhaustive list, but provides some key examples of the ways in which a message can be spun, either intentionally or accidentally, at the expense of credibility.

2.1.1 Simple Mistakes

In 2011 the United States military, under the leadership of Barack Obama, carried out a successful mission to assassinate Osama Bin Laden. When the news was announced that Osama had been killed the Fox News Network made the mistake of running the very unfortunate headline: "Obama Bin Laden is dead" [11] [1]. This is an example of the most trivial class of media error: the simple mistake. Indeed some of the most common errors involve swapped names, incorrect titles, typos, and erroneous numbers [66].

2.1.2 Intentional Manipulations (Author)

Consumers can be misled through the triggering of biases, unfair framing, ideological cues, and loaded language. The difference between two takeaways might be as simple as a modified headline. Publishers with agendas are well aware of the ways that people process information and have the ability to take advantage of every trick in the book.

2.1.3 Intentional Manipulations (Newsmaker)

There are also situations where newsmakers have an incentive to knowingly stretch the truth. Politicians benefit from making promises even if they know they will not be able to

[1] Just to emphasize how easy it is to make this type of mistake, readers of this thesis were unable to notice the error without re-reading the headline multiple times. Hint: "Obama" vs "Osama"

follow through, newsrooms benefit by making stories more dramatic, and advertisers benefit by making their products look better than the reality. This type of misinformation is less likely to receive a correction because it is generated and spread intentionally. These make up the majority of the statements that Truth Goggles is intended to identify.

2.1.4 Poor Sources

Of course it is also possible for content authors to make serious mistakes without an intent to mislead. A journalist might get his or her information from an unreliable source or forget to double check a key aspect of a story. This mistake is easy to make today given the number of crowd driven sources. For example, evolving information repositories like Wikipedia make it possible to completely invent truth: a false claim on Wikipedia can be picked up and transmitted through credible sources in a way that circles back to give undeserved credibility to the original claim [46].

2.2 Fact Checking

Today there are many organizations, journalists, and collectives explicitly dedicated to promoting credibility. The landscape contains participants of all kinds: information sources and information distributors, partisans and non-partisans, professionals and crowds, traditional and experimental operations. This section describes some of the organizations and projects that have paved the way for Truth Goggles.

2.2.1 Fact Checkers Today

Newspapers and magazines started to hire staffers explicitly responsible for fact checking as early as the 1920s [66], but the concept of a "fact check" as a regular, audience-facing feature or newspaper column didn't start until 1985, when Ronald Reagan tried to convince America that trees generated more pollution than cars [21]. From that point forward several media outlets began to seek out bogus claims and careers began to form around correcting political misinformation. Below is an overview of some of the more prominent fact checking operations today.

PolitiFact Politifact is "a project of the Tampa Bay Times and its partners" [4] and is built around professional reporters and news organizations. Politifact tracks claims made by politicians, pundits, and participants in congressional hearings [3]. They are powered by a network of journalists, so it is no surprise that their processes reflect the practices of traditional journalism. When confronted with a potentially dubious claim they do research, cite sources, and attempt to piece together an analysis. That analysis is eventually made public and published alongside their final verdict. The verdicts are placed on a proprietary scale called the "Truth-o-Meter" which can hold the values of True, Mostly True, Half True, Mostly False, False, and the ever degrading "Pants on Fire".

PolitiFact's data set is particularly accessible for toolmakers due to the fact that they have a developer API. The API is closed, but anyone with permission can use it to interface directly with their corpus of fact checked claims. This also means that they maintain their content in a fairly structured format. Fact checked claims are called "ruling statements" which, for the purposes of a tool like Truth Goggles, represent the phrases that should be flagged in a credibility layer.

In addition to generating fact checks, PolitiFact also tracks changes in stance ("flip flops") and monitors the status of political promises, monitoring whether they have been kept and broken over time. All of this information is accessible over their API and the web site, although Truth Goggles only uses their fact check information. Politifact scrutinizes conservatives and liberals alike, and has no stated partisan agenda.

FactCheck.org FactCheck.org is a project of the Annenberg Public Policy Center of the University of Pennsylvania [33]. Their mission is very similar to PolitiFact's, although they take a slightly different approach. FactCheck.org produces detailed analysis surrounding broader issues, discussions, and events as opposed to focusing on specific statements that embody a particular dispute. They analyze their topics in a way that combines "the best practices of both journalism and scholarship" [34], and the result is a full article that explores the issue at hand including an academic-style bibliography.

Surely by design, FactCheck does not make it easy for a reader to reach a conclusion at a glance in the same way that PolitiFact does. There is no equivalent to the Truth-o-Meter on FactCheck.org. They also do not provide an API and do not structure their information in a way that makes it easy to identify the specific pieces of content being fact checked. For these reasons Truth Goggles does not currently use FactCheck as a data source despite the quality of their work.

Washington Post's Fact Checker The Washington Post Fact Checker is a political column and blog based at the Washington Post. It focuses on claims surrounding "issues of great importance, be they national, international or local" [37]. As this statement suggests, entries are written about particular issues with an explanation of the context, an analysis of the fact, and ultimately a rating of truth. Truth values are assigned on a "four pinocchio" spectrum, with zero pinocchios representing a very true statement and four pinocchios representing a very false statement.

Like FactCheck.org, the Washington Post presents content in a blog format and without a semantically accessible structure. It is difficult to automatically parse out the specific statements being fact checked because the entries tend to be about wider issues. As a result, Truth Goggles does not currently incorporate this corpus.

Snopes Snopes is a family run website that began in 1995 and has evolved into a well known digital resource [43]. The authors focus their attention on folkore, urban myth, and viral emails with categories that range from "Autos" to "Weddings." They also have sections that are dedicated to the analysis of political myths. The cases they explore yield a verdict ("true," "false," "multiple truth values," "undetermined," or "unclassifiable veracity") along with an article to explain the verdict.

Snopes does not provide an API and is fairly protective of their content. They do, however, organize their analyses in terms of the specific claims and instances of potential misinformation and reach a categorized verdict, making their corpus a potentially great fit for Truth Goggles.

Additional Sources There are several other fact checking operations beyond the sites, blogs, and organizations described above, and each puts forth an alternate take on the concept of debunking. One particularly intriguing example is Truth Squad, which offered a crowd sourced approach to fact checking. Truth Squad was hosted by a now defunct organization called News Trust [23]. The service yielded articles with a well organized format and structure quite similar to the one pioneered by PolitiFact, but the process behind that output relied on users to contribute perspectives and arguments, and thought leaders to synthesize that input into a verdict and article.

Another interesting set of sources are those which are admittedly partisan. This includes organizations like Media Matters [24], Newsbusters [48], and Fairness and Accuracy in

Reporting [60]. These sources tend to produce content that has been explicitly designed to speak to a specific type of audience and further a partisan agenda. As a result their output regularly contains charged language and focuses on topics that fit into the organization's broader agenda to promote a particular political attitude.

2.2.2 Credibility Layers

Truth Goggles is not the first credibility layer to exist for the Internet and it is not the only one being actively developed today. This section provides an overview of several past and present efforts to help readers understand the truth behind a statement, email, or online article.

Dispute Finder Dispute finder is a javascript-driven interface developed at Intel Labs and published in 2009. The tool was created to help "users know when information they read on the web is disputed" [22]. Their script modifies web pages to insert a highlight over any piece of content that had been identified as "disputed." Users are then able to click on disputed phrases to view a list of external content that supports the phrase as well as a second list of external content that opposes it.

Dispute Finder is now a retired project, but when it was active it resulted in a short paper describing the system and the process of building associations between snippets and claims [22]. It relied on its users to help identify claims, associate evidence to those claims, and organize that evidence. It incorporated algorithms to help those crowds find potential evidence, but for the most part was reliant on humans. The system was able to exist despite this shortcoming because certain classes of user claimed to "already spend a lot of time performing similar tasks." More specifically, "activists" and idea champions were already researching claims and posting their results across the internet in the name of spreading evidence towards their cause.

Dispute Finder is an excellent example of a credibility layer and provides a starting point for Truth Goggles. There are several important concepts borrowed and lessons learned from Dispute Finder, including the planned use of the crowd to build associations, the implementation of the client side, and the development of a flexible API. The goals of Dispute Finder and Truth Goggles are similar, but there is an important difference: Dispute Finder attempts to identify disputed content as a convenience for specific user types, and Truth Goggles attempts to guide all consumers through the process of critical thought.

Truth Goggles and Dispute Finder also diverge in their specific implementations. Truth Goggles places a much greater emphasis on presenting the evidence surrounding a claim in a way that will mitigate the risks of bias and backfire effects and is intended for a more general audience. Dispute Finder, with “skeptics” and “activists” as the primary user types [22], is designed for people with very specific motivations. These motivations are heavily reflected in the tool’s interface.

The Truth Goggles credibility API is also more restrictive about what is considered a valid source. In Dispute Finder sources come from the open Internet and, as a result, is much more broad in deciding what can be considered “disputed.” In Truth Goggles the sources are not contributed by the end user but are instead added to the system only after satisfying the guidelines set forth by the API’s host. Finally, although both tools rely on humans to form associations, Truth Goggles relies more strongly on algorithms to trigger the human interaction. In Truth Goggles users help by occasionally verifying associations, but the identification of those associations is left to the software.

Hypothes.is Hypothes.is [74] is a crowd-driven annotation layer in early stages of development. It is walking in the footsteps of similar concepts such as Third Voice, which closed its doors in 2001 [41]. The team behind this service is attempting to create credible user generated notes at the sentence level across any website on the Internet. For example, when viewing a legal document (e.g. a site’s terms of service) Hypothes.is could help explain what a particular phrase means. The vision is to create a tool where “wherever we encountered new information, sentence by sentence, frame by frame, we could easily know the best thinking on it.” [74]

The major challenge for Hypothes.is is the same as the challenge behind its predecessors:

ensuring credibility and procuring reliable information from authoritative sources. This is also where most of their effort is being spent. The Hypothes.is team is aware that people will attempt to game the system and that controversial content in particular will yield questionable annotations without an incredibly sophisticated algorithm. Compared to Truth Goggles, Hypothes.is is attempting to solve a slightly different and more general problem.

WikiTrust WikiTrust [76] is a browser extension designed specifically for Wikipedia [75] that takes advantage of the way wikis work to estimate the credibility of user contributions. The tool is built around author reputation, and authors are given a score based on how many of their edits are preserved and how often their edits are immediately removed [6]. Using this heuristic, WikiTrust uses this heuristic to highlight content on Wikipedia and reflect predicted credibility.

The highlighting interface is cited as inspiration for Dispute Finder, which by transitivity means it is inspiration for the highlighting interfaces in Truth Goggles. Everything else about this tool, however, is fairly specialized to situations where sources can be assigned a credibility rating and where it is possible to track what content came from those sources. This author-driven approach could potentially work in a broader credibility layer, but for now it is primarily applicable to Wikipedia.

LazyTruth LazyTruth [68] is another credibility layer being developed at the MIT Media Lab's Center for Civic Media, but unlike Truth goggles it is designed specifically for email. It is an inbox widget that automatically detects correspondence containing debunked claims, of which a significant amount are related to politics and political figures. The widget warns recipients about those claims, provides links to debunking resources, and suggests responses. Because LazyTruth is based in email it needs to be able to adapt to messages that evolve over time. It must also be able to work with more casual interactions and account for a completely different consumer mindset than Truth Goggles.

Truth Goggles and LazyTruth share a common goal with different use cases. LazyTruth interfaces have been designed with email communication in mind while Truth Goggles is meant to integrate within a less personal consumption experience. From an algorithmic standpoint, LazyTruth has the benefit that email chain letters follow more specific patterns and, currently, the tool is focused on email-level corrections as opposed to sentence-level. If it were not for this subtlety it would be very likely that LazyTruth and Truth Goggles would already be powered by the same credibility API. As it stands the tools both use similar data sources

Skeptive Skeptive is a credibility interface that highlights any content on the page that goes against the user's personal definition of truth [27]. The developers built Skeptive in response to negative experiences with Dispute Finder, where they felt the tool did not provide relevant information or highlight phrases that they personally held negligible interest in. The system allows users to "train" the interface and create a profile. That profile is used to generate targeted results based on sources those users personally trust. As a result this tool helps people more effectively consume information from their favorite trusted sources.

The unfortunate side effect of this design is that the interface is far less likely to explicitly challenge the user's world view. It will even report different definitions of truth depending on the individual's preferred sources. For instance, a conservative is likely to favor conservative sources and might be told that there are 16 million "illegal immigrants" in the United States while a liberal might be told that there are only 13 million "undocumented immigrants." The end result is an interface that risks the perpetuation of a user's personal vision of truth and makes it easier to disregard challenging information.

Skeptive is driven by users who identify and rate sources in addition to flagging new disputes. The client is implemented as a browser bookmarklet that attempts to detect instances of disputes in the content being consumed across the web, and provides evidence surrounding the dispute according to the user's sources. Although Skeptive is technologically intriguing, the tool has a very different philosophy about truth when compared to Truth Goggles.

2.3 Mental Models

It is hard to disagree with the social value of credibility, but not everyone is as convinced about the impact of fact checks and corrections. There are many ways that media consumers fool themselves when reading information. These traps can cause fact checks to not only fail, but to backfire and result in stronger conviction about false beliefs [49]. Mental pitfalls can dramatically hinder corrective information even among consumers who believe they are actively seeking out accurate information and considering other perspectives. This section describes some of the key biases and backfire effects that were considered in the design of Truth Goggles.

2.3.1 Biases and Effects

Belief Perseverance and Continued Influence It is often more difficult to change someone's mind than it is to convince them of something completely new. There are two psychological quirks which help explain this phenomenon: belief perseverance and continued influence. Belief perseverance is the tendency to continue believing a piece of information despite corrective evidence. The continued influence effect is similar but takes root at a later stage; continued influence refers to the way that false beliefs can affect decisions and analyses even after those beliefs have been corrected.

These effects do not always work the same way, and there are many interaction effects based on the type of information and the context surrounding it. For instance in the case of belief perseverance people are more likely to preserve negative information about political figures. If negative information is corrected the correction is likely to fail, but when positive information is corrected the correction is likely to succeed [19].

For fact checkers the key takeaway is that belief perseverance and continued influence are serious and complicated challenges and that there is no clear and universal solution. While it is useful to understand their nature, the best response may simply be to avoid these complications completely by protecting consumers from situations where the false information would be perpetuated in the first place. This "bullet proof vest" approach has already been introduced and discussed among the leading thinkers in the fact checking space [44].

Confirmation Bias and Disconfirmation Bias Confirmation bias and disconfirmation bias are two sides of the same selective exposure coin. They describe "gut reactions" that content consumers can have when exposed to a piece of information where their opinions are already formed. In the case of confirmation bias an individual is less likely to question (and more likely to accept) a claim that supports his or her existing world view. The result of this effect is that people might not stop and question certain pieces of content even if it deserves a second thought, and will be less likely to challenge their world views.

Disconfirmation bias is the opposite effect. It refers to the tendency to spend more time and mental energy thinking about why information you are inclined to disagree with is wrong. By spending time figuring out how to tear down a piece of information the individual becomes more embedded in their prior beliefs and less likely to consider alternatives. Between these two effects people are less likely to challenge their own world views and more likely to lambast the world views of others.

Enhanced Negation Effect The enhanced negation effect occurs when people reject corrections because the corrections attempt to make far stronger claims than necessary. This was observed in a study where researchers provided two corrections regarding a chemical fire in a warehouse. One group was told that the reports were simply incorrect and that there was no fire. The other group was told that not only was there no fire, but the warehouse has never once contained paint or chemicals. The effectiveness of the more extreme correction was far lower than the more casual correction [16].

The explanation for this effect is that the enhanced negation is much more difficult to believe than the casual false story. In other words, it is easier to believe there was a fire than it is to believe that the warehouse never held chemicals, even if the former was false and the latter was true. For this reason it is important for fact checkers to avoid overselling their corrections.

Familiarity Backfire / Illusion of Truth Effect The familiarity backfire effect refers to the way humans associate familiarity with truth. If a person is able to remember a statement they are more likely to believe that statement. This becomes a problem because even if someone accepts a correction, fact check, or other piece of information one day they could easily forget that correction over time. Months later the individual might remember the false information more strongly than the fact check, and revert to their false belief [67]. This is why it is important to minimize the number of times a consumer reads a fact checked phrase, as each repetition will increase the chances of remembering that phrase instead of the correction.

This effect is the reason that negations are a particularly ineffective form of correction [51]. Words like "not" are very easy to erase from memory, which means fact checks incorporating these words could accidentally re-enforce the very claim that is being disputed. For example, if someone were to assert that "Milhouse is a meme" a wise debunker would correct the statement using completely different language (e.g. "nobody cares about Milhouse"). Unfortunately, many people make the mistake of correcting through negation (e.g."Milhouse is not a meme") [2]. By using negations instead of alternate language it is more likely for participants to remember the core statement (e.g. "Milhouse" and "meme") which is exactly the association that the debunker was attempting to reject.

Hostile Media Perception People with opposing viewpoints can look at a piece of content that has been explicitly written to be neutral and everyone can believe the content is biased against them [73]. This is known as hostile media perception, and it reflects the tendency for passionate individuals to view coverage of polarizing issues as biased in favor of their ideological counterparts regardless of the reality. Hostile media perception makes it particularly difficult to present information about many popular and important issues, where fact checks would likely have the most potential impact and relevance.

There is evidence that hostile media perception may truly be a *media* perception. This would mean that it does not exist in more casual information settings and that the collective finger pointing may be restricted to traditional media reports [28]. The study that first introduced this potential caviat did not explore additional text sources such as blogs or social media, but it is an interesting question in the context of credibility interfaces: if information is not presented as mass media it is possible that people consume it with fewer ideological defenses.

Ideology, Identity, and Race Many of these effects indicate that when people consume information they are struggling to maintain their identity. The more that people are reminded of how new information might challenge the things they identify with, the less likely they are to be open to new information. For example, writing that contains ideological cues (e.g. references to political figures or charged language) will often incite reactions along party lines [15]. Similar effects have been observed with other identifying factors such as race and age [39]. Even something as simple as the photographs presented alongside an article can have a significant influence over the way information is perceived [51].

Motivated Reasoning Some of the leading theories on how we interpret new information is based on the premise that people are driven by existing beliefs when trying to form meaning. When exposed to new claims people attempt to find a way to fit that new information into the complicated internal network of understandings that coalesce to form their world view. This process is known as motivated reasoning, which more generally refers to the idea that people let motivations affect their ability to interpret information. It explains why many people struggle to change their perspectives when faced with evidence that runs counter to their beliefs.

In the context of political information, consumers appear to be torn between the desire to be accurate [9] and the desire to perpetuate their partisan beliefs [40]. Political information tends to trigger partisan motivations to the point where even well intended readers find it "impossible to be fair-minded" [70]. The situation seems dire, but it has been shown that people are more likely to consider corrections if steps have been taken to calm motivating factors associated with identity protection [50].

Overkill Backfire Effect When readers feel overwhelmed with too much information or when an explanation is more difficult to process it becomes less likely that they will accept a correction or fact check [20]. This phenomenon is known as the overkill backfire effect, and it exists because people tend to associate simplicity with truth. More complicated explanations take more effort to process and, as a result, are less likely to have a productive impact.

This effect is particularly troublesome because some false beliefs stem from a lack of information. Fact checkers and credibility interfaces need to present enough information to explain the correction, but not so much that the reader might be confused or have to spend too much effort understanding. The Debunking Handbook suggests an approach that allows the reader to pick the level of detail based on their interest, thus making it possible to satisfy the curious while still addressing the key concerns for those with less attention to spend [20]. It has also been shown that use of graphical information can help corrections stick [50].

Selective Exposure and Filter Bubbles Selective exposure refers to the natural inclination to seek out sources that reinforce one's existing world view. This is similar to confirmation bias, but is distinct in that selective exposure actually affects what content people consume as opposed to affecting how they interpret that content. Selective exposure may be invoked at the moment of source selection (e.g. selecting a news source with a known slant or perspective), or it may not propagate until the moment people decide which content to view (e.g. reading a specific article based on an attractive headline).

Content sources and curators explicitly take advantage of selective exposure. Eli Pariser's "Filter Bubble" [52] refers to algorithmic embodiments of selective exposure. As web services incorporate demographic and activity metrics (e.g. click rates) to predict interest in new content, consumers are not even given a chance to make the mistake of selective exposure themselves. The experience becomes personally tailored and devoid of the "informational veggies" that human editors used to be able to include themselves [52].

2.4 Alternate Responses

When people are unhappy with information they have several possible courses of action. In many cases they do nothing, or simply complain to the media source through emails, social platforms, or angry comments. In other cases they might decide to stop using the source entirely, and indeed it has been shown that unhappiness with mainstream media leads to increased adoption of alternate sources for information, such as social media [71]. There are also more creative steps, and this section describes a few examples.

2.4.1 Content Modification

In 2008 the homophobic "American Family Association" developed a content filter that would automatically replace the word "gay" with the word "homosexual" to incite their visitors with divisively charged language [25]. The manipulation became infamous when their script failed to take into account the fact that "Gay" is also a last name. This oversight resulted in laughable headlines like "Tyson Homosexual wins 100m dash." Regardless of ethical intent or quality of execution, unhappy content consumers can take matters into their own hands, and there are several tools that exist to empower them.

Reamweaver Reamweaver is a tool published by the dtournement group The Yes Men. This project, whose name is a play on the popular web publishing tool Dreamweaver [7], "allows users to instantly 'funhouse-mirror' anyone's website in real time" [63]. It can download a site's content, copy it locally, and run a search-and-replace operation to change key words and phrases as per the user's command to create a parody version. For example a user might want to remove charged or partisan language from an article in order to make it feel more balanced.

The tool is no longer available and was originally released to prove how easy it is for anyone to remix web content [42], but the motivation remains true: people on the Internet do not need to accept the framing decisions made by corporations and third parties.

NewsJack NewsJack is a project I developed alongside Truth Goggles with a related goal: to increase media literacy and remind people of the importance of critical analysis. NewsJack is a "media remixing tool" that makes it possible for anyone to change the content of any web site and publish their work [65]. There are two types of user within NewsJack: the content producer and the content consumer.

Consumers exposed to remixed content through shared URLs have ten seconds to consider what they see on the page without any explicit warnings about what is happening. After that time expires a NewsJack banner is rendered on the top of the screen alerting everyone to the fact that the page is the result of a remix. Until that banner appears the tool hopes to inspire reflection about the world or the media's representation of that world[2]. Once the temporal window is closed the hope is that users will go on to be more thoughtful about all content on the internet regardless of formatting or apparent source.

Producers have an opportunity to challenge framing decisions made by third parties. To this end NewsJack is being used by a local Boston organization called Press Pass TV [72] to help urban youth challenge the way their communities are represented in the media. During critical media literacy workshops the kids are given the chance to rewrite articles about violent crime in their neighborhoods and asked to change the language to reflect the way they would have preferred to be represented. NewsJack hopes to remind users that framing is important and help them become more aware of framing decisions as they read content.

2.4.2 Self Censorship

Another class of response is to undergo self censorship by artificially blocking out categories of content from articles, television programming, or any other medium. This is a fairly popular practice in response to unwanted advertisements on the Internet through tools like the AdBlock browser extension [5], which automatically detects and removes advertisements

[2] It is more likely that users will reflect on the world's vast supply of amusing pictures of cats, and how much more amusing those pictures become when plastered on the front page of the New York Times. [32]

from web sites and videos. Some developers have tried to replicate the concept in other contexts.

Arduino TV Mute The Arduino TV Mute is a small hardware hack that is able to analyze a closed captioning stream and respond to certain key phrases by simulating a remote "mute" signal to mute the TV [61]. The intended application is for celebrity names such as "Kardashian" or "Snooki" but the concept is clearly generalizable. For instance a person could easily program this to mute any time a loaded political term is used.

Browser Plugins There is a class of browser plugin that will automatically detect and block out specific phrases and links referencing those phrases. For instance, when Charlie Sheen was making headlines in 2011 one frustrated consumer created the Tinted Sheen browser plugin to block his name from the browser [29]. Similar tools have been created for political celebrities like Sarah Palin and Ron Paul. These tools have not caught on and appear to be more of a social statement than a social movement, but they exist to remind people that consumption, even at the sentence level, is a choice.

Chapter 3

Design

With so much prior work in the spaces of fact checking, credibility, and mental models, Truth Goggles is not starting from a blank slate. This section describes how past efforts and research have informed the design and goals of the system.

3.1 System Goals

3.1.1 Mission

The primary goal of Truth Goggles is to help people think critically at the moments when it matters most. Users of this tool should hold beliefs because they took a moment to really think about the issue at hand. Below is an explanation of what this mission means in practice.

Triggering Critical Ability Truth Goggles is not a vehicle for truth dissemination. Fact checks should be presented for the user's consideration, but they should not be presented as "universal truths" or as though the user is expected to simply accept their verdicts. Such a presentation would result in a system that is no better than existing consumption processes. Users who trust the sources would believe the results while users who are skeptical of the sources would reject the results.

Instead Truth Goggles is a vehicle for critical ability dissemination. It leverages the critical thinking of third parties to trigger the critical ability of the user. It then provides additional context to help that user more easily act on the red flags if they choose. This is a far less authoritarian approach, and is one that should result in a more savvy consumer base instead of a more complacent one.

Popping Filter Bubbles There are already many ways for people to re-enforce their existing beliefs. One of the goals of Truth Goggles is to help users actively question information that they might have previously taken for granted. In order to do this, Truth Goggles needs to help bypass "gut feelings" such as confirmation and disconfirmation bias. It also needs to motivate the user to explore information that they might naturally have ignored. These goals are reflected in the user interface, but they are also the reason that users are not able to choose their sources. Everyone is exposed to the same set of content and the same diversity of perspectives. Anything less would make it easy for users to maintain a narrow understanding of the issues.

3.1.2 User Types

Anyone consuming content is a potential user of Truth Goggles, either through explicit activation or automatic embedding. These users will approach the Truth Goggles interface with a wide variety of mindsets and expectations. Below is a list of user types that represent people who may be particularly passionate about their experience with the tool.

Media Skeptics Many people are skeptical of the content they see online and in the mainstream media. They simply want to be able to benefit from narratives while still viewing content that they can trust. Truth Goggles offers these people a way to help constructively address some of those trust concerns. As they read content online, they will be able to more effectively question claims and in some cases get answers to those questions.

Curious and Open Minded Users Some people want to better understand additional perspectives about the content they are reading. They are not necessarily as concerned about accuracy and conformity to their world view so much as they are concerned about seeing what others have to say. Open minded users are going to use Truth Goggles as an exploratory interface to help them dive deeper into a given claim in the name of becoming more informed.

Opinionated Users Many users will already hold strong opinions about the information highlighted by Truth Goggles, and these will be the most difficult use cases. These are people who may have activated Truth Goggles out of curiosity, because they are seeking out more evidence to support their opinions, or because Truth Goggles is embedded in the web pages they are viewing. It is also possible that these users think they are open minded but there are some particular claims or issues they feel particularly passionate about.

3.1.3 Use Cases

Bookmarklet One way to activate Truth Goggles is to consciously decide to "turn it on." This is the bookmarklet use case, which requires active participation from the user in order to function. In this situation the user must remember to use Truth Goggles, which means that person is either always activating it or that something about a page's content causes him or her to want to activate it. This use case may be less effective at combating biases because the decision to invoke Truth Goggles will also be subject to those biases.

Browser Plugin Another way to activate Truth Goggles is through a browser extension that automatically runs the tool on every web page. This has the benefit of providing a passive activation experience, meaning the user would not have to consciously think about credibility for the tool to have an effect. Unfortunately this requires a user that is dedicated enough to the tool to install it into their browser.

Publisher Activation Content publishers might realize that accuracy and credibility is something their readers value and decide to embed Truth Goggles onto their pages. This would automatically activate Truth Goggles for everybody viewing their content without having to trigger or install it. This use case in particular means that Truth Goggles must be intuitive even to people who have never heard of it before.

3.1.4 Open Source

Truth Goggles provides an important civic service, so it is appropriate that the system has been developed in the open. It is likely that instances of Truth Goggles will be used by fact checking organizations to better share their work with the world, but even more importantly the concepts and systems behind Truth Goggles may help future researchers and developers improve or modify the system for the betterment of the global community. Because of this, Truth Goggles has been designed to be easy to set up, install, and contribute to.

3.2 Design Principles

3.2.1 Responding to Effects

In order to be successful Truth Goggles needs to cut through some of the biases and backfire effects that plague the meaning making process, specifically when it comes to corrections and potential misinformation. Below is a list of some of the concepts incorporated into the design of Truth Goggles. Several of these principles came from published best practices [20] [51], while others are more experimental.

Minimize Repetition In response to the familiarity effect, Truth Goggles will attempt to minimize the number of times a fact checked claim is repeated. This is accomplished by using the instances of the claims that appear in the page content instead of listing the phrases in a separate navigational interface. Some variations on the interface go so far as to censor the phrases from the original content until the user is ready to consider them more carefully.

Avoid Overwhelming the User Users can feel intimidated and confused by information excess and overcomplicated interfaces. Truth Goggles attempts to create a simple user experience that avoids presenting more information than necessary to trigger thought. The Debunking Handbook suggests that the best approach is to provide different levels of information [20], allowing more curious users to explore the additional information while protecting the more casual user from information overload.

Prevent Selection Biases Skeptive shows that it is possible for a credibility layer to facilitate selection, confirmation, and disconfirmation biases, but Truth Goggles aspires to avoid this practice. There are two points in the system where a user could undergo selection bias:source selection and phrase exploration. To avoid biases at source selection users are not able to tailor the sources available in Truth Goggles based on brand or political leaning, as Truth Goggles does not support "disabling" or "enabling" sources. Truth Goggles also tries to put only minor emphasis on the sources of content in order to help prevent user-driven selection biases.

Avoiding biases at the point of phrase exploration is more difficult, but still possible. The more convenient it is to skip the exploration process, the more power a user has to invoke biases. For instance one could imagine an interface where users were forced to spend exactly five minutes looking at the context surrounding every fact checked claim regardless of how they felt about the content of the claim. This would not allow them to skip over content they already agreed with or spend more time finding flaws with content they already disagreed with. Of course, this would also be a terrible user experience. Truth Goggles must strike a balance between convenience and protection.

Do Not Conflate Partisan and Non-Partisan Sources The Hostile Media Effect makes it easy to call credible information biased so Truth Goggles needs to do as little as possible to legitimize these concerns. This means that any source that is widely considered partisan should not be included in the tool at the same level as sources explicitly attempting to be independent [20]. For this thesis only non-partisan sources are to be considered for inclusion. Eventually users might be able to switch between "partisan" and "non-partisan" modes, where partisan mode would replace independent sources with liberal *and* conservative sources.

Do Not Threaten Identities When people are reminded that information might go against concepts they identify with they often respond by rejecting the incoming information. The more that Truth Goggles can remove information from potentially manipulative framing and reduce ideological cues, the more effective it will be at creating an environment where it is safe to have an open mind. It is also important that users realize the goal of the tool: Truth Goggles does not care *what* people think, it just cares *that* they think. Users of Truth Goggles should not feel threatened or manipulated by the presentation.

3.2.2 Responding to Technology

Human mistakes are far more forgivable than algorithmic mistakes, but the paraphrase detection processes powering Truth Goggles will never be flawless. There will be phrases that are highlighted incorrectly, claims that are incorrectly associated with fact checked content, and context that is completely irrelevant to a given phrase. It is important that the Truth Goggles interface is designed with this reality in mind. For example, it would be inappropriate to say that a piece of content is false based on a mistaken association. It is much more forgivable, however, to say that a piece of content is interesting based on a mistaken association. This is yet another reason why Truth Goggles has been designed to inspire thought as opposed to dictate it.

3.3 Existing Approaches

Truth Goggles is fortunate to have a rich background to work from outside of theories and ideas. Several of the concepts behind the overall system structure and even certain interface components are inspired by past projects. The idea of using highlights to trigger attention and the incorporation of users as a method of building associations between fact checks and new content was borrowed from Dispute Finder [22]. Skeptive also applies some of these techniques, but the key takeaway from Skeptive was one of contrast. Skeptive helped me realize that allowing users to choose sources might facilitate more polarized world views.

Dispute Finder's system architecture provides an incredibly helpful starting point. The project is broken into a javascript client that can either be invoked by the user as a bookmarklet, by the browser as an extension, or by the producer as a piece of embedded code. The API is also flexible enough to support more than one type of client, and to be able to harness the crowd to identify new paraphrases [22]. Truth Goggles was moving in these directions already, but these successful practices from Dispute Finder's architecture were quickly incorporated.

Chapter 4

Implementation

This section explains the Truth Goggles system in detail, from the client side design to the scrapers and data access APIs. It also explains the decisions and explorations that led to the final prototype as well as notes from prior iterations.

4.1 Methodology

4.1.1 Processes

This project began with significant questions and significant risks. What would be considered truth? How accurately would the system be able to automatically find paraphrases? What pieces of the puzzle could be reliably crowd sourced? Would users trust the system? Would users trust the sources? With so many different ways that the idea could fail it was important to pursue an architecture that would make it easy to change course and follow the most promising paths.

This need for flexibility affected the system design, the development process, and ultimately resulted in a more elegant architecture. Risks were clustered into decoupled components that could be iterated on in isolation. This meant that individual road blocks and failed experiments would not hurt the larger development goals. For instance, the fuzzy matching algorithms were separated from the credibility API, which itself was also designed to support more than just consumer-facing credibility layers.

4.1.2 Timelines

Truth Goggles was researched, developed, and evaluated over the course of eight months. The first three months were spent researching the rich landscape of fact checking, bias, critical media consumption, and credibility. This research was applied over the next four months to direct the design and implementation of the system itself. The remaining time was spent evaluating the system.

4.1.3 Infrastructure

Truth Goggles was written using PHP [58], Python [59], and JavaScript (jQuery in particular) [36]. More specifically, the client side was written with JavaScript, the credibility API was written with PHP, and the more sophisticated Natural Language Processing (NLP) algorithms / fuzzy matching API was written in Python. The data surrounding the credibility API is stored in a MySQL database [47]. Version control was done using Git and the code is publicly accessible on GitHub [64]. The service is served on a traditional LAMP stack being kindly hosted at the MIT Media Lab.

4.2 System Structure

Truth Goggles is separated into three primary components: the client, the credibility API, and the fuzzy matching engine. Each component represents a set of major challenges and have been decoupled from one another and iterated on separately.

4.2.1 Client

The client interface is implemented as a browser bookmarklet written in JavaScript and jQuery. When activated, the bookmarklet injects the latest Truth Goggles scripts and stylesheets into the web page being actively viewed. The script extracts all text from the page and sends the content to the server side "credibility API." This API identifies existing and potential fact checked phrases as described in the section of this document titled "Credibility API."

Once the phrases of interest have been returned the client script modifies the page's Document Object Model (DOM) to inject several classes of inline wrapper around all relevant content. At this point all known claims, potential claims, and plain content in the DOM are programmatically accessible as interactive objects. The wrappers also associate data with that content, such as the ID of the relevant fact checked claims and the page-level index of the most recent fact checked claim.

Wrapper Classes There are three types of wrapped content which make up the Truth Goggles credibility layer.

Snippets Snippet wrappers contain previously known instances of fact checked claims. Depending on the interface goal, this content may be highlighted, blocked out, or otherwise treated specially. A snippet represents a claim worth thinking about more carefully from the perspective of the system.

Potential Snippets Potential snippet wrappers contain page content that has been identified as being a possible match with a known fact checked claim. The accuracy of the match is not yet strong enough to be considered an actual snippet. Depending on the interface goal this content may be incorporated into a crowd-based validation system or it might be presented to the user as a claim that is worth thinking about more carefully from the perspective of the system.

Interim Content Content that appears between snippet wrappers is also accessible by code. These wrappers are labeled in terms of the snippets that appear before them, which makes it possible to do things like blur out the text that appears after an unexplored phrase.

4.2.2 Credibility Backend

The credibility backend consists of an API, a scraper, and a data management system. This component is responsible for making fact checks accessible to front ends, scraping the fact checks from credible sources, collecting crowd sourced data, and interacting with the fuzzy matching API. This portion of Truth Goggles was designed with the intent to eventually power more than just the Truth Goggles web client.

The entire service is written in PHP and content is stored in a MySQL database. The API itself is RESTful [62] and all calls are accessible through basic HTTP GET and POST requests. Interactions with the fuzzy matching API make it possible to detect potential snippets.

4.2.3 Matching Engine

The matching engine is responsible for identifying likely paraphrases of known fact checked claims. It is accessed directly by the credibility backend, as opposed to being a client facing API. This section of Truth Goggles has undergone two experimental iterations. The first version was written using PHP and took a statistical approach. The second version was implemented in Python using the Natural Language Toolkit (NLTK) [56] and Luminoso [54].

Neither version of the matching engine currently yields useful results as a fully automated system, but they both represent important steps towards the goal of identifying new instances of fact checked claims. I also propose a future design that combines the output of the matching engine with human assistance in order to create a semiautomated snippet detection process.

4.3 Client

This section explains the various components incorporated into the Truth Goggles user interface. From the perspective of the users these components formed a unified experience.

4.3.1 Control Pane

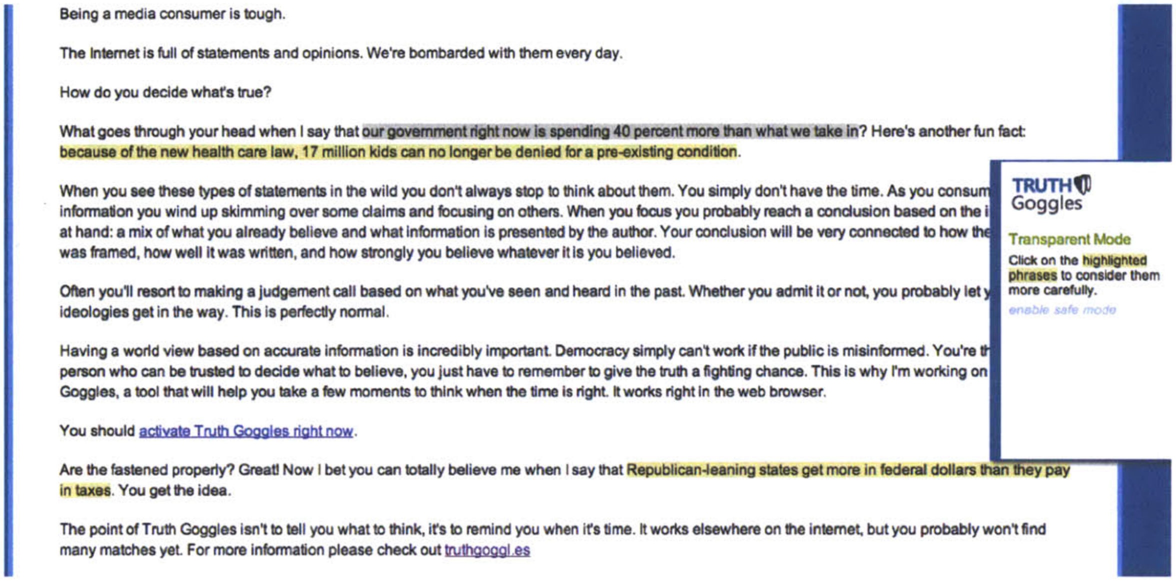
The first thing a Truth Goggles user sees upon activation of the tool is the control pane. This collapsable pane is rendered immediately and displays the status of Truth Goggles as well as instructions for the user. After the server responds with results, this pane is refreshed to provide information about the active credibility interface. Although the interface components are meant to be intuitive, the control pane contains simple instructions to guide the user through the process of interacting with snippets. This component also contains basic control interfaces which allow the user to toggle through the various "modes" of Truth Goggles. In the study, participants couldn't change modes.

4.3.2 Credibility Layer

The bulk of the time spent on the Truth Goggles interface was focused on the nuances of the credibility layer itself. This refers to fact-checked content that is "called out" to the user inline, and at the point where the user would have consumed it under normal circumstances. The following is a list of some of the cues considered: Font weight, font color, font family, font legibility, font size, background color, addition of "badges" or "flags" around the phrase, animation or movement (e.g. shaking, bouncing letters, etc), and addition of hyperlinks.

For this thesis I explored three different versions of the credibility layer. Each layer strikes a different balance between protection and convenience. For instance, an interface that forces a user to interact with fact checked claims as they read the page may prevent selection biases, but might also create a more disjointed experience.

Highlight Mode Highlight mode is intended to be the least obtrusive of the three interfaces. It highlights fact-checked phrases on the page in a light yellow background. Users can decide to click on a highlighted phrase to activate the inspection pane and access additional information. Once the inspection process is complete the highlight is "dulled" and turned gray to indicate that the details have already been explored. Any other reference to the related phrase that appears on the page is also considered "clicked" and becomes dulled.



Being a media consumer is tough.

The Internet is full of statements and opinions. We're bombarded with them every day.

How do you decide what's true?

What goes through your head when I say that our government right now is spending 40 percent more than what we take in? Here's another fun fact: because of the new health care law, 17 million kids can no longer be denied for a pre-existing condition.

When you see these types of statements in the wild you don't always stop to think about them. You simply don't have the time. As you consum information you wind up skimming over some claims and focusing on others. When you focus you probably reach a conclusion based on the i at hand: a mix of what you already believe and what information is presented by the author. Your conclusion will be very connected to how the was framed, how well it was written, and how strongly you believe whatever it is you believed.

Often you'll resort to making a judgement call based on what you've seen and heard in the past. Whether you admit it or not, you probably let y ideologies get in the way. This is perfectly normal.

Having a world view based on accurate information is incredibly important. Democracy simply can't work if the public is misinformed. You're th person who can be trusted to decide what to believe, you just have to remember to give the truth a fighting chance. This is why I'm working on Goggles, a tool that will help you take a few moments to think when the time is right. It works right in the web browser.

You should activate Truth Goggles right now.

Are the fastened properly? Great! Now I bet you can totally believe me when I say that Republican-leaning states get more in federal dollars than they pay in taxes. You get the idea.

The point of Truth Goggles isn't to tell you what to think, it's to remind you when it's time. It works elsewhere on the internet, but you probably won't find many matches yet. For more information please check out truthgoggl.es

TRUTH Goggles

Transparent Mode

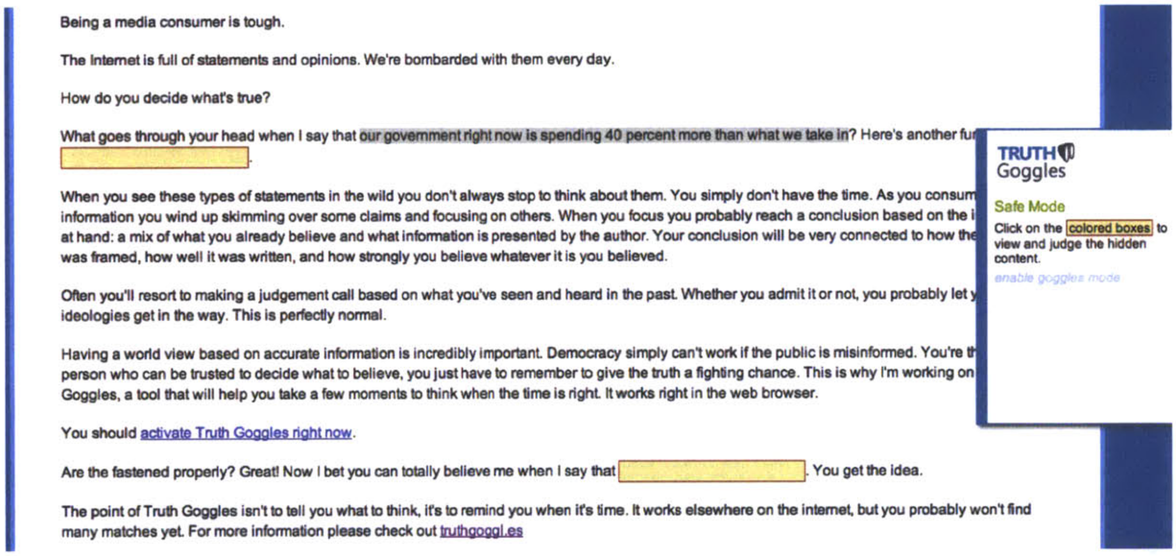
Click on the highlighted phrases to consider them more carefully.

Figure 4-1: An example of highlight mode.

Compared to the other interfaces highlight mode takes the fewest number of steps to prevent backfire effects and biases. For instance, it would be easy for the user to ignore the highlight completely, as there is nothing forcing an interaction. By providing what is essentially the bare minimum it gives the users more control over their decisions to pursue content. This allows more room for selection bias, confirmation bias, and disconfirmation bias. Users could decide to "click through" based on how informed, uninformed, or passionate they feel about particular claims.

Because highlight mode is minimally interfering, it may also be the most user friendly. If less control over consumption process creates a more pleasant user experience and increased adoption rates for Truth Goggles then this could make up for the weaker mechanisms. A greater number of users could cause a net increase in defense against overall misinformation despite being less effective from a theoretical perspective.

Safe Mode Safe mode is designed to take the strongest steps toward preventing the spread of misinformation. It goes beyond simply highlighting content by actually blocking it out. Each detected phrase is replaced with an opaque colored box which, when clicked, reveals the hidden content along with the inspection pane. After the phrase is considered, the block is removed and the phrase appears in its original context with the same gray highlight as the other modes.



Being a media consumer is tough.

The Internet is full of statements and opinions. We're bombarded with them every day.

How do you decide what's true?

What goes through your head when I say that our government right now is spending 40 percent more than what we take in? Here's another fu
.

When you see these types of statements in the wild you don't always stop to think about them. You simply don't have the time. As you consum
information you wind up skimming over some claims and focusing on others. When you focus you probably reach a conclusion based on the i
at hand: a mix of what you already believe and what information is presented by the author. Your conclusion will be very connected to how the
was framed, how well it was written, and how strongly you believe whatever it is you believed.

Often you'll resort to making a judgement call based on what you've seen and heard in the past. Whether you admit it or not, you probably let y
ideologies get in the way. This is perfectly normal.

Having a world view based on accurate information is incredibly important. Democracy simply can't work if the public is misinformed. You're th
person who can be trusted to decide what to believe, you just have to remember to give the truth a fighting chance. This is why I'm working on
Goggles, a tool that will help you take a few moments to think when the time is right. It works right in the web browser.

You should activate Truth Goggles right now.

Are the fastened properly? Great! Now I bet you can totally believe me when I say that . You get the idea.

The point of Truth Goggles isn't to tell you what to think, it's to remind you when it's time. It works elsewhere on the internet, but you probably won't find many matches yet. For more information please check out truthgoggl.es

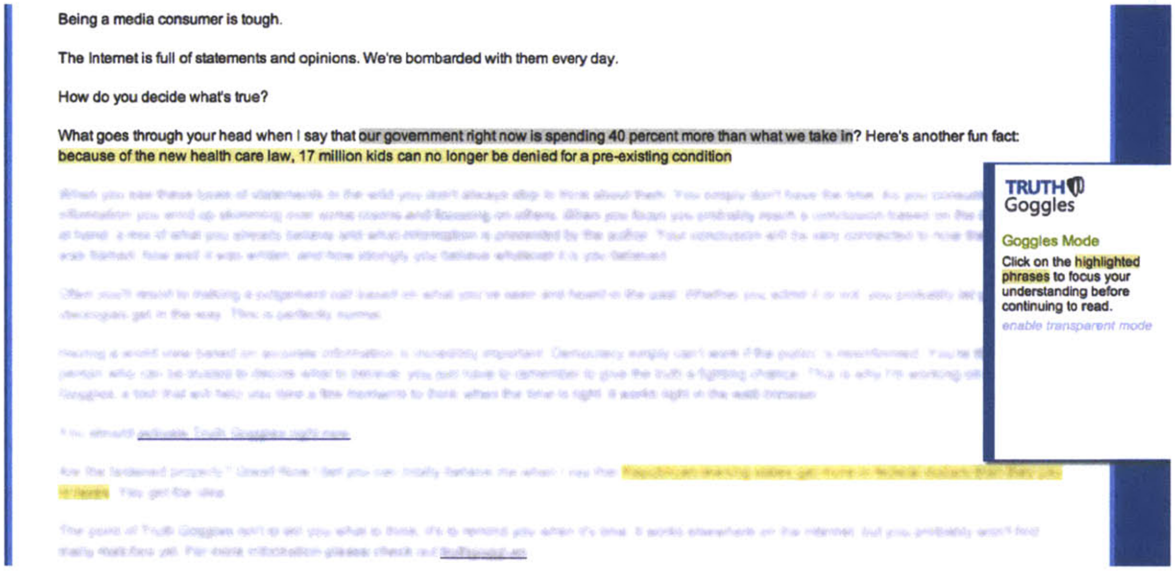
Click on the colored boxes to view and judge the hidden content.

Figure 4-2: An example of safe mode.

This interface protects the user from exposure to potentially dubious content until the point where they are ready to think about it carefully. This helps prevent the familiarity backfire effect by minimizing the number of times a user is exposed to that content. It also helps combat confirmation, disconfirmation, and selection bias by separating the content of a claim from the user's decision to explore the claim. This interface also weakens the content author's ability to frame the phrase, as the user is first exposed to the content outside of the article context, and is able to reach a more independent conclusion.

The obvious downside to safe mode is that it does not conform to the way many people read information online. It forces users to either go through the article searching for colored boxes before reading the content or to interrupt their reading mid-sentence to explore a claim. There are possible additions to both the control pane and the inspection pane which could help mitigate these concerns. For instance, making it easy to discover the boxes and phrases from the control pane would promote the former use case, while incorporation of additional article context into the inspection pane could dampen the effects of the latter.

Goggles Mode Goggles mode is very similar to highlight mode with a key difference: it forces users to interact with the claims in order to continue reading the rest of the article. The hope is that goggles mode can guide the user through the critical thinking process. Just as before, fact-checked phrases on the page are highlighted in a light yellow color. All content following the highlighted phrase is blurred out and made illegible. In order to focus the rest of the article the user must click on the phrase and activate the inspection pane. After closing the inspection pane the "dulling" effect occurs on the highlighted phrase and the blur effect is removed from the rest of the text up to the end of the next unexplored claim.



Being a media consumer is tough.

The Internet is full of statements and opinions. We're bombarded with them every day.

How do you decide what's true?

What goes through your head when I say that our government right now is spending 40 percent more than what we take in? Here's another fun fact: because of the new health care law, 17 million kids can no longer be denied for a pre-existing condition

TRUTH Goggles

Goggles Mode

Click on the highlighted phrases to focus your understanding before continuing to read.

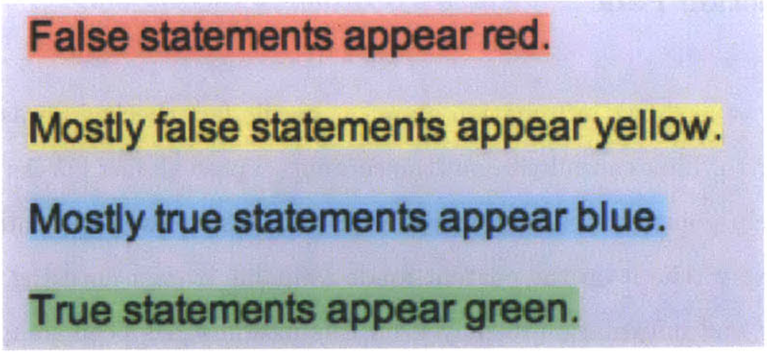
enable transparent mode

Figure 4-3: An example of goggles mode.

Goggles mode is intended to be a compromise between highlight mode and safe mode. Users are able to read the article much more naturally, completing the end of a sentence before entering the inspection pane, but they are still forced to interact with the claim in order to continue reading. This achieves similar protections against confirmation, disconfirmation, and selection bias without as many negatives. It would be possible to create a "skip" button which would allow the user to bypass the inspection pane while still continuing through the article, but I feel that this is too similar to highlight mode to be interesting.

Common Concerns There are several concerns that guided the design of all three interfaces, particularly surrounding the amount of information to present at the level of a credibility layer. All three credibility layers explicitly avoid any situation where the user could believe they are reaching an informed conclusion through Truth Goggles based on the credibility layer alone. The only information a user can get from the credibility layer is that a phrase was fact checked by someone.

This is not necessarily true for all credibility layers. For instance, one of the original designs of highlight mode used colors to reflect different levels of gauged accuracy for fact checked phrases. This was ultimately replaced with the universal application of a more neutral light yellow.



False statements appear red.

Mostly false statements appear yellow.

Mostly true statements appear blue.

True statements appear green.

Figure 4-4: The original highlight interface.

As described earlier, the key motivation behind the shift towards a single highlight style was the concern that users would use the more informative interface to think even less carefully than they normally would. Users who trusted the sources behind Truth Goggles could simply accept or reject a claim at a glance according to what color it was highlighted. Since the goal of Truth Goggles is to enhance critical thought, rather than simply disseminate third party perceptions of what is true, this would have been a catastrophic outcome.

Another reason to standardize the highlight color is that a blunt interface could easily exacerbate backfire effects, increase polarization, and subsequently decrease trust in the tool. For instance, if the phrase "global warming is real" were highlighted in green for "true" then climate skeptics might be less likely to explore that highlight, might assume that the tool is biased, and might become less trusting in all future highlights. Truth Goggles attempts to prevent this type of response by creating a separation between the thought trigger and the supporting information.

The final reason for weakening the highlight interface is in order to mitigate risks surrounding improperly identified content. No matter how accurate the paraphrase detection process becomes there will always be room for error. It is far worse to mislabel a phrase as false than it is to mislabel a phrase as worth thinking about. In the multi-colored highlight interface those mistakes would have been far more damaging.

4.3.3 Inspection Pane

The inspection pane shows information about a specific fact checked phrase. It is designed to help the user to think carefully about the content, expose all fact checks surrounding the content, and help point the reader towards additional resources. The content appearing in this pane is deeply reliant on the content made available by fact checking services, and all of the language and information displayed on this interface has been scraped from a third party.

There are several ways that pane could have been implemented. It could be a widget that floats off on one side of the screen, a box that appears next to a piece of content in response to some type of user action such as a hover, or even an information summary embedded inline. After considering the options I decided to implement the inspection pane as a modal window, something that would remove focus from all other content on the page and force the user to home in on the Truth Goggles content.

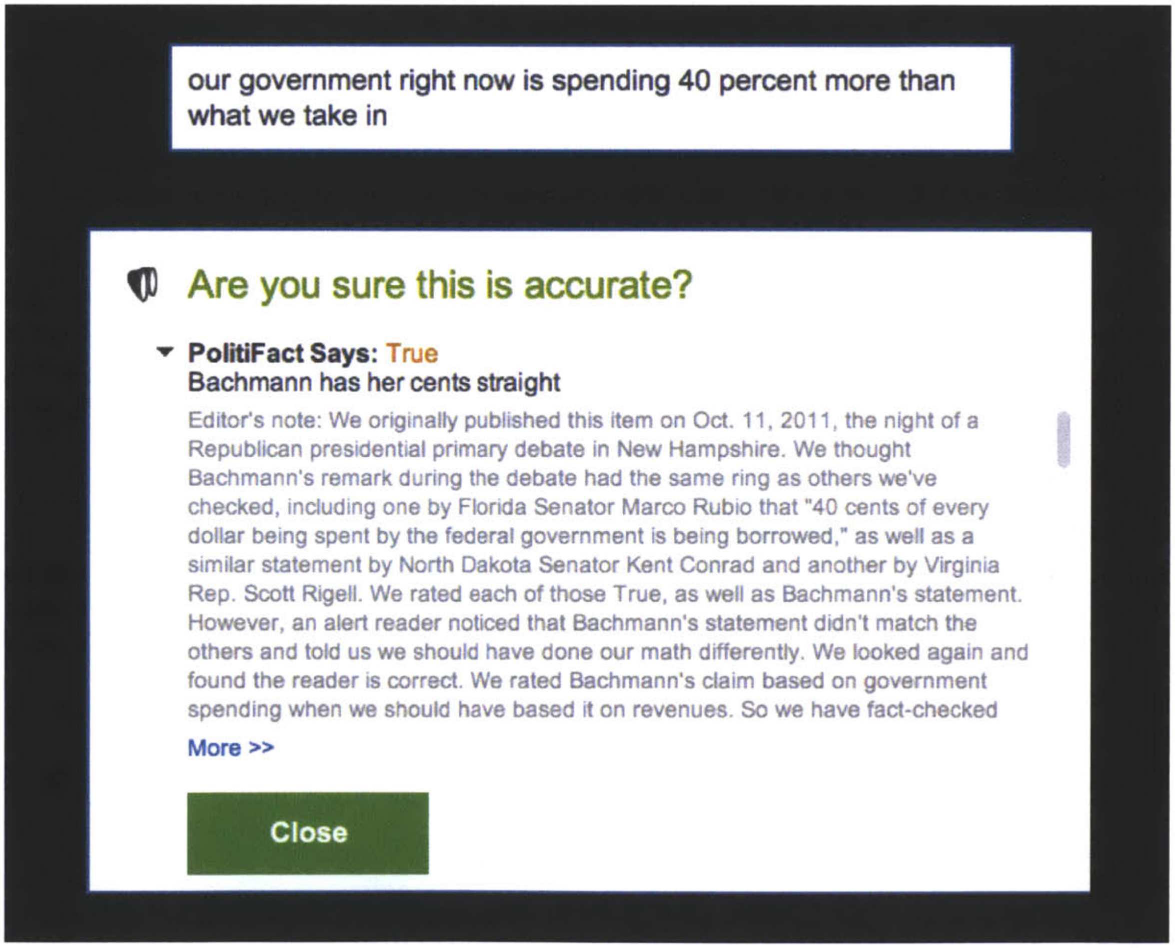
This choice was made in order to decrease the risk of ideological backfire effects and weaken the impact of manipulative framing. For instance if the buildup to the claim was being spent supporting or lambasting the claim (e.g. "can you believe that some moron said X?"), it is likely that removing the leading pretense during the point of consideration would enable a more fair analysis. It should also help the reader avoid issues that arise from known manipulations such as images and ideological cues.

Triggering the pane The trigger mechanism for the pane went through two iterations. The original implementation of Truth Goggles used a simple "hover" effect to trigger the content, meaning a user could initiate the analysis without even having to click. Once the decision was made to treat the pane as a modal it became immediately clear that this was far too easy to trigger accidentally, and since modal views require a click to close this caused a poor user experience. Ultimately the decision was made to require a click to activate the content. This decision aligns with the hyperlink metaphor that internet users are already quite familiar with.

There are other ways that the information pane could have been triggered if the information pane had a non-modal format. For instance if the content were appearing inline then a time-based trigger (when a claim has appeared on screen for a certain amount of time) or a scroll-based trigger (when a claim reaches a certain point on the screen) might work perfectly. Neither of these triggers would have required explicit user action, however, which would have shifted the tone of Truth Goggles from an exploratory tool to an information dissemination tool.

Content in the pane The inspection pane contains a claim, a prompt, and a list of evidence. The claim that appears is the fact checked claim as rated by the third party service. It has been removed from the article context in order to minimize risk of ideological cues that might get in the way of careful consideration. The user is then asked if they are sure the statement "is accurate" in an attempt to instill healthy skepticism.

To help the user act on that question they are shown a list of third party verdicts surrounding the claim. Each item on the list shows the source, the source's verdict (e.g. true, half true, etc), a short phrase describing the reason for that verdict, and the beginning of the long explanation of the reasoning that led to their decision. Users are then able to click "More" to be taken to the source's explanation, if they are interested in reading the entire reasoning process. As has already been mentioned the Truth Goggles prototype uses only one source, PolitiFact, although multiple sources are supported by the system.



our government right now is spending 40 percent more than what we take in

Are you sure this is accurate?

▾ **PolitiFact Says:** True
Bachmann has her cents straight

Editor's note: We originally published this item on Oct. 11, 2011, the night of a Republican presidential primary debate in New Hampshire. We thought Bachmann's remark during the debate had the same ring as others we've checked, including one by Florida Senator Marco Rubio that "40 cents of every dollar being spent by the federal government is being borrowed," as well as a similar statement by North Dakota Senator Kent Conrad and another by Virginia Rep. Scott Rigell. We rated each of those True, as well as Bachmann's statement. However, an alert reader noticed that Bachmann's statement didn't match the others and told us we should have done our math differently. We looked again and found the reader is correct. We rated Bachmann's claim based on government spending when we should have based it on revenues. So we have fact-checked

Figure 4-5: An example of the inspection pane.

The decision to reveal source names (e.g. PolitiFact) at such a high level was not trivial.

Indeed, it may be safer to obfuscate this information in order to prevent lazy acceptance or rejection of verdicts based on variant levels of brand trust. For the current iteration the brands are made prominent in order to make it clear that the verdict is not intended to be an absolute truth, but is merely a judgement made by a third party. Presenting the source so clearly helps communicate that it is not Truth Goggles making the claim. Truth Goggles is simply presenting the claim for the user's consideration.

Third Party Concerns The inspection pane, more than any other piece of the interface, is limited by the information available from the data sources. Almost all content appearing on the information pane has been taken from a third party. This means that the tool is reliant on those third parties to follow best practices with regards to biases and effects. For instance if a fact check explanation repeats paraphrases of the fact checked claim, negates the original claim, or uses ideological cues then this could have an adverse effect on the effectiveness of Truth Goggles.

4.4 Credibility Backend

This section outlines the data behind Truth Goggles, the API served to access and modify that data, and the services used to populate it.

4.4.1 Models

4-6 An Entity Relationship diagram of the Truth Goggles database.

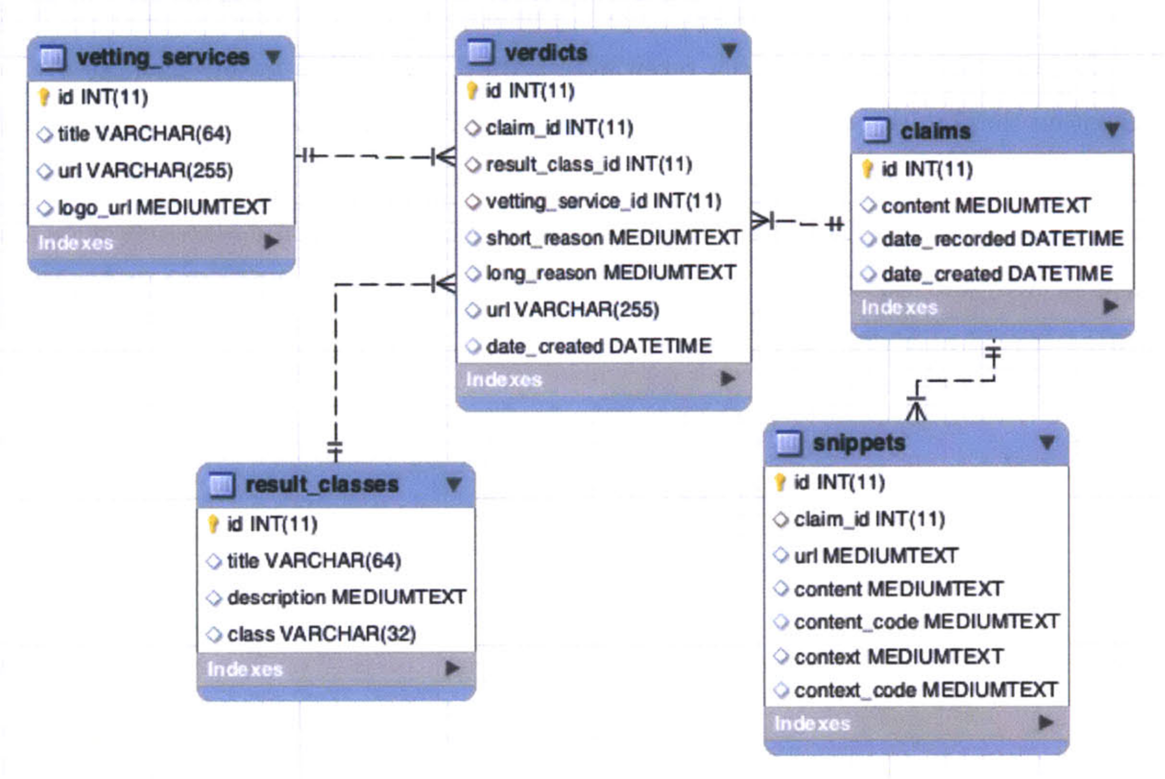
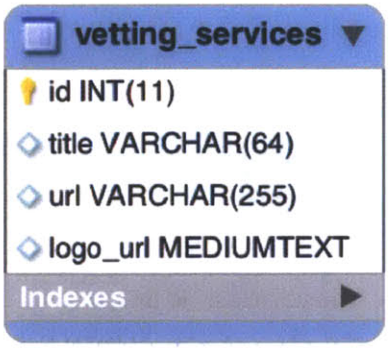


Figure 4-6: An Entity Relationship diagram of the Truth Goggles database.

Vetting Service Vetting services represent sources of "truth" in the system. A vetting service might be an organization (e.g. PolitiFact), a website (e.g. Wikipedia) or an algorithm (e.g. Twitter mining). This information makes it possible for interfaces to provide additional information and credit about a source.



vetting_services

id INT(11)

title VARCHAR(64)

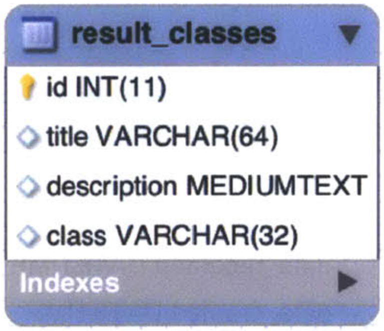
url VARCHAR(255)

logo_url MEDIUMTEXT

Figure 4-7: The vetting service model.

The vetting service model

Result Class Result classes represent the possible values of truth that a verdict can contain, for instance "true" or "mostly true." Each result class has a description explaining the meaning of that class, as well as a "class" field which contains a valid CSS class name that could be used by the client if desired. This list will grow as additional sources are added.



result_classes

id INT(11)

title VARCHAR(64)

description MEDIUMTEXT

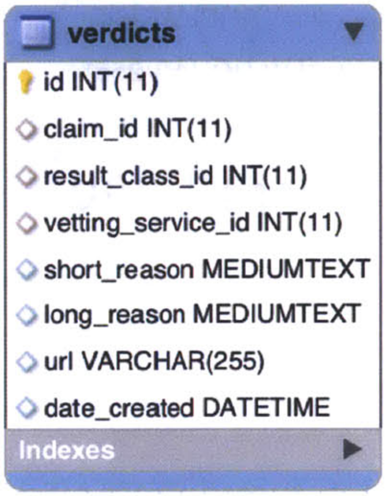
class VARCHAR(32)

Figure 4-8: The result class model.

The result class model

By including a specific "class" field it is possible to create groups of result classes with the same "class." For instance if Politifact rates something "Pants on Fire" the developer could decide to programmatically treat that in the same way as the Washington Post Fact Checker's "Four Pinocchios" [37] while still recognizing the different rating language used.

Verdict Verdicts represent applications of a result class to a fact-checked claim. They contain a short summary, a long summary, and a link to a full and original verdict source.



verdicts

id INT(11)

claim_id INT(11)

result_class_id INT(11)

vetting_service_id INT(11)

short_reason MEDIUMTEXT

long_reason MEDIUMTEXT

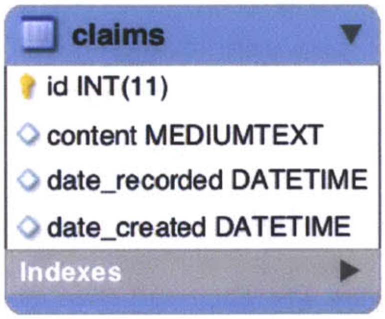
url VARCHAR(255)

date_created DATETIME

Indexes

Figure 4-9: The verdict model.

Claim A claim is a piece of content that has been fact checked. Claims and verdicts are separated because it is possible that a given claim has been fact checked by more than one source.



claims

id INT(11)

content MEDIUMTEXT

date_recorded DATETIME

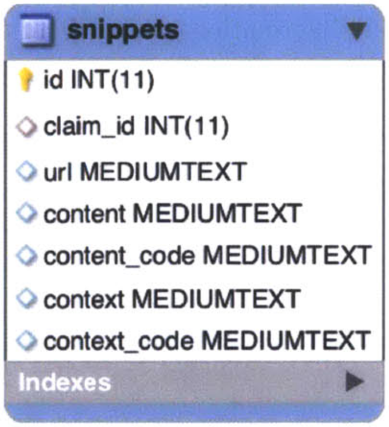
date_created DATETIME

Indexes

Figure 4-10: The claim model.

Snippet Snippets represent potential instances of fact checked phrases. These are paraphrases that were either flagged by a user or by an automated system.

The snippet model



snippets

id INT(11)

claim_id INT(11)

url MEDIUMTEXT

content MEDIUMTEXT

content_code MEDIUMTEXT

context MEDIUMTEXT

context_code MEDIUMTEXT

Figure 4-11: The snippet model.

4.4.2 Credibility API

The Truth Goggles credibility API is a partial implementation of a more substantial proposal for a complete credibility API. Since the Truth Goggles tool is only focused on one portion of the entire fact checking process it does not implement the API calls that have to do with creating and collecting new fact checks. Below is the proposed credibility API, including the proposed create and update methods. The unimplemented methods are marked as "*(proposed)*" to clarify their status.

Content Access (GET)

- **get_claims** takes in a string of text and returns a list of all claims that might appear in that text as determined by the fuzzy matching API. It also provides more specific information about where those claims exist within the content (i.e. character positions). The threshold for what is considered to a potential match is determined by the fuzzy matching API, so the accuracy of the results will depend on the specific implementation.
- **get_snippets** takes in a string of text and returns a list of all snippets that exist in that text as semi-perfect matches (i.e. matches discounting case and punctuation). It

also provides more specific information about where those snippets appear. What is considered to be a verified snippet will depend on the specific implementation of the credibility backend. It could be based purely on algorithms or could incorporate some form of human input. For Truth Goggles a snippet requires human confirmation to be considered verified.

- **get_verdicts** takes in a claim ID and returns a list of all verdicts surrounding that claim.
- **get_result_classes** takes no parameters and returns a list of all result classes represented in the system.
- **get_vetting_services** takes no parameters and returns a list of all vetting services represented in the system.

Content Creation (POST) It is suggested that the API implementation includes some form of authentication protocol such as OAuth [57] before allowing content creation. This would make it possible to verify contributions, track reputation of contributors, and otherwise prevent abuse of the API.

- **add_claim***(proposed)* takes in a claim object and inserts it into the database. This might be used to allow requests for fact checks, as a claim could be inserted without any verdicts.
- **add_snippet***(proposed)* takes in a snippet object and inserts a new snippet into the database. Depending on the implementation this snippet may be considered verified or unverified by default.
- **add_verdict***(proposed)* takes in a verdict object and inserts it into the database. Depending on the implementation this verdict may be considered credible, or it may need to undergo additional scrutiny.

- **add_result_class** *(proposed)* takes in a result class object and inserts it into the database. Obviously result classes are only relevant if they actually get used by a verdict.

Content Updates (POST)

- **update_claim** *(proposed)* takes in a claim ID and a new claim object and updates the claim.
- **update_snippet** *(proposed)* takes in a snippet ID and a new snippet object and updates the snippet.
- **update_verdict** *(proposed)* takes in a verdict ID and a new verdict object and updates the verdict.
- **update_result_class** *(proposed)* takes in a result class ID and a new snippet object and updates the verdict.

Content Moderation (POST) These methods all allow the crowd to weigh in on the relevance or appropriateness of a specific piece of content within the system. Depending on the implementation, the system might then adapt to reflect this feedback.

- **flag_claim** *(proposed)* takes in a claim ID and a flag type ("reject", "approve") and registers the information.
- **flag_snippet** *(proposed)* takes in a claim ID and a flag type ("reject", "approve") and registers the information.
- **flag_verdict** *(proposed)* takes in a claim ID and a flag type ("reject", "approve") and registers the information.

4.4.3 Scrapers vs API

The Truth Goggles credibility backend does not support the ability to add claims to the database over the API. Instead it uses scraper scripts which were custom built to interface with Politifact's developer API. The reason for this decision is that the focus of Truth Goggles is in the exploration of the client experience, which does not require more than one source. The challenges surrounding tracking the credibility of unknown sources represent a significant research question in itself.

It will eventually be necessary to modify Truth Goggles to allow for external submissions of new fact checks. As described in the API overview, this would require additional logic to ensure that only credible content made it to the Truth Goggles interface. This could be accomplished through a combination of moderation, crowds, reputations, and bias detection algorithms.

4.4.4 Additional Applications

This backend and credibility API can power more than just a browser based consumer interface like Truth Goggles. This API processes arbitrary text. That text could come from a web page just as easily as it could come from a processed photograph of a newspaper, a transcript of a live debate, a text field in the comments section of a blog, or the word processor of a journalist. With such a flexible system there are possibilities for many different types of credibility layers across a diverse set of media.

One such application is being explored by PolitiFact. They intend to create a mobile application with the capacity to highlight fact checks surrounding political advertisements at the moment that those advertisements are being played on television. PolitiFact would be able to process the transcript using this credibility API and identify potentially relevant prior fact checks for specific moments in the video.

Another often mentioned application of the credibility API is a content producer's "fact check" tool. This would result in a credibility equivalent to spell and grammar check. As the author types, their words would be sent to the credibility API to check for any fact checked claims that might have been referenced. The author would then have an opportunity to correct or clarify as needed well before that information was pushed out to the world.

4.5 Matching Engine

The matching engine was lightly explored and partially implemented, but ultimately did not yield strong results. Given the focus on credibility layers and interfaces for critical thought, the relatively ineffective matching engine did not hinder the research process. The user study and prototype was able to be run using hand-entered paraphrases, and the interface never incorporates results from any matching engine even in cases where results exist. This subsection provides a description of the approaches explored, the lessons learned from failed experiments, and some hopeful directions for future work.

4.5.1 Iteration 1: Tokenization

The first attempt at a fuzzy matching algorithm was written entirely in PHP as part of the credibility backend. It became immediately apparent that this portion of the system was complex enough to warrant being formally separated and iterated on. The first iteration was important despite the fact that it was ineffective because it inspired the concept of a semiautomated system.

Under this process all claims and snippets entered into the system are tokenized into lower case strings separated by whitespace and stripped of punctuation. The tokens are compared to a list of stop words in order to remove common English words, yielding a final set of meaningful tokens for each snippet. When new text is submitted for processing it is also tokenized by the same rules. A subset of snippets is filtered from the full pool using an inexpensive heuristic: the system identifies all snippets where at least 50% of their tokens appear in the new text's token pool. For instance the snippet "phrase context detection" would pass the filter if the new text contained at least two of the three tokens from the set ["phrase", "context", "detection"].

Once the pool of snippets is decreased the system begins the more expensive operation of walking through the new content for each snippet and comparing the snippet's tokens with ordered subsets of the new text's tokens. The size of these subsets is based on the size of the original snippet and contains a set of words. For instance if the set size was five the set list would contain the sets t[0:4], t[1:5], t[2:6], etc. where t is the set of all tokens. The number of stopword and non-stopword matches are then calculated for each window and the values are associated with the first word in the window. This process results in a "heat map" of potential references to a claim.

The goal of this algorithm was not to identify snippet phrases but was to power an interface that would identify areas of the page that are more likely to contain a fact checked phrase. In this sense it succeeded, as more potentially relevant sections of the text had higher numbers of matches. Unfortunately this algorithm yielded numerous false positives. For instance, any article about healthcare spending would flag areas of the page with potentially relevant instances of several fact checked phrases concerning Obama's healthcare reform bill, regardless of the actual relevance.

4.5.2 Iteration 2: Luminioso

Luminoso is a Python project that originated at the MIT Media Lab [54] and has continued on as an open source project. The tool provides a semantic analysis engine which is able to simulate linguistic understanding by learning from a massive corpus of word associations called the Common Sense Reasoning Database. The second iteration of the matching engine attempts to use Luminoso to identify instances of fact checked claims. Unlike the first iteration of the matching engine it has not been fully integrated into the Truth Goggles workflow, as results were never actually bubbled up to a client side script.

The Luminoso-powered matching engine is written in Python and regularly scrapes the Truth Goggles credibility API to get the latest set of claims and all content surrounding those claims (e.g. snippets, verdicts, the articles describing those verdicts, and the articles containing known snippets). All of this content is broken into paragraphs and processed by Luminoso, resulting in an updated language model. At this point each claim, snippet, and paragraph exists in the matching engine as a separate document within Luminoso.

When provided new text the algorithm breaks the content into paragraphs and converts those paragraphs into a vectors using Luminoso's "vector_from_text" method. That vector is then compared with the previously seeded document space using Luminoso's "docs_similar_to_vector" method. This process results in a list of documents that are most similar to each paragraph. The hope was that these documents, which were each associated with a claim, would reveal potential claims within an entire article (or within a paragraph of that article).

This approach failed for several reasons and ultimately the matches were not particularly useful. The primary issue was that it assumed that similarities between documents would reflect similarity between phrases and sentences. Under this approach all content in an article containing a fact checked phrase would have been associated with that phrase despite the fact that much of it would have been completely unrelated. Even if the corpus had been made more accurate and included only the most relevant paragraphs, identifying similar paragraphs is not necessarily comparable to finding instances of paraphrases.

It became clear that "document similarity" was not the correct metaphor to use when working with Luminoso for the purposes of Truth Goggles. Instead the language model would need to be seeded with general content and that claims and snippets would need to be vectorized at a sentence level and then stored for future use. New content would then be chunked into *sentences* not paragraphs. Those sentences would be vectorized as well, and those vectors compared using a mathematical heuristic. At this point, as automated matching was not the first priority for this thesis, attention was turned back to the user interface. However, I expect to continue exploring this space in future, and so the next proposed iteration is described found below.

4.5.3 Next Steps: Luminioso, Entities, and Crowds

The first two iterations, although unsuccessful, help us better understand what might be possible in the quest for a paraphrase detection engine. In this section I attempt to leverage those lessons and offer a proposal for a fuzzy matching process. The proposed process utilizes a sentence-level invocation of Luminoso, more general NLP techniques, and the crowd to ultimately create a semi-automated set of steps capable of identifying new instances of fact-checked claims.

The Tools Below are the components that would be vital for a Truth Goggles paraphrase detection process. Each tool offers a new piece of information which can then be used to ultimately determine the presence of paraphrased fact checks in new text.

Anaphora Resolution Anaphora refers to situations where pronouns are used to reference nouns introduced earlier in a body of text. For instance a news article about John Doe might first introduce John by name, but would go on to refer to him through pronouns or professional titles. When considering a claim it is helpful to understand all of the nouns to which it is referring. NLTK has algorithms to assist in anaphora resolution.

Entity Extraction Entity Extraction is a process that determines the key nouns found in a piece of text. These might be names, locations, brands, organizations, or otherwise noteworthy segments of text. Knowing that two pieces of content are referring to the same entities provides a significant clue about the potential link between content. This is especially true in the space of political fact checking, where the origin of a claim is often a famous person who is likely to be referenced nearby.

Sentence Detection Separating a block of text into sentences is not impossible, but it is also non-trivial. One might assume that punctuation as a separator would be enough, but this immediately gets disrupted by things like acronyms and abbreviations. Fortunately this is a solved problem and NLTK provides methods that are able to separate content into sentences.

Luminoso The key contribution of Luminoso is the ability to take a corpus of text (e.g. 200 pre-scraped news articles), combine that corpus with known associations from the common sense reasoning database, and use that understanding to develop a mathematical space representing the language. Luminoso is then able to take new text as an input and return a vector representing that text and where it fits in the language space. It is possible to estimate similarity between statements by comparing their vectors.

Fact Checking Organizations Fact checking organizations have an incentive to identify instances of claims out in the wild. They are also more trustworthy than a random anonymous Internet user. These professionals might be able and willing to donate a few minutes a day to offer final confirmation on low volume / high credibility tasks such as making determinations about disputed associations if it would result in increased traffic to their articles.

Truth Goggles Users Users are not a personal army [8], but it is possible to design interfaces and experiences that are valuable to both the user and the system. In this way the Truth Goggles fuzzy matching process could get most of the way towards identifying matches through automated processes, but still rely on the end user to take the final step.

The Dispute Finder project already showed evidence that people are willing to take significant action in the name of spreading information they are passionate about. If the interface is simple enough, Truth Goggles can harness its users to confirm or deny the accuracy of algorithmically suggested phrase associations. The reverse is also true, as Truth Goggles users can suggest specific phrases that are worth fact checking and those phrases can be specially processed by the system or shown to fact checking organizations.

The Process None of these tools would be able to solve the problem of paraphrase detection independently. Each NLP technique addresses part of the puzzle, but even if they were combined the result would not be accurate or consistent enough for full automation. The human approach simply will not work without supplemental processes or the incorporation of impossibly addictive interfaces or substances. Users would quickly be overwhelmed by being asked to look for needles in haystacks. The solution is to combine the two approaches to create an asynchronous snippet detection process.

Corpus Processing In order to prime Luminoso and NLTK a corpus of domain content must first be collected, stored, and processed. In the case of Truth Goggles this would consist of 200 or more news articles. Once the scraping is complete those documents must then be processed by the NLP engines so that they can update their models. The result is a set of tools primed and ready to begin analyzing new content.

Claim Processing The Truth Goggles claim database is dynamic: it grows and updates as fact checking organizations create and edit their claims, and as new kinds of paraphrases are identified. When a new claim or claim instance is added to the system or an old claim is updated that content needs to be processed and converted to a form that will work with the paraphrase detection components. This means creating a list of the entities associated with the content and generating a Luminoso vector for the content, which will eventually be compared to new vectors.

Text Processing Processing text to detect potential matches is expensive and will not be possible to do synchronously. Instead, new text will need to be scraped and processed by the system. The scraping can be done traditionally through a tool like MediaCloud [55] which is already set up to generate a stream of content as it is published, or submitted automatically as users navigate the internet (i.e. human-driven scraping). The benefit of the former is that content would generally be processed before a Truth Goggles user visited a page while the benefit of the latter is that only content being viewed by users would be processed.

Regardless of origin, once a piece of content has been added to the processing queue it will be modified to resolve anaphors, key entities will be extracted, and it will be broken into sentences. The sentences will then be combined into sentence trigrams (in order to provide additional context) and those trigrams will be converted to vectors by Luminoso.

At this point we can begin comparing this evidence with the claim vectors and claim entities. The specific constraints of this algorithm would need to be explored, but it will generally involve the generation of an entity similarity score and a vector similarity score. Sentences whose similarity scores with claims meet a certain threshold would then be considered viable candidates for the final step: human verification.

Phrase Candidate Vetting Once the list of likely paraphrases is generated it is further processed to remove any content that had already been vetted. This list of possible matches can, from this point forward, be instantly and synchronously identified in new content through simple string searches. The Truth Goggles client would then need to be modified to handle a new type of content: the potential snippet. When potential snippets are identified on a page the user would be presented with the opportunity to "Help identify claims" at which point they would be asked to confirm or deny associations.

After enough confirmations the phrase would graduate into becoming a full snippet. After enough rejections, however, the phrase would be marked as a known mismatch and would no longer be presented to regular users.

Chapter 5

Evaluation

A user study was run to test the effectiveness of Truth Goggles as an instigator of critical thinking. The study design, results, and reflections are described in this section. Links to the specific questions and articles and claims used in the study can be found in the appendix.

5.1 Study Design

This study was performed in order to understand Truth Goggles's effectiveness as a catalyst for critical thinking.

5.1.1 Measures of Success

There are many ways that a tool like Truth Goggles could be considered successful. One might hope that users familiar with the functionality would prefer it to the non-augmented consumption experience. Another measure of success might reflect the number of claims that users explored when the tool was enabled or the quality of that exploration. These questions are interesting and worth exploring, but they all require different inputs from the user. The following is a list of dimensions potentially worth testing along with a description of the study features those tests might be require.

1) Did people use Truth Goggles? It is difficult to accurately measure the use of a tool when working with a "captive" audience. This means that in order to test this question fully we would have to expand beyond a closed experiment and see if users installed the tool and activated it in more casual situations. Unfortunately Truth Goggles does not contain enough facts to be generally relevant in a non-sandboxed environment. Within a sandbox, however, it might be possible to measure use regardless. Activation rates will not be an accurate measure of use because several treatments involve automatic activation of Truth Goggles, but in these cases it might be useful to track the amount of time spent viewing content in the Inspection Pane, as during those times there is no question that the user is "using" Truth Goggles.

2) Did people enjoy using Truth Goggles? By presenting post-use questionnaires about the user experience it will be possible to get a sense of how happy participants were with the user experience. It will also be possible to gauge happiness with the tool by giving users a choice to either activate or disable Truth Goggles at some point during the study, after the tool has been exposed to them. Finally, it will be helpful to keep track of how many participants would choose to install Truth Goggles onto their browser at the conclusion of the experiment.

3) Were users exposed to more fact checks In order to compare a change we must have a baseline and the ability to measure exposure. This could be accomplished with something as simple as a survey question asking users to self report the number of fact checks they tend to view when reading an article, or by tracking the number of times the user opts to explore a fact checked claim more deeply when given a choice (e.g. in highlight mode).

4) Did users engage with the fact checks? To understand levels of engagement the tool would need to keep track of what content was actually read and comprehended, as opposed to what content was simply rendered on a screen. One way to measure this could involve reading comprehension questions to see whether or not users noticed specific pieces of the fact check. Another might involve hiding an "easter egg" button within a specific fact check which users might see if they read the content carefully.

5) How well did Truth Goggles enable critical thinking? Although critical thinking does not require a change of opinion it seems reasonable to believe that a change of opinion does indicate some type of thought. By measuring the drift in beliefs about fact checked claims after using Truth Goggles it would at least be possible to capture the ability of Truth Goggles to trigger updated beliefs.

6) Did Truth Goggles affect levels of trust in consumption experiences? This question is deeply relevant, but given the format of this study it will be quite difficult to measure trust accurately. Asking for self-reported measurements in changes in the amount of trust the reader has may yield at least some insight towards this question.

Decisions The study design reflects aspects of each of these questions; however, questions 2, 4 and 5 are most important given the motivations of this thesis ("Did people enjoy using Truth Goggles," "Did users engage with fact checks," and "How well did Truth Goggles enable critical thinking"). Question 6 ("Did Truth Goggles affect levels of trust in consumption experiences") is also interesting, but represents a piece of information that will be difficult to collect beyond self reporting.

5.1.2 Preparation

Before the study began I selected, tagged, and pre-processed ten political articles. The articles were then placed into a digital sandbox. Articles did not necessarily share the same publication month, topic, source, or author. A list of the articles used in this study can be found in the Appendix.

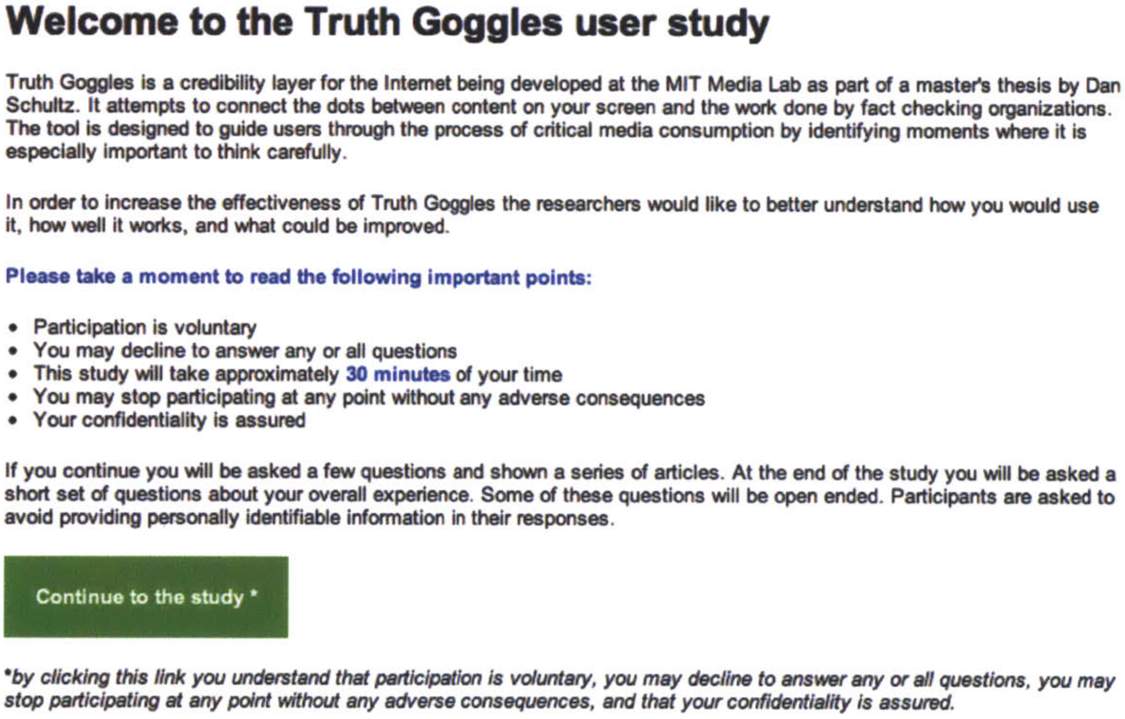
In addition to the content preparation there were also several tracking features which needed to be added to Truth Goggles. The most important feature was click tracking, which allowed every interaction within the study sandbox to be recorded with timestamps. There was also the addition of a feature to "lock into" different viewing modes. The standard Truth Goggles tool allows the reader to switch between interface styles at will based on personal preference. Such a feature would potentially dilute the study results.

A set of simple study scripts were created to guide users across the multiple articles and enable the properly randomized test conditions. The scripts were designed to randomize the presentation of the questions (when appropriate) and the ten articles. They also prevented users from being able to "go back" to view earlier portions of the study.

5.1.3 Process

Recruitment Users were recruited through digital communication services such as email and Twitter [30] and directed to a landing page. It is important to mention that this approach does not yield a truly random sampling and it is very likely that it recruited a disproportionate number of friends, individuals who were already familiar with the concept of Truth Goggles, and professionals who are already aware of the challenges surrounding media literacy. This is especially important because it means that the political biases of participants will not be evenly distributed.

Landing Page The landing page provided a brief summary of Truth Goggles and disclosed the survey participant's rights (See Appendix). The study did not begin until participants confirmed their understanding of this information.



Welcome to the Truth Goggles user study

Truth Goggles is a credibility layer for the Internet being developed at the MIT Media Lab as part of a master's thesis by Dan Schultz. It attempts to connect the dots between content on your screen and the work done by fact checking organizations. The tool is designed to guide users through the process of critical media consumption by identifying moments where it is especially important to think carefully.

In order to increase the effectiveness of Truth Goggles the researchers would like to better understand how you would use it, how well it works, and what could be improved.

Please take a moment to read the following important points:

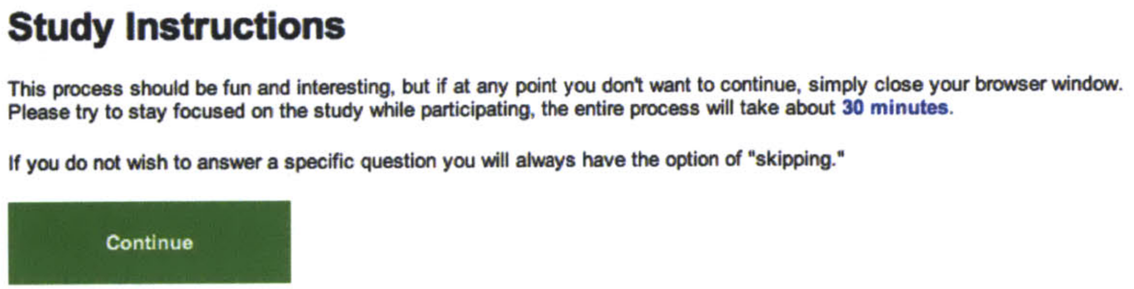
- Participation is voluntary
- You may decline to answer any or all questions
- This study will take approximately **30 minutes** of your time
- You may stop participating at any point without any adverse consequences
- Your confidentiality is assured

If you continue you will be asked a few questions and shown a series of articles. At the end of the study you will be asked a short set of questions about your overall experience. Some of these questions will be open ended. Participants are asked to avoid providing personally identifiable information in their responses.

**by clicking this link you understand that participation is voluntary, you may decline to answer any or all questions, you may stop participating at any point without any adverse consequences, and that your confidentiality is assured.*

Figure 5-1: The landing page.

Initial Instructions Before beginning the Prior Belief Survey users were reminded of the logistics of the study. In particular they were told how to end participation if they did not wish to complete the study and were asked to complete the study in one sitting.

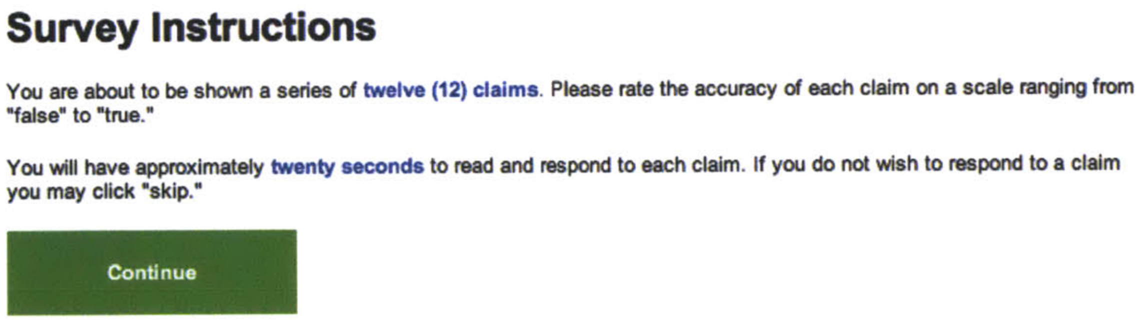

Study Instructions

This process should be fun and interesting, but if at any point you don't want to continue, simply close your browser window. Please try to stay focused on the study while participating, the entire process will take about 30 minutes.

If you do not wish to answer a specific question you will always have the option of "skipping."

Figure 5-2: The initial instructions.

Prior Belief Survey In order to gauge how effectively Truth Goggles helps users think critically and update their beliefs we must first understand what they believed before using the tool. To collect this information the study began with a brief survey. Participants were shown a series of statements and asked to rate the truthfulness of each on a five point scale ranging from "False" to "True." Because the goal was to get a "gut reaction" as opposed to a thoughtful analysis, the user was only given a few seconds to make each decision.


Survey Instructions

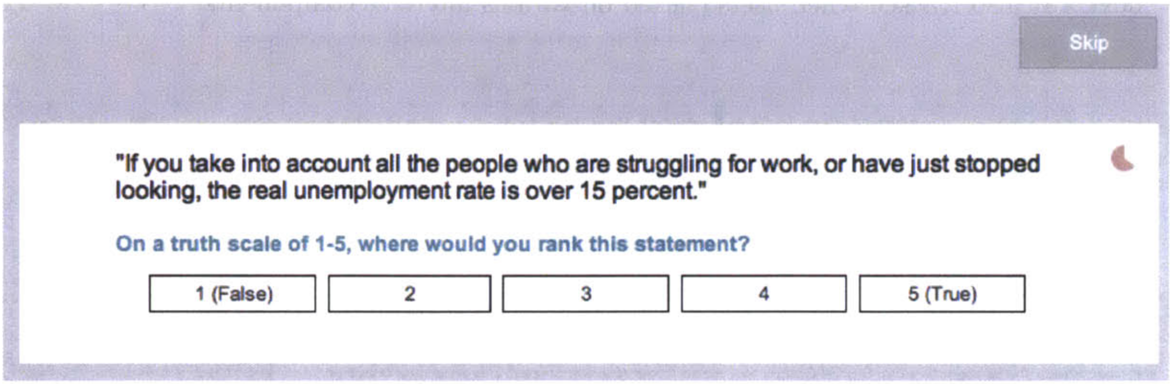
You are about to be shown a series of twelve (12) claims. Please rate the accuracy of each claim on a scale ranging from "false" to "true."

You will have approximately twenty seconds to read and respond to each claim. If you do not wish to respond to a claim you may click "skip."

Figure 5-3: The prior belief survey instructions.

Articles The bulk of this study involved presentation of actual news articles which had been pre-selected to contain known fact checked claims. Ten articles were utilized in this study and their presentation order was randomized. The order of the treatments was semi-randomized. After viewing a short set of instructions the user began the article section.

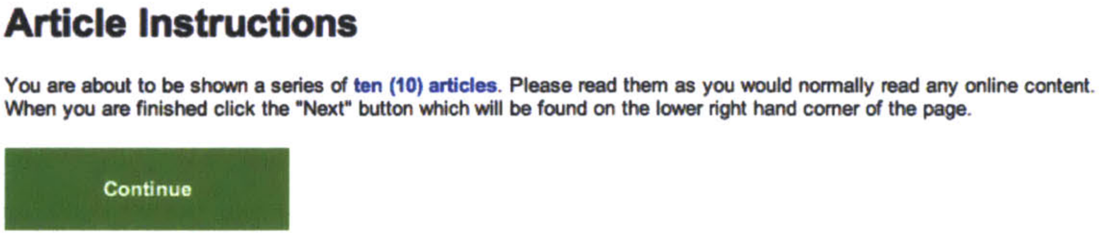
The first article pair was presented with no credibility layers. The second, third, and fourth



"If you take into account all the people who are struggling for work, or have just stopped looking, the real unemployment rate is over 15 percent."

On a truth scale of 1-5, where would you rank this statement?

Figure 5-4: The prior belief survey interface.

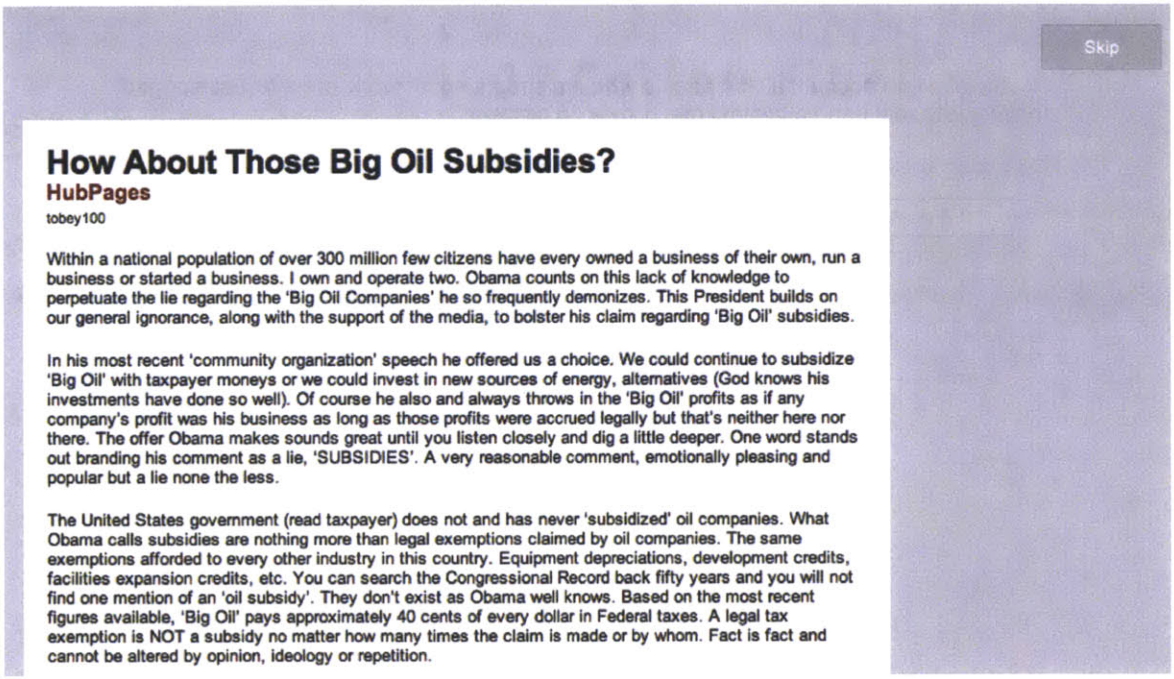


Article Instructions

You are about to be shown a series of **ten (10) articles**. Please read them as you would normally read any online content. When you are finished click the "Next" button which will be found on the lower right hand corner of the page.

Figure 5-5: The article instructions.

pairs were presented with one of the three Truth Goggles interfaces (each interface appearing once in random order). The fifth pair allowed the user to choose which experience he or she prefers. The system tracked how much time was spent on each article before continuing to "next." It also tracked when users clicked on various interface components.



How About Those Big Oil Subsidies?

HubPages

tobey100

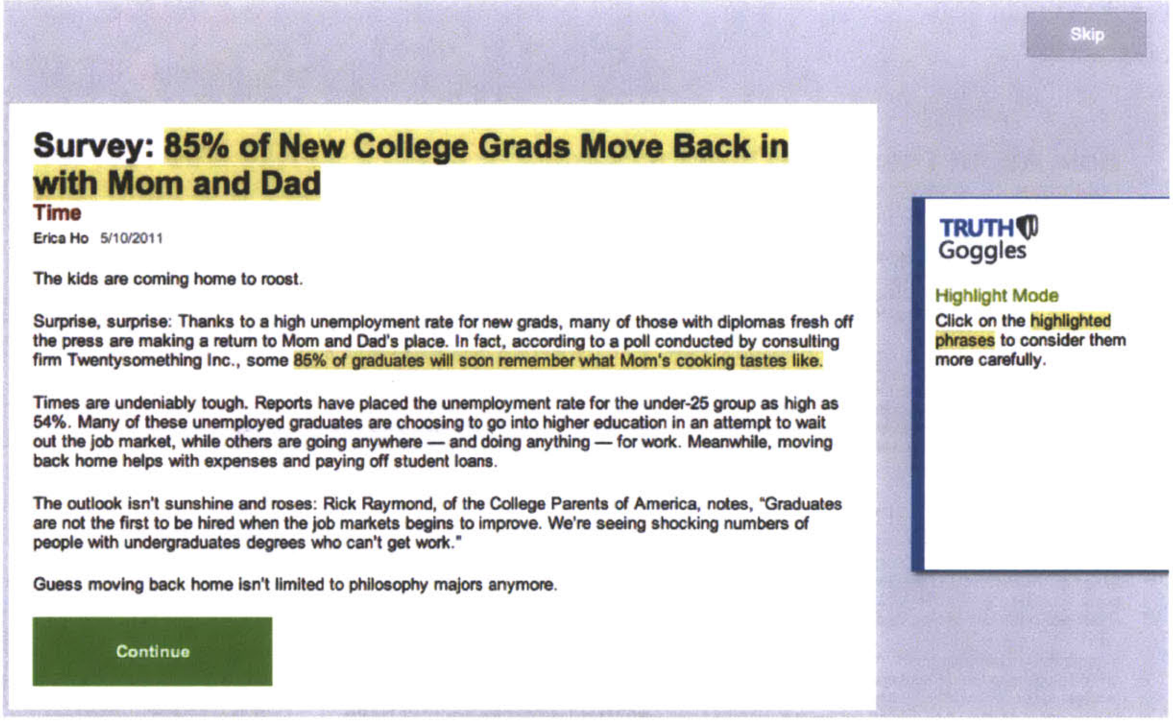
Within a national population of over 300 million few citizens have every owned a business of their own, run a business or started a business. I own and operate two. Obama counts on this lack of knowledge to perpetuate the lie regarding the 'Big Oil Companies' he so frequently demonizes. This President builds on our general ignorance, along with the support of the media, to bolster his claim regarding 'Big Oil' subsidies.

In his most recent 'community organization' speech he offered us a choice. We could continue to subsidize 'Big Oil' with taxpayer moneys or we could invest in new sources of energy, alternatives (God knows his investments have done so well). Of course he also and always throws in the 'Big Oil' profits as if any company's profit was his business as long as those profits were accrued legally but that's neither here nor there. The offer Obama makes sounds great until you listen closely and dig a little deeper. One word stands out branding his comment as a lie, 'SUBSIDIES'. A very reasonable comment, emotionally pleasing and popular but a lie none the less.

The United States government (read taxpayer) does not and has never 'subsidized' oil companies. What Obama calls subsidies are nothing more than legal exemptions claimed by oil companies. The same exemptions afforded to every other industry in this country. Equipment depreciations, development credits, facilities expansion credits, etc. You can search the Congressional Record back fifty years and you will not find one mention of an 'oil subsidy'. They don't exist as Obama well knows. Based on the most recent figures available, 'Big Oil' pays approximately 40 cents of every dollar in Federal taxes. A legal tax exemption is NOT a subsidy no matter how many times the claim is made or by whom. Fact is fact and cannot be altered by opinion, ideology or repetition.

Figure 5-6: The no layer interface.

Post-Article Survey The participant was presented with a short survey after viewing all articles. The survey asked the user to rate the truth value of the claims that could have been found in the articles in addition to two claims that did not appear in any articles as a control. This functioned in the exact same way as the prior belief survey, presenting the participants with the series of claims and asking them to rate the truth value of those claims according to what they believe. Once again they had a limited time to respond to each claim, as the purpose of this survey was to understand how the reading experience did or did not affect their final beliefs.



Survey: 85% of New College Grads Move Back in with Mom and Dad

Time

Erica Ho 5/10/2011

The kids are coming home to roost.

Surprise, surprise: Thanks to a high unemployment rate for new grads, many of those with diplomas fresh off the press are making a return to Mom and Dad's place. In fact, according to a poll conducted by consulting firm Twentysomething Inc., some 85% of graduates will soon remember what Mom's cooking tastes like.

Times are undeniably tough. Reports have placed the unemployment rate for the under-25 group as high as 54%. Many of these unemployed graduates are choosing to go into higher education in an attempt to wait out the job market, while others are going anywhere — and doing anything — for work. Meanwhile, moving back home helps with expenses and paying off student loans.

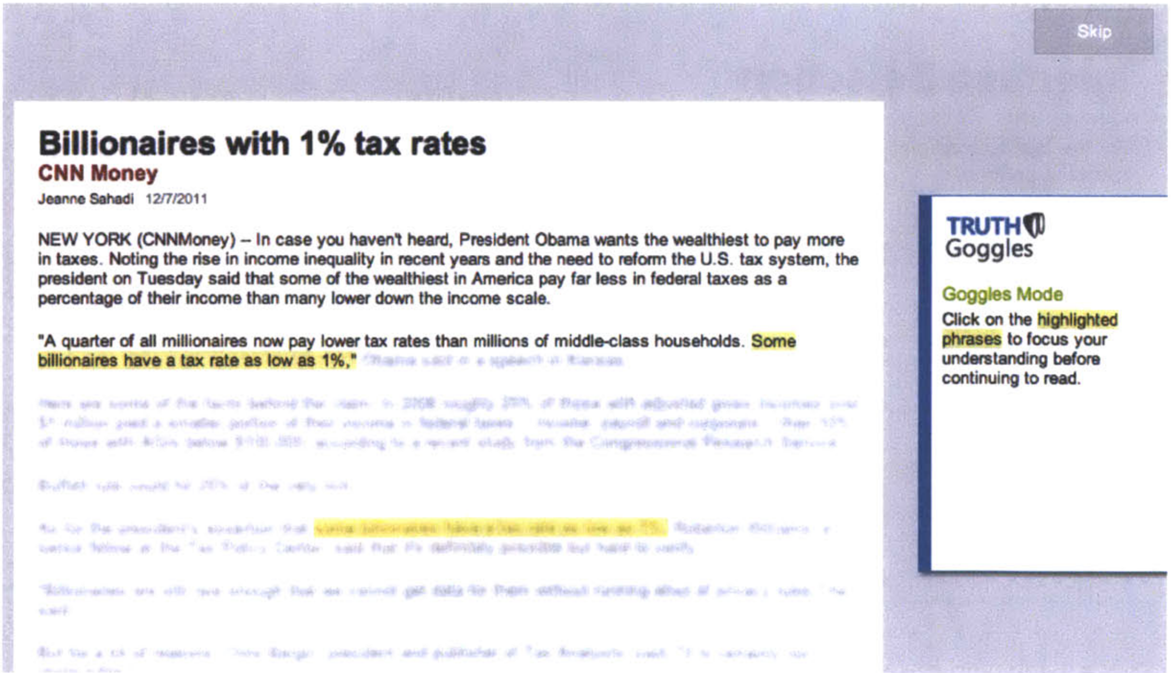
The outlook isn't sunshine and roses: Rick Raymond, of the College Parents of America, notes, "Graduates are not the first to be hired when the job markets begins to improve. We're seeing shocking numbers of people with undergraduates degrees who can't get work."

Guess moving back home isn't limited to philosophy majors anymore.

Highlight Mode

Click on the highlighted phrases to consider them more carefully.

Figure 5-7: The highlight mode interface.



Billionaires with 1% tax rates

CNN Money

Jeanne Sahadi 12/7/2011

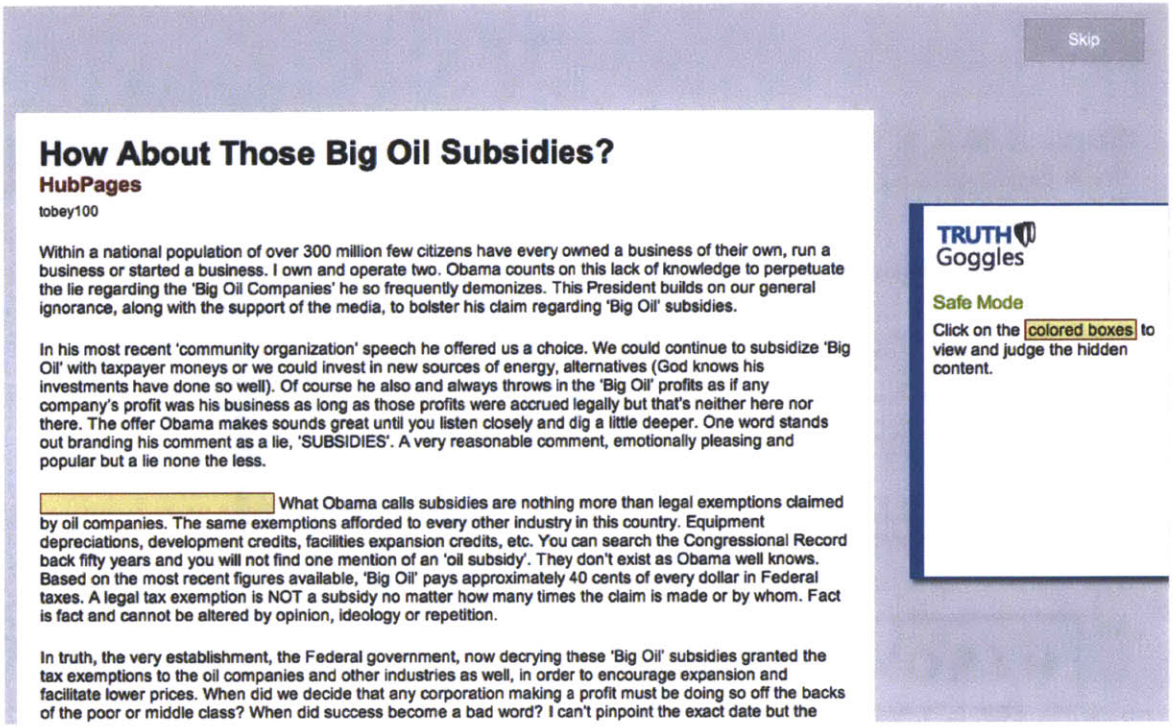
NEW YORK (CNNMoney) -- In case you haven't heard, President Obama wants the wealthiest to pay more in taxes. Noting the rise in income inequality in recent years and the need to reform the U.S. tax system, the president on Tuesday said that some of the wealthiest in America pay far less in federal taxes as a percentage of their income than many lower down the income scale.

"A quarter of all millionaires now pay lower tax rates than millions of middle-class households. Some billionaires have a tax rate as low as 1%," [illegible]

Goggles Mode

Click on the highlighted phrases to focus your understanding before continuing to read.

Figure 5-8: The goggles mode interface.



How About Those Big Oil Subsidies?

HubPages

tobey100

Within a national population of over 300 million few citizens have every owned a business of their own, run a business or started a business. I own and operate two. Obama counts on this lack of knowledge to perpetuate the lie regarding the 'Big Oil Companies' he so frequently demonizes. This President builds on our general ignorance, along with the support of the media, to bolster his claim regarding 'Big Oil' subsidies.

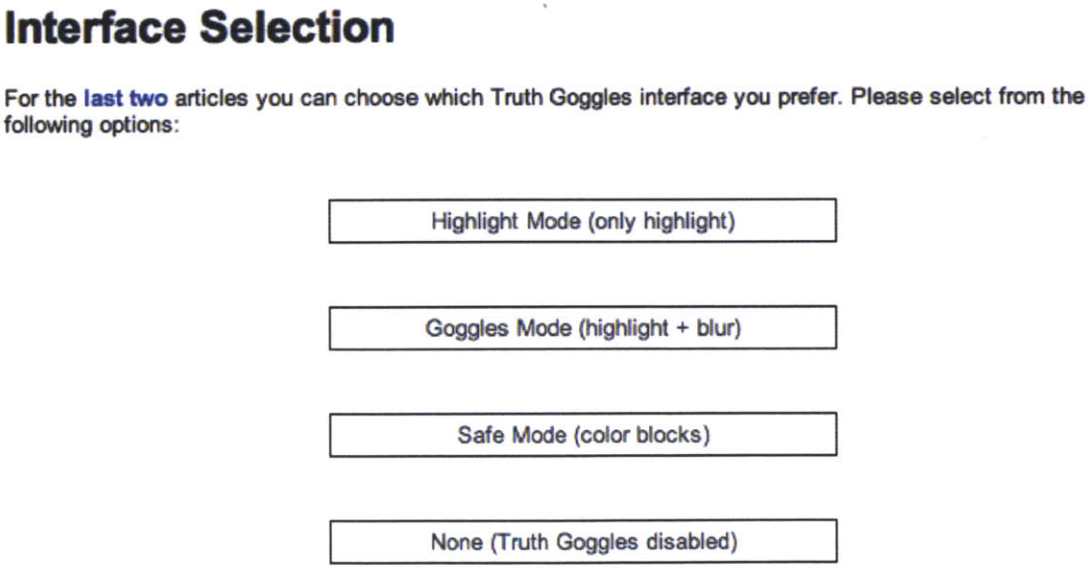
In his most recent 'community organization' speech he offered us a choice. We could continue to subsidize 'Big Oil' with taxpayer moneys or we could invest in new sources of energy, alternatives (God knows his investments have done so well). Of course he also and always throws in the 'Big Oil' profits as if any company's profit was his business as long as those profits were accrued legally but that's neither here nor there. The offer Obama makes sounds great until you listen closely and dig a little deeper. One word stands out branding his comment as a lie, 'SUBSIDIES'. A very reasonable comment, emotionally pleasing and popular but a lie none the less.

What Obama calls subsidies are nothing more than legal exemptions claimed by oil companies. The same exemptions afforded to every other industry in this country. Equipment depreciations, development credits, facilities expansion credits, etc. You can search the Congressional Record back fifty years and you will not find one mention of an 'oil subsidy'. They don't exist as Obama well knows. Based on the most recent figures available, 'Big Oil' pays approximately 40 cents of every dollar in Federal taxes. A legal tax exemption is NOT a subsidy no matter how many times the claim is made or by whom. Fact is fact and cannot be altered by opinion, ideology or repetition.

In truth, the very establishment, the Federal government, now decrying these 'Big Oil' subsidies granted the tax exemptions to the oil companies and other industries as well, in order to encourage expansion and facilitate lower prices. When did we decide that any corporation making a profit must be doing so off the backs of the poor or middle class? When did success become a bad word? I can't pinpoint the exact date but the

Click on the colored boxes to view and judge the hidden content.

Figure 5-9: The safe mode interface.

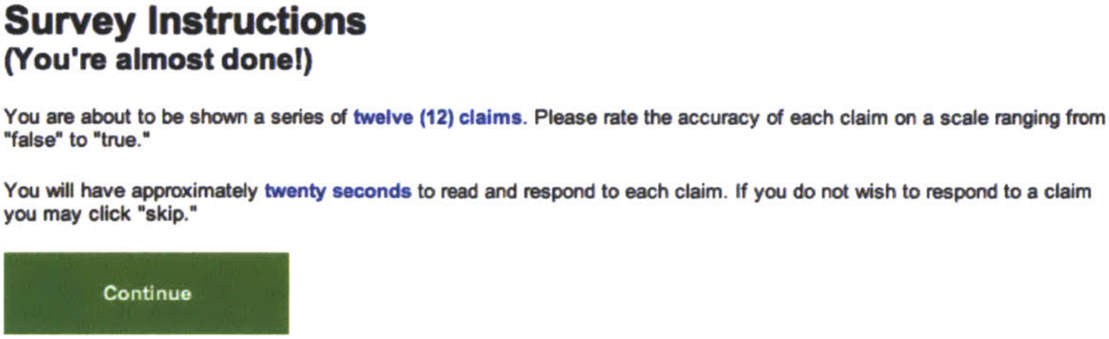


Interface Selection

For the last two articles you can choose which Truth Goggles interface you prefer. Please select from the following options:

Highlight Mode (only highlight)
Goggles Mode (highlight + blur)
Safe Mode (color blocks)
None (Truth Goggles disabled)

Figure 5-10: The optional activation interface.

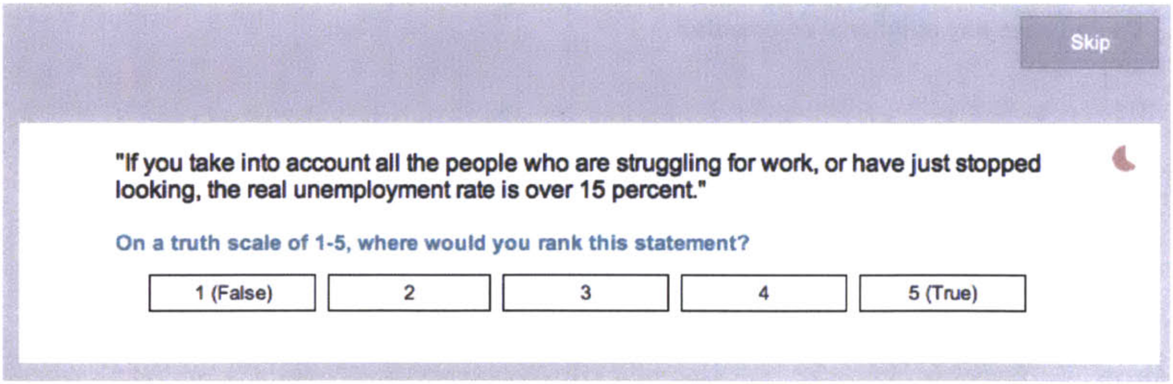


Survey Instructions
(You're almost done!)

You are about to be shown a series of **twelve (12) claims.** Please rate the accuracy of each claim on a scale ranging from "false" to "true."

You will have approximately **twenty seconds** to read and respond to each claim. If you do not wish to respond to a claim you may click "skip."

Figure 5-11: The post-article survey instructions.

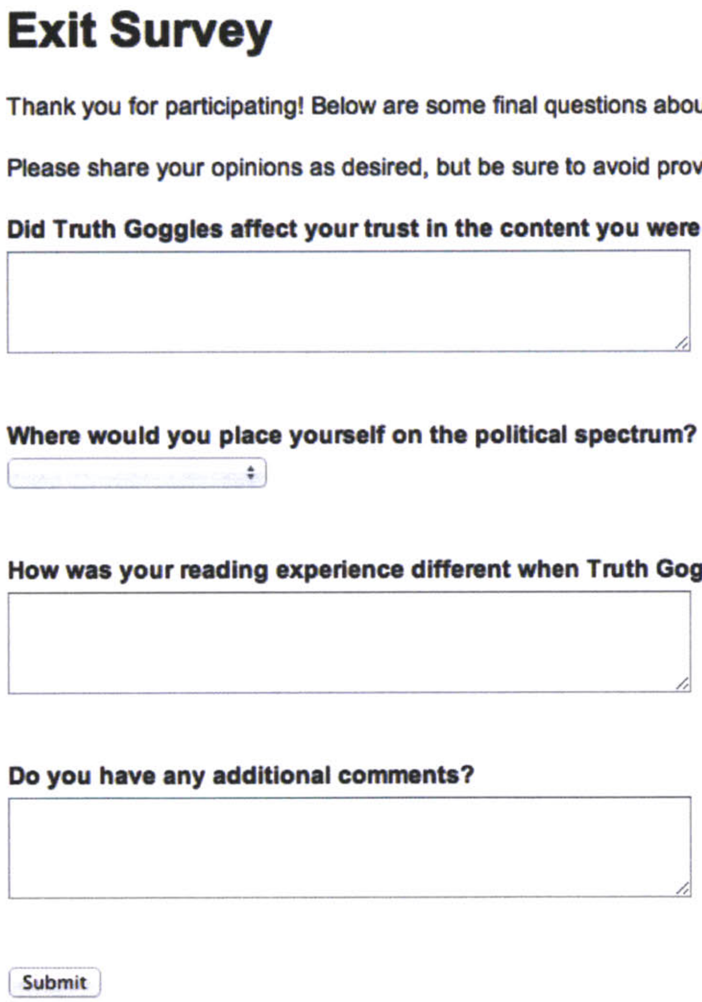


"If you take into account all the people who are struggling for work, or have just stopped looking, the real unemployment rate is over 15 percent."

On a truth scale of 1-5, where would you rank this statement?

Figure 5-12: The post article survey interface.

Exit Survey Once all of the articles had been reviewed and the post-article survey was complete the participant was thanked for his or her time and asked to provide feedback on the experience. These questions focused on opinions of the Truth Goggles interface and overall experience.



Exit Survey

Thank you for participating! Below are some final questions about your experience.

Please share your opinions as desired, but be sure to avoid providing personally identifiable information.

Did Truth Goggles affect your trust in the content you were reading? Please explain.

Where would you place yourself on the political spectrum?

How was your reading experience different when Truth Goggles was enabled?

Do you have any additional comments?

Figure 5-13: The exit survey interface.

The "Final Survey" interface.

5.2 Results

The study was conducted online over the course of five days. This section describes the data, the analysis of that data, and the lessons learned from the results. Although Truth Goggles does not generally assume that its sources necessarily contain "truth," in the name of data evaluation some ground truth needed to be considered. Lacking a better metric, the source verdicts (i.e. PolitiFact's ratings) were used as grounding for "accuracy" for this analysis.

5.2.1 Participants

There were a total of 219 participants, 88 of whom completed the entire process[1]. The participant pool consisted of both U.S. and non-U.S. citizens. This analysis only considers results from participants that completed the entire study. The analysis omits two samples due to comments that cite technical difficulty and, in one case, inebriation.

Reported Ideology	Count	Percent
Unspecified	11	13%
Strong Conservative	1	1%
Moderate Conservative	3	3%
Independent	8	9%
Moderate Liberal	42	48%
Strong Liberal	23	26%

Table 5.1: The ideological breakdown of study participants.

Ideological Breakdown The ideological breakdown of full participants in Table 5.1 shows that the vast majority of the study participants who completed the process were strong and moderate liberals. This fact was expected due to the recruitment methods, but nevertheless significantly limits the potential impact of this study.

[1] There were actually 430 participants with a total of 200 completions, but only the first 219 participants were considered in the data pool due to potential bias resulting from third party press about the study. The comments from the full 430 were considered for the feedback section.

Completion Rates The sub-50% completion rate warrants explanation, but there are several reasons why the number might be so low.

- **1: Lack of interest in the tool**. It is possible that participants decided, after reaching the point of the study where they got to see one of the Truth Goggles interfaces, that they were no longer interested in continuing.
- **2: Lack of interest in the study content**. Asking people to read ten full articles that they did not select themselves is challenging. It is possible that some participants simply did not care about the content. Furthermore, many of the participants were not United States citizens and all of the articles and fact checked claims were based around US politics.
- **3: Lack of time**. Participants who stopped in the study may have simply decided they did not have the time or available attention to participate in the study. Some participants may have returned and completed the survey, but this was not tracked.
- **4: Lack of interest in the study**. Some participants exited the study after the treatment section was complete. These people finished using Truth Goggles and stopped just short of completion, which indicates they may have simply gotten bored of the process on the whole.

Progresss	Count	Percent
Pre Goggles	94	43%
Mid Goggles	32	15%
Post Goggles	5	2%
Complete	88	40%

Table 5.2: The points at which participants ended their progress.

In Table 5.2 “Pre Goggles” means that participants exited before ever being shown a Truth Goggles interface. This may have been during the prior belief survey or during the “No Interface” article treatment. “Mid Goggles” indicates participants exited before completing the article phase, but after they had seen the Truth Goggles tool. “Post Goggles” refers to participants exited during the post article survey.

The breakdown reveals that the vast majority of the non-completions were due to an early exit. This is promising, as it implies that participants exited either due to time or a realization that the study was focused on United States political claims.

5.2.2 Variables

There are several types of data collected in this study. The responses to the prior belief and post article surveys are collected on an interval scale, while things like survey responses, activity, and interface preferences are not on any type of scale.

The data analysis for each treatment category is separated into three parts: pre-treatment, post-treatment, and drift. There is also some more straightforward data such as direct survey responses and final treatment preference selections. The pre- and post-treatment analysis is built around aggregate responses to the prior-belief and post-article surveys respectively. The drift section reflects changes in responses on a per-participant level.

Inaccuracy Measures One of the most important aspects of the study is understanding how well the aggregate participants were able to specify truth values for fact-checked claims. The first measure, **average inaccuracy saturation**, is based on the number of "truth units" that a participant's stated beliefs differed from Politifact's, on a truth scale from 1 to 5. For instance a vote of "false" (1) would contribute two inaccuracy units for a claim that PolitiFact rated "half true" (3) and four inaccuracy units for a claim that PolitiFact rated "true" (5) . The inaccuracy saturation is calculated by dividing the number of inaccuracy units by the maximum number of possible inaccuracy units.

This takes into account the fact that a claim with a real verdict of "true" has the potential to contribute four inaccuracy units (with "false" being the furthest from accurate), while a claim with a real verdict of "half true" only has the potential to contribute two.

The second measure, **average inaccuracy distance**, is simply the average number of inaccuracy units across all participants.

The inaccuracy saturation and average inaccuracy variables both reflect how accurate the entire population was overall. Both measures are defined in terms of the potential and the number of participants. Inaccuracy distance will necessarily be smaller for claims whose verdicts fall in the middle of the truth spectrum, as those claims are restricted to smaller maximum inaccuracy distances (a maximum of two units on either side instead of four units in a single direction).

Over, Under, and Exact The over, under, and exact scores reflect how participants voted on claims in relation to PolitiFact's verdicts.

- **Over** shows the number of people who believed a claim to be more true than it actually was.
- **Under** shows the number who believed a claim to be more false than it actually was.
- **Exact** shows the number of people who reached the same verdict as PolitiFact.

For example if the over rating were 40, the under rating were 10, and the exact rating were 30 this would mean that across all votes on all claims appearing in a given treatment, 40 votes were over-trusting, 10 votes were over-skeptical, and 30 votes were exactly right.

The over and under scores come with corresponding **saturation** and **bias saturation** ratings. The saturation rating takes into account the fact that some claims are already at their maximum or minimum values and therefore could not possibly be over or under. For example a claim with a real truth rating of "false" (1) could never yield a response that falls under it. Thus the saturation measure is more appropriate for drawing conclusions.

The bias saturation measure indicates, within the over or under category, how much of the possible distances in that direction were "filled" by the responses. There is also a measure called **total bias saturation** which reflects, within the participants who voted over or under, how much of the possible distances were "filled" by the responses. A high level of bias saturation implies increased polarization.

Intended, Backfire, and Neutral Drift Participants were asked to rate the same claims twice during the study. The first time was during the prior belief survey, and the second was after the treatment, in the post article survey. This means that it is possible to track changes in response for each pairing.

There are several ways a participant's response to a fact checked claim can drift in relation to PolitiFact's verdicts. For the most part it can move closer to accuracy, it can move further from accuracy, or it could have no change at all. The drift metrics help capture the nature of response changes by comparing each participant's prior belief survey responses with the post article survey responses.

- **Perfect Drift** counts the number of responses which drifted between surveys from being over or under to being exact.
- **Intended Drift** counts the number of responses which yielded a smaller number of inaccuracy units between the surveys.
- **Backfire Drift** counts the number of responses which yielded a larger number of inaccuracy units between the surveys.
- **Neutral Drift** counts the number of responses which did not change between the surveys.
- **Neutral Drift (Exact)** counts the number of responses which were exact and did not change between the surveys.

For some pairs it simply is not possible to drift in a given direction (i.e. if the prior belief response was "true" or "false"). This is accounted for by the "Intended Drift Saturation" and "Backfire Drift Saturation" metrics. These numbers are calculated by considering only responses that could have possibly yielded the relevant type of drift.

For instance if the total number of intended drifts was 15 out of 30 total pairs but of those 30 pairs 10 were already exact matches in the prior belief survey (and therefore could not possibly have yielded a good drift), then the Intended Drift Saturation would be 15 out of 20, or .75.

Positive and Negative Overcompensation It is also possible for a respondent to overcompensate in a way that moves towards correction but goes too far and becomes equally or more incorrect but from the opposite perspective. For example if a participant were to say that a half true claim (3) had a truth rating of mostly false (2) in the prior belief survey but then change their answer to mostly true (4) or true (5) in the post article survey this would be considered overcompensation.

If the new value is closer to the accurate value it is still considered an "Intended Drift" even if the respondent "switched sides." For instance if a participant were to say that a half true claim (3) had a truth rating of false (1) and drifted to mostly true (4), this would be an "Intended Drift." If the respondent had drifted all the way to true (5) however, this would have been an overcompensation.

Overcompensation is considered positive[2] if the drift is towards "true" (e.g. the prior belief rating was lower than the post article survey rating). Similarly, it is considered negative if the drift is towards "false" (e.g. the prior belief rating was higher than the post article survey rating). There were so few instances of overcompensation that deeper analysis was not performed in the data analysis section.

5.2.3 Data

All tests were built around a participant sample size of 85, but the number of responses within each section fluctuated because participants were able to skip questions as they pleased. Individual skip responses (or timeouts) were removed from the data pool. In addition, one claim was removed from the study due to an unrealized disparity between the verdict provided by PolitiFact's API and the verdict published on the PolitiFact website. Several participants noticed the difference and it is unclear which verdict to use when attempting to consider accuracy.

The data cited in the data analysis section can be found in the Appendix.

[2] By positive and negative I am referring to numeric value, not sentiment.

5.2.4 Analysis

Treatment Equivalence Due to an error in the study software, all participants were exposed to the same two claims for the no-information control treatment. This means that the no-information control treatment is nonequivalent and cannot be compared with the other treatments. The metrics are still calculated and shown here because they do reflect no aggregate change, as expected, between the two surveys.

All other treatments (no-layer, highlight, goggles, and safe) used the same article and claim pool, which were randomly sorted for both the belief tests and treatments. The only design difference between these treatments is that the no-layer mode was always presented first, while the layer treatments were always randomly ordered. The treatments showed no significant differences in average ground truth rating, with the average ground truths tending to fall between "mostly false"(2) and "half true" (3).

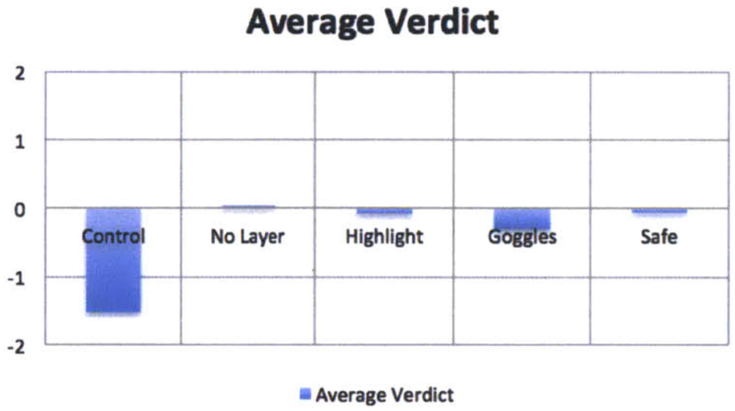


Figure 5-14: The average core verdicts across treatments.

The effects of this mistake is made incredibly clear by viewing the average verdicts across treatments, as shown in figure 5-14. The control group had an average verdict between -1 (mostly false) and -2 (false), while the other four treatments had averages closer to the neutral 0 (half true)

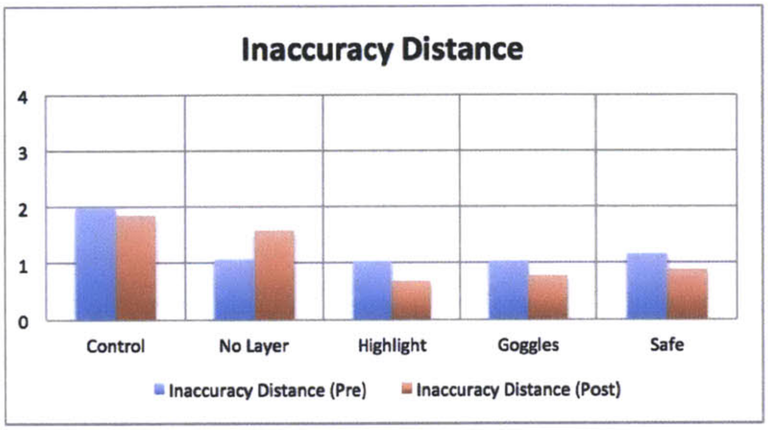


Figure 5-15: The average inaccuracy distance before and after treatments.

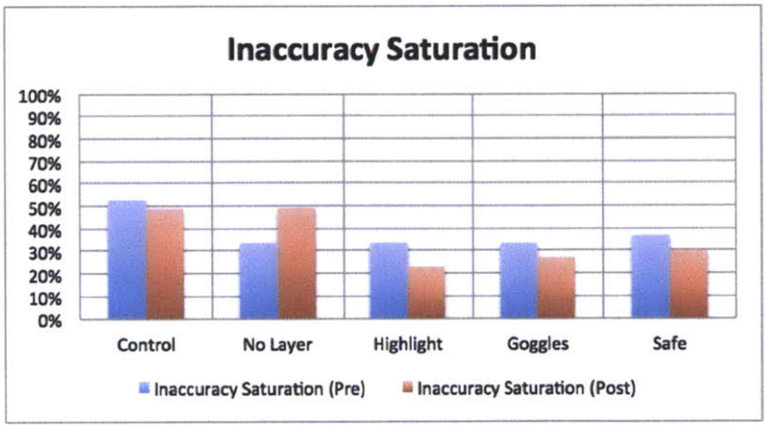


Figure 5-16: The average saturation of inaccurate responses before and after treatments.

Accuracy Figures 5-15 and 5-16 reflect the two inaccuracy measures tested. Inaccuracy distance shows average number of inaccuracy units that responses fell from a given claim. Inaccuracy saturation shows, of the total inaccuracy possible, how much inaccuracy occurred across the population.

As expected, there was no statistically significant change in accuracy for the no-information treatment. This makes sense, as participants were given no evidence with which they could have updated their beliefs. Very interestingly, participants became *less* accurate both in terms of saturation ($p < .01$) and distance ($p < .01$) during the no-layer treatment. This indicates that simply reading the articles referencing fact checked claims caused participants with no guidance to develop less accurate beliefs about those claims.

The highlight, goggles, and safe treatments resulted in an increase in accuracy both in terms of saturation (highlight $p < .01$, goggles $p < .05$, safe $p <.05$) and distance (highlight $p < .01$, goggles $p < .01$, safe $p < .01$). This indicates that all three variations of the Truth Goggles interface were effective at helping consumers reach more accurate conclusions about fact checked claims.

There was no statistically significant difference in accuracy saturation shift or accuracy distance shift between the three Truth Goggles treatments. This indicates that there is no reason to believe that one is more effective than another in terms of promoting accuracy. Compared to the no-layer treatment, all three layers had a difference in accuracy saturation shift (highlight $p < .01$, goggles $p < .01$, safe $p < .01$) and accuracy distance shift (highlight $p < .01$, goggles $p < .01$, safe $p < .01$). This indicates that all three modes were more effective than the no-layer treatment in helping participants form accurate beliefs.

Over, Under, and Exact Once again, as expected, there was no statistically significant change in over, under, or exact saturations for the no-information treatment. The no-layer treatment showed a decrease in exact saturation ($p < .01$) and an increase in over saturation ($p < .01$). There was not a statistically significant change in under saturation for the no-layer treatment. This indicates that participants who were not already overly skeptical of a

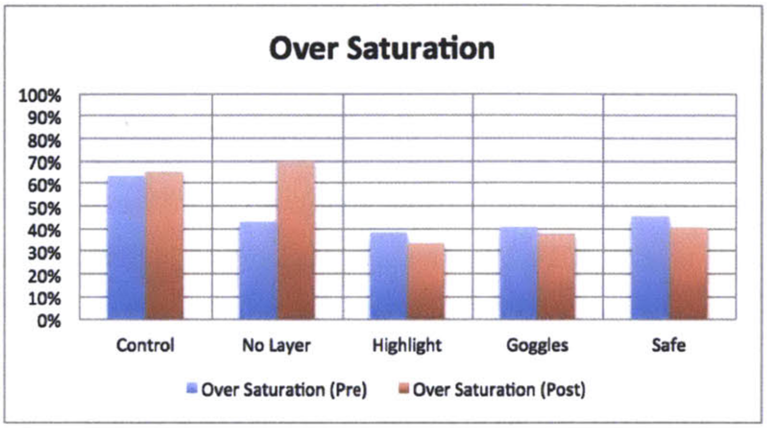


Figure 5-17: The percentage of overly trusting responses out of the possible pool before and after treatments.

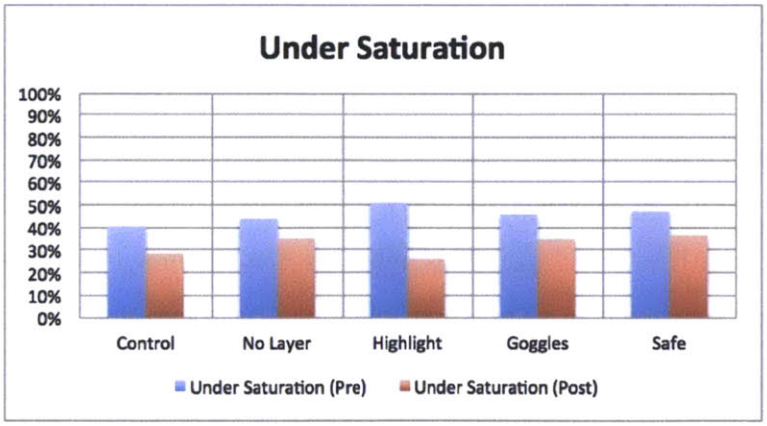


Figure 5-18: The percentage of overly skeptical responses of the possible pool before and after treatments.

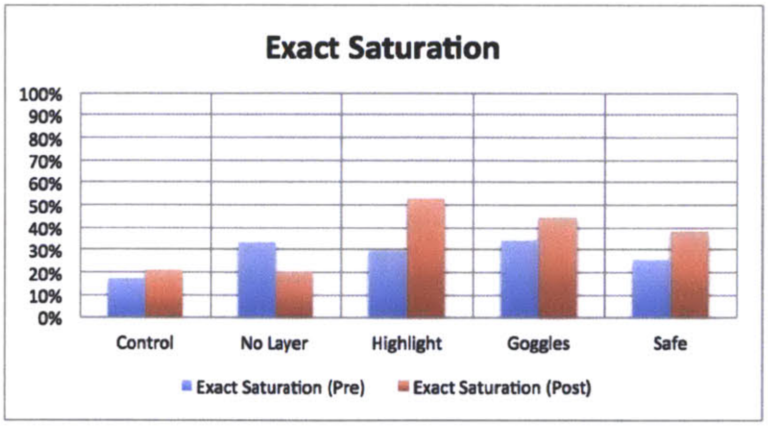


Figure 5-19: The percentage of exactly correct responses before and after treatments.

claim would tend to overly trust that claim when exposed to the corresponding article with no credibility layer.

All three Truth Goggles treatments yielded an increase in exact saturation (highlight p < .01, goggles p < .05, safe p < .01), a decrease in under saturation (highlight p < .01, goggles p < .05, safe p < .01), and no statistically significant change in over saturation. This implies that participants who were not already overly trusting of a claim would tend to update their beliefs in a way that resulted in more accuracy when using a credibility layer.

The three Truth Goggles treatments showed no statistically significant change in total bias saturation, under-bias saturation, or over-bias saturation. This implies that participants did not tend to shift their beliefs towards extremes (e.g. "true" or "false" as opposed to "mostly true" or "mostly false") when consuming content with a credibility layer. Without the credibility layer there was an increase in overall bias saturation (p < .01) and over-bias saturation (p < .01), and no statistically significant change in under-bias saturation.

Drift Figure 5-22 shows the average participant ratings across all claims before and after treatments juxtaposed with the average verdicts of those claims. It should be noted that these averages do not represent accuracy, but rather represents overall drift, as a rating of

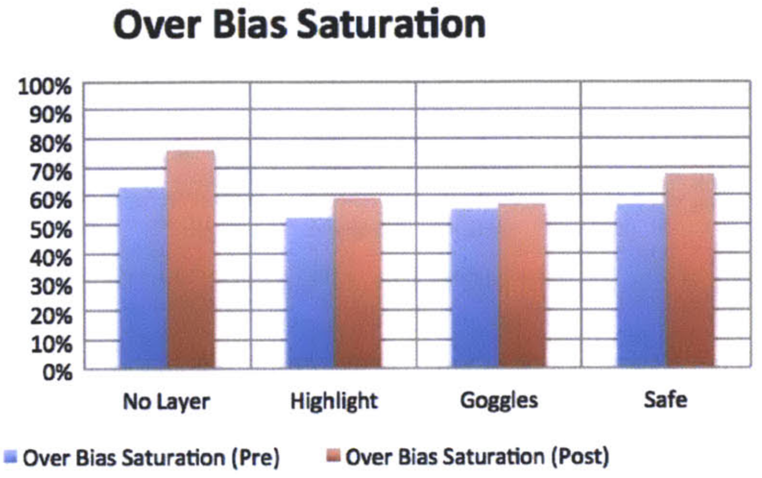


Figure 5-20: The bias saturation among responses that were over trusting before and after treatments.

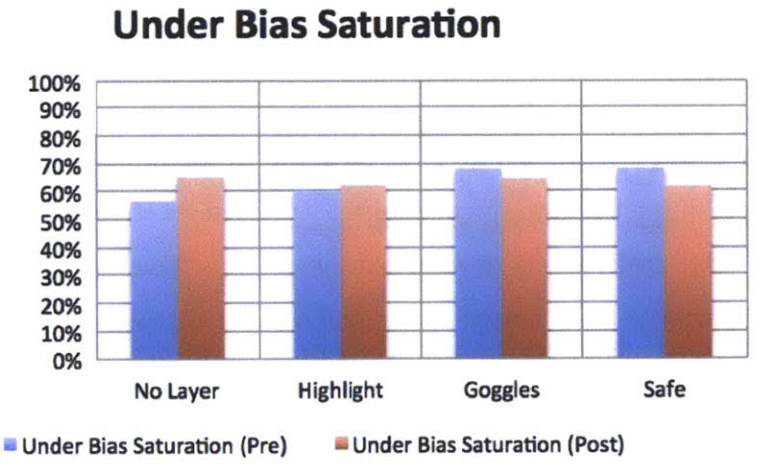


Figure 5-21: The bias saturation among responses that were over trusting before and after treatments.

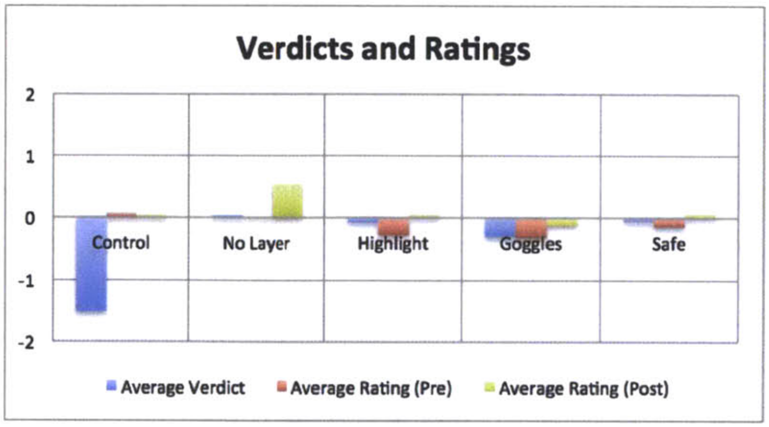


Figure 5-22: The average actual verdicts compared to average participant ratings before and after treatment.

2 (mostly false) and 4 (mostly true) would have the same average as a rating of 1 (false) and 5 (true) contributing different numbers of inaccuracy units for a given verdict.

As expected there was no statistically significant direction in drift in the no-information treatment. There was, however, a non-zero average absolute drift (p < .01) in the no-information treatment. This means that responses tended to shift (albeit a minor amount) despite no additional information, but the direction of that shift was not consistent.

All other treatments also have a non-zero absolute drift (p < .01). Their drifts were larger than the no-information drift (no-layer p < .01, highlight p < .01, goggles p < .01, safe p < .01), although they were all statistically equivalent to one another.

Both the no-layer treatment and the highlight treatment showed a positive drift (no-layer p < .01, highlight p < .01), while both the safe and goggles treatment showed no significant directional trends. This implies that for the no-layer and highlight treatments the participant population generally became more trusting, while for the goggles and safe treatments the participant population generally remained just as trusting or skeptical as they were before treatment.

This unique disparity between highlight treatment and the other layer treatments could be a reflection of the fact that highlight mode allows participants to ignore fact checks of claims they already believe to be true. This would have made it less likely for over-trusting participants to correct in a way that would yield a negative drift. Meanwhile safe mode and goggles mode do not give participants the ability to choose to explore a claim based on its content (i.e. over-trusting and over-skeptical participants are likely to be challenged more equally), which could explain the more neutral drift.

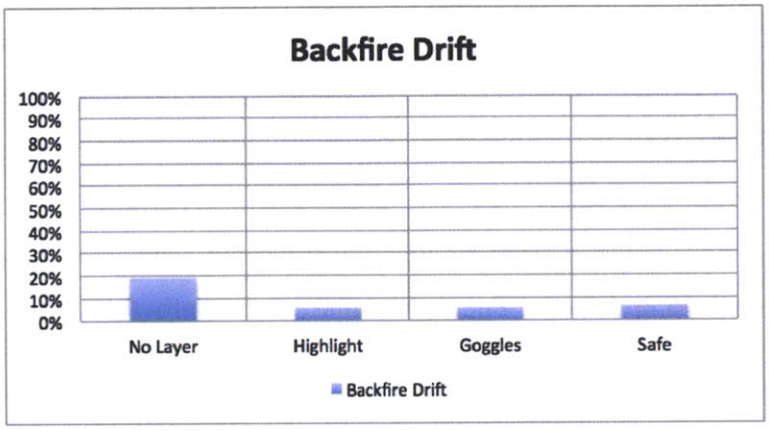


Figure 5-23: The percentage of backfire drift among responses that could have backfired before and after treatments.

Backfire Drift The credibility layers yielded a decrease in proportion of backfire drift when compared to no-layer treatment (highlight p < .01, goggles p < .01, safe p <.01). There was no significant difference in backfire drift across the three layer treatments. The rate of backfire drift among layer treatments generally ranged from around 2% to around 9% (p < .05). These numbers show significant improvement over backfire measurements in past studies [49], but because this experiment was not designed to explicitly measure backfire effects the comparison is irrelevant. For the purposes of this study it is enough to note that there was a lower backfire drift in the layer treatments when compared to the no-layer treatment.

5.2.5 Reflections

Selected Interface	Count	Percent
Highlight	62	71%
Goggles	16	18%
Safe	8	9%
None	2	2%

Table 5.3: The breakdown of interface preferences for the final treatment.

Did People Enjoy Using Truth Goggles? After being exposed to all three interfaces, participants were allowed to choose a final treatment from the four information treatment options. As shown in Table 5.3, when given this choice only two (2) out of the 88 participants who completed the survey chose to view their final two articles without using some variation of Truth Goggles. The vast majority of participants (62) selected "highlight mode." This evidence alone should not be used to conclude preferences, as it was never explicitly stated to the participants that they were being asked to select their favorite mode during this portion of the survey.

Did Users Engage with the Fact Checks? Several treatments essentially forced the user to enter the inspection pane (the modal window that appears when a claim has been clicked) for fact checked claims in order to read the articles presented to them. This means it would not be appropriate to make assertions based on the number of clicks. Furthermore, because this study was conducted online and anonymously it would not be appropriate to consider the amount of time spent viewing the inspection pane, as participants may have temporarily paused their participation due to external activities.

Thus the only measurement we have to understand engagement with fact checked claims is the number of times the "more" button appearing within the inspection pane was clicked. Throughout the study and across all treatments the claim inspection pane was rendered 664 times. Of those renderings the "more" link was clicked 176 times, or approximately 25% of the time. This means that each participant deeply engaged with, on average, two fact checked claims during their participation in the study.

Considering the fact that users were never prompted to click the "more" link and had enough information with just the information pane to reach conclusions without viewing the full story, 25% feels like a successful exploration rate. There is no way to tell how much more or less the levels of engagement was different from normal given the study design.

Another indication that users engaged with fact checks is that several participants noticed that for one claim the rating provided in the Truth Goggles information pane was different from the rating provided on the PolitiFact website. The rating had apparently been recently corrected and the correction had not made it into the Truth Goggles database in time for the study. This resulted in the removal of that claim's data from the analysis, but it shows just how carefully many participants were consuming information.

Did Truth Goggles Enable Critical Thinking? This was the primary research question of the study and the answer appears to be a resounding yes. Participants became more accurate, less polarized, and generally more informed in their beliefs when using Truth Goggles when compared to reading information without a credibility layer. Surprisingly there did not appear to be a significant difference in invoked critical thinking between the three interface designs, although the drift disparity between highlight mode and the other two modes indicated that participants may have been less likely to correct over-trusting beliefs when using highlight mode over goggles or safe mode.

Comments were generally strongly supportive of Truth Goggles, although some participants commented that Truth Goggles actually hindered their ability to process the article by prompting a processing of the fact checks instead.

Did Truth Goggles Affect Levels of Trust? Of the 88 participants asked, 33 explicitly said "yes" (that Truth Goggles did affect their trust in the article content) and 8 participants explicitly said "no" (that Truth Goggles did not affect their trust in the article content). Many participants noted that they were far more skeptical of all claims in articles and that they would often decide to stop reading articles based on false claims.

Unfortunately, because of the way the question was phrased, participants did not comment on their trust in the information experience on the whole. Several participants did suggest that they would feel much more comfortable with the experience if Truth Goggles pulled from multiple fact checking services, noting that they did not necessarily trust that PolitiFact was flawless or unbiased.

5.3 User Feedback

The user study generated over 400 comments, which were submitted through the exit survey by participants who had gone through the entire study process. In this section I attempt to describe some of the themes that were spread across the feedback.

5.3.1 Use Cases and Experiences

Almost every participant used the comments as an opportunity to describe their experience with Truth Goggles and how it changed the way they read content.

Determining Article Credibility

> Being able to quickly cross-reference the "sound-bytes" allows a quick evaluation of the article - and therefore the author's diligence."

> If a fact was marked as 'false' or 'barely true' I immediately became skeptical of the article on the whole."

Many participants saw the Truth Goggles interface as a way to judge the accuracy of an entire article based on its reliance on a key fact checked claim. For instance one commenter noted "when the main crux of the article was only partly true I didn't trust the remainder of the article, assuming it was also only half true at best." For some, this attitude was then used to determine the value in continuing to read at all: "it made me more likely to ignore an article that hinged on false information." For others this approach was used to determine the appropriate level of trust, which they used to fine-tune their reading process as opposed to rejecting the content outright.

Conveniently Accessing Context

> I am already in the habit of checking this stuff, but the goggles make it easier.“

> Truth Goggles actually empowered me to be able to check these claims for myself.“

Many users saw Truth Goggles as an efficiency tool, allowing them to more effectively take exploratory actions that they already take (or at the very least aspire to take). They saw Truth Goggles as a more convenient way to “refute numbers that sounded like bullshit without having to dig through a bunch of google searches.” For others it was simply nice to have additional context within a few clicks.

Skimming Fact Checks

> I skimmed right to the disputed fact to discover if it was true or not. I did not read the whole article, which is what I did with the first, unhighlighted, articles.“

Some readers used Truth Goggles highlighting as a metric to guide their habit of skimming articles, although it is quite possible that this was because they were using them in a study context. Many users felt that the nature of the highlight made skimming feel more natural. Some found themselves inadvertently becoming more interested in the fact check than the article itself.

Increasing Confidence

> It made me feel more secure that I could read the pieces —especially the op-ed ones —without as much risk of being misled as before.“

Truth Goggles helped some users feel more confident that they could approach information that otherwise would have been avoided due to concern about misinformation. "I would probably read more online articles if I had a way of checking their validity as I was reading. Checking later is onerous, at the least, and frankly I don't really know how to go about it."

Increasing Critical Ability

> This was a good reminder to be skeptical of sources I trust."

As one might have guessed based on the title of this thesis, increasing critical ability is the primary goal of Truth Goggles, making this particular line of comments particularly important. Several participants noted that Truth Goggles reminded them to stay skeptical of claims that were highlighted but more importantly some users found themselves "wondering about whether other statements were true, trying to look for claims that sounded iffy."

Some participants even found themselves "scrutinizing the PolitiFact pieces," which is a pretty good indicator that they were being inspired to think carefully. In his public analysis of the user study, Andrew Phelps said that "After using the goggles for awhile, it was impossible to read articles without a skepticism bordering on incredulity" [53].

Increasing Open Mindedness

> I thought less about trying to detect the political motivations of the author as a gauge of truthfulness, and more about the fact vs fiction aspect, which also got me into a sort of meta-thinking about the real issues."

An unintended but equally thrilling effect was the occasional increase in open mindedness. Some users indicated that the ability to determine accuracy through an independent party's analysis occasionally served as a gateway into self reflection. For instance, one participant reported that they were able to consider opposing viewpoints when they discovered that claim used as evidence by their ideological opponent, which they expect to be a lie, was true according to PolitiFact.

5.3.2 Concerns

The comments, while generally very positive and supportive, also gave participants an opportunity to voice their complaints and concerns about the tool. Although many of these issues are already known, they are well worth repeating.

Skepticism of the Sources

> I felt wary of Truth Goggles, not knowing the inclinations of its author(s)"

One concern, which was actually a little too uncommon considering its relevance, was in the credibility of Truth Goggles and the diversity of its sources. Some of the comments boiled down to a distrust in PolitiFact specifically, some cited a more general distrust in everything, and others simply felt that the tool needed more than a single source to be considered trustworthy.

These concerns indicated that the Truth Goggles interface could be improved to make it more clear that the sources behind Truth Goggles are not above scrutiny. In many ways it is just as dangerous to blindly accept a verdict as it would be to blindly accept a statistic in a news article. In general the comments calling for more than one source are absolutely correct: there needs to be more credible data and more sources powering Truth Goggles before it can be considered a trustworthy tool.

Distracting

> I tend to skim articles anyway, so highlight mode made it hard for me to read anything except"

The most common complaints had to do with distraction and flow interruption caused by the three interface modes. People were generally less fond of the blur and safe modes because they did not appreciate being forced to interact with content and wanted the freedom to ignore the Truth Goggles claim. One comment declared that the participant "didn't like the blur. I'm anti blur."[3] Beyond complaints about the ways that blur and safe mode forced certain types of interactions, there were clearly fundamental aspects of the credibility layer which were a very distracting point for many people.

There are ways to identify claims without a bright highlight, and indeed a less attention-grabbing approach may be more appropriate from a theoretical perspective, as it would not generate extra attention and increased familiarity towards fact-checked claims. For example, a subtle icon placed before or after a claim or even a less vibrant highlight color might make it easier for a person to use Truth Goggles without having their entire experience dominated by fact checks. It is a fine line, however, as getting too subtle might make it easy for consumers to absorb fact checked information without knowing there was contextual information.

The forced context switching, caused by the modal information pane, was also raised as a concern. This is more difficult to address because that context switch was by design, but it is possible to find ways to dampen the negative effects. For example it might be possible to expose users to the fact checks before they begin to read.

Focus on Fact Checks

> I scanned straight to the highlighted bit. Intentionally or not, that works as a beacon, calling attention to disputed facts."

A large number of people reported that they went straight for the fact checks and in many cases ignored the article content completely. This could easily be due to the study environment, since participants had not selected these articles and presumably they were more curious about the Truth Goggles interface than they were about month old political articles. This reenforces the previously mentioned concern with highlighting.

[3] Given the researcher's respect for anonymity it is important to note how overwhelmingly unlikely it is that this participant was actually a character from Seinfeld.

Excuse for Laziness

> Similar claims within the same articles were not challenged, and I didn't know whether that meant they had been vetted."

Some users reported that Truth Goggles made them "think less" because they assumed that all dubious claims had been identified for them. Although this was not a common trend it is absolutely a dangerous one, and steps need to be taken to ensure that users understand that Truth Goggles is not, and never will be, a replacement for their brain.

5.3.3 Ideas and Requests

People were excited and hopeful that it would soon become a fully deployed product. Aside from requests like "make sure it will be free"[4], there were requests for two major features specifically: a highlight mode that used colors to reflect truth value, and the ability to submit requests for additional fact checks.

> Maybe go directly to highlighting colors for truthfulness, as opposed to clicking on the highlight to get a popup"

While this idea sounds useful at first, there are many reasons why a blatant color coding in terms of truth would be dangerous even in a system that is technologically perfect for the many reasons described earlier in this thesis. That said, use of color and saturation to convey information is low hanging fruit, and there might be a way to use color or color intensity to reflect the likely importance of a fact checked claim since color should not be used to reflect the value.

> There were many other 'facts' that I would have liked to see verified or disputed."

[4] It will be free.

The ability to highlight claims and request fact checks is clearly one that would be trivial to implement and incredibly useful for both consumers and fact checkers. Organizations like PolitiFact have already expressed an interest in being able to find fact checked claims. This is reflected in the API design, which proposes the ability to submit requests for new fact checks.

5.3.4 Hall of Fame

There were some very amusing comments as well. Some declared love for turtles [38], others declared glory to the hypnotoad [26], and one even warned the researcher that the participant was "vaguely drunk" and "didn't read the articles." It seems appropriate to conclude this analysis with what is undoubtably my favorite comment of all.

> Truth Goggles helped me to think more critically about the references to data in the news articles. Without the goggles, I would normally read a statistic and probably ignore it, since I don't really have access to a way of checking it, although sometimes you read a statistic and believe it because it seems like its probably true. I like that Truth Goggles makes you consider every claim, so that you don't believe something that could be very inaccurate. I often find myself reading articles and thinking about how this is a story all about how my life got flipped turned upside down, and I'd like to take a minute just sitting right there and tell you how I became the prince of a town called Bel-Air." [1]

Chapter 6

Conclusion

I have explained many of the considerations that should be on the mind of developers and designers hoping to create more credible information experiences. I have also introduced and tested Truth Goggles, my first attempt to apply those considerations, and collected hundreds of comments that show people are receptive to and excited by the idea that technology has the capacity to help promote information grounded in fact. This warm reception is backed by study results, which show that Truth Goggles was able to significantly improve the way users process information and reach conclusions.

I also offer lessons learned and several suggested next steps. The implementation of a complete credibility API will make it possible to collect additional fact checks and expand the scope of Truth Goggles. The continued exploration of fuzzy matching algorithms will increase the relevance and impact of Truth Goggles by identifying more instances of fact checked claims. Finally, improvements to the credibility layer itself, in response to the criticisms identified during the study, will ultimately provide a better and potentially more effective user experience.

When I started this project there were major concerns that the biases, effects, and basic shortcomings of human nature would make technologically triggered critical thought a virtually impossible task. In this thesis I have shown that effective credibility layers, able to guide users through the process of critical thought, are absolutely achievable. The promising results from the user study combined with the support and passion of a population eager for credible information give serious reason for hope.

Without trust it is easy to dismiss content as manipulative, biased, or otherwise ignorable. Comments from the user study also revealed that credibility layers, by facilitating trust, have the potential to help people become more open to messages that challenge their world views. It would seem that the reintroduction of credibility is essentially the reintroduction of respect, and it allows users to consider new perspectives.

Journalists continue to grapple with the constantly evolving implications of the Internet. The industry is built around its ability to generate credible information. If it can leverage that ability and create universally trustworthy consumption experiences it would not only make their product more attractive: it could change the nature of political discourse in the 21st century. Through Truth Goggles I have shown that these experiences are possible, what they might look like, and how ready people are to finally start trusting again.

Appendix A

Appendix

A.1 COUHES Application

A.1.1 Purpose of Study

Truth Goggles is a credibility layer for the Internet being developed at the MIT Media Lab as part of a master's thesis by Dan Schultz. It attempts to connect the dots between content online and the work done by fact checking organizations. The tool is designed to guide users through the process of critical media consumption by identifying moments where it is especially important to think carefully.

In order to increase the effectiveness of Truth Goggles the researchers would like to better understand how users might use it, how well it works, and what could be improved.

A.1.2 Experimental Procedures

This study will be conducted online through a sandboxed demo. Participants will be presented with a landing page disclosing their rights, describing Truth Goggles, and briefly explaining the goals of the study. After acknowledging this information the participant will be presented with a list of links to political articles and blog posts that have been copied and loaded into the Truth Goggles sandbox.

After clicking one of the links, the participant will be presented with the article's content. Depending on the treatment condition the participant may view the content with the assistance of a version of Truth Goggles. If Truth Goggles is active the user will see additional information presented on top of the article.

Information about the user's activity within the controlled sandbox, such as click rates, exploration of claims, and interactions with the tool, will be collected. Since the content pool is restricted to the articles and blog posts copied into the sandbox there are no privacy risks involved.

Once the participant is finished with an article, he or she may be presented with a small survey asking about the experience. Once the participant has finished interacting with the survey he or she will be returned to the list of articles. From here the participant may either continue exploring additional articles or may choose to end the study and be taken to a closing survey.

Once the participant has finished interacting with the closing survey he or she will be given the opportunity to subscribe to a mailing list about Truth Goggles if he or she is interested in learning more about the project as it progresses.

The pool of survey questions can be found in the attached materials. None of the survey questions ask for personally identifying information, although some of the survey responses are open format. Open format would allow the participants to accidentally write personally identifying information. Users will be reminded of this risk in the instruction page and asked to avoid providing this type of information accidentally.

A.1.3 Procedures for Obtaining Informed Consent

Because this experiment is being administered online it will not be possible to collect physical signatures. In order to participate in the study users will be required to view a landing page explaining the study, the tool, and the rights of the participant. Participants will not be allowed to continue into the study until after viewing this landing page and clicking a link to indicate understanding.

The landing page copy is attached.

A.1.4 Processes to Ensure Confidentiality

There are no user accounts, no names, and no explicitly personally identifiable information collected, although there will be open ended questions in the survey results. The only collected data will be basic use patterns on the sandboxed site and the survey results. Users will be reminded of this risk in the instruction page and asked to avoid providing this type of information accidentally.

Responses with personally identifiable information will be censored to remove the information in question before being included in publication.

A.2 User Study Materials

A.2.1 Study Recruitment

Below is a compilation of all official digital communication copy used to recruit participants in the user study testing Truth Goggles. Twitter and email will both be used as a recruitment technique with tracked click through rates.

Twitter

Tweet A Ive set up a test bed for Truth Goggles! Have a minute to help out by participating in a quick usability study? [Link here]

Tweet B If you have a chance please spread the word about the Truth Goggles user study: [Link here]

Tweet C Im collecting information about a few Truth Goggles prototype interfaces. Can you help? Check it out: [Link here]

Email

Email A Could you help me test out Truth Goggles? Ive set up a test bed and would love if you could help me out by trying Truth Goggles and answering a few questions as part of a usability study. You can learn more about participating here: [link to landing page].

Many thanks, Dan

Email B Would you be willing to spend a few minutes helping me test an online tool Im developing for my masters thesis? The entire process will take anywhere from 2 to 30 minutes, depending on how much you want to explore. You can learn more about the tool and about participating here: [link to landing page].

Many thanks, Dan

A.2.2 Landing Page

Truth Goggles is a credibility layer for the Internet being developed at the MIT Media Lab as part of a master's thesis by Dan Schultz. It attempts to connect the dots between content on your screen and the work done by fact checking organizations. The tool is designed to guide users through the process of critical media consumption by identifying moments where it is especially important to think carefully.

In order to increase the effectiveness of Truth Goggles the researchers would like to better understand how you would use it, how well it works, and what could be improved.

Please take a moment to read the following important points:

- Participation is voluntary
- You may decline to answer any or all questions
- This study will take approximately 20 minutes or less of your time
- You may stop participating at any point without any adverse consequences
- Your confidentiality is assured

If you decide to continue, you will be asked a few questions and shown a series of articles and blog posts. After reading an article you may be asked a short set of questions about your experience. At the end of the study you may be asked a short set of questions about your overall experience. Some of these questions may be open ended. Participants are asked to avoid providing personally identifiable information in their responses.

[Link to the study]

Note: by clicking this link you understand that participation is voluntary, you may decline to answer any or all questions, you may stop participating at any point without any adverse consequences, and that your confidentiality is assured.

A.2.3 Instructions

The following instructions were presented to the participants to explain how to continue. They are separated in terms of which portion of the study they appeared.

Introduction Instructions Thank you for participating in the Truth Goggles user study. This process should be short and interesting! If at any point you don't want to continue, simply close your browser window. If you take a long break after beginning the study this window will close automatically.

If you do not wish to answer a specific question you will always have the option of "skipping."

Prior Belief Survey You are about to be shown a series of claims. You will be asked to rate the accuracy of each claim on a scale ranging from "false" to "true." You will have approximately ten seconds to read and respond to each claim. If you do not wish to respond to a claim you may click "skip."

Articles You are about to be shown a series of short articles. Please read them as you would read any online content. When you are finished click the "Next" button which will be found at the end of the page.

Post-Article Survey You are about to be shown a series of claims. You will be asked to rate the accuracy of each claim on a scale ranging from "false" to "true." You will have approximately ten seconds to read and respond to each claim. If you do not wish to respond to a claim you may click "skip."

Final Survey Thank you very much for participating in the Truth Goggles user study! Below you will find a series of final questions about your experience. Please share any opinions as desired, but be sure to avoid providing personally identifiable information.

A.2.4 Questions

The following questions make up the pool which study participants may be exposed to. They are separated in terms of which portion of the study they appeared.

Prior Belief Survey Q: Please rate the accuracy of the following claim: [INSERT A CLAIM FROM THE CLAIM POOL HERE] A: True — Mostly True — Half True — Mostly False — False

Post-Article Survey Q: Did you notice that Truth Goggles was enabled? A: Yes — No

Q: Please rate the accuracy of the following claim: [INSERT A CLAIM FROM THE CLAIM POOL HERE] A: True — Mostly True — Half True — Mostly False — False

Final Survey Q: Did Truth Goggles affect your trust in the content you were reading? A: I generally felt I could trust the content more when Truth Goggles was enabled — I generally felt I could trust the content less when Truth Goggles was enabled — There was generally no change in how much I trusted the content when Truth Goggles was enabled

Q: Were you surprised by any of the content exposed by Truth Goggles? Please explain. A: Yes — No A2: [open format]

Q: Where would you place yourself on the political spectrum? A: Strong conservative — Moderate conservative — Independent — Moderate liberal — Strong liberal

Q: How was your reading experience different when Truth Goggles was enabled? A: [open format]

Q: Do you have any suggestions for Truth Goggles? A: [open format]

Q: Do you have any additional comments? A: [open format]

Unused Q: What pieces of information from the article do you remember most clearly? A: [short open format]

Q: How much time do you think you spent thinking about the content of this article? A: Less than a minute — 1-5 minutes — 5-10 minutes

Q: How much time do you think you spent thinking about the highlighted claims in this article? A: Less than a minute — 1-5 minutes — 5-10 minutes

Q: Did Truth Goggles affect your trust in the article author? A: I felt I could trust the article author more — I felt I could trust the article author less — There was no change in how much I trusted the article author

A.2.5 Articles

- **Republicans ramp up 'war on women' debate** url(http://news.yahoo.com/republicans-ramp-war-women-debate-172702689.html)
- **How About Those Big Oil Subsidies?** url(http://tobey100.hubpages.com/hub/How-About-Those-Big-Oil-Subsidies)
- **UC students worry about loan interest rate hike could add thousands to post-graduation debt** url(http://www.wcpo.com/dpp/news/local_news/uc-students-worry-about-loan-interest-rate-hike)

- **Obama goes back to 2008 playbook: Blame Bush** url(http://www.rogerhedgecock.com/story/1 times-obama-goes-back-to-2008-playbook-blame-bush)

- **Survey: 85% of New College Grads Move Back in with Mom and Dad** url(http://newsfeed.time.com/2011/05/10/survey-85-of-new-college-grads-moving-back-in-with-mom-and-dad/)

- **Without health care reform, 20-somethings out of luck** url(http://money.cnn.com/2012/04/16 care-young-adults/index.htm)

- **Jon Wills gift** url(http://www.washingtonpost.com/opinions/jon-will-40-years-and-going-with-down-syndrome/2012/05/02/gIQAdGiNxT_story.html)

- **Stronger Families, Stronger Societies** url(http://www.nytimes.com/roomfordebate/2012/04/24/ family-values-outdated/stronger-families-stronger-societies)

- **All measures of unemployment are falling** url(http://www.washingtonpost.com/blogs/ezra-klein/post/all-measures-of-unemployment-are-falling/2011/08/25/gIQAX5U5tQ_blog.html)

- **Billionaires with 1% tax rates** url(http://money.cnn.com/2011/12/07/news/economy/obama_tax

A.2.6 Fact Checks

- **Under Republican economic policies, "the typical American family saw their incomes fall by about 6 percent."** url(http://www.politifact.com/truth-o-meter/statements/2012/apr/26/barack-obama/barack-obama-says-family-incomes-have-fallen-6-per/)

- **"Women account for 92.3 percent of the jobs lost under Obama."** url(http://www.politifact. o-meter/statements/2012/apr/10/mitt-romney/romney-campaign-says-women-were-hit-hard-job-losse/)

- **"If you take into account all the people who are struggling for work, or have just stopped looking, the real unemployment rate is over 15 percent."** url(http://www.politifact.com/truth-o-meter/statements/2012/feb/08/mitt-romney/mitt-romney-says-broader-measure-national-unemploy/)

- **“After four years of a celebrity president 85% (of recent college grads are) moving back in with their parents.”** url(http://www.politifact.com/truth-o-meter/statements/2012/may/01/american-crossroads/american-crossroads-ad-says-85-percent-recent-coll/)

- **“Over 40 percent of children born in America are born out of wedlock.”** url(http://www.politifact.com/truth-o-meter/statements/2012/feb/24/rick-santorum/rick-santorum-says-over-40-percent-children-are-bo/)

- **After prenatal diagnosis, “90 percent of Down syndrome children in America are aborted.”** url(http://www.politifact.com/truth-o-meter/statements/2012/feb/27/rick-santorum/rick-santorum-says-90-percent-down-syndrome-childr/)

- **The oil industry subsidies that President Barack Obama is attacking dont exist** url(http://www.politifact.com/ohio/statements/2012/may/04/bill-johnson/bill-johnson-says-subsidies-oil-companies-barack-o/)

- **Because of the new health care law, “2.5 million young adults now have coverage.”** url(http://www.politifact.com/truth-o-meter/statements/2012/mar/16/barack-obama/barack-obama-film-touts-coverage-25-million-young-/)

- **The average Ohio student graduates from a four-year college or university with nearly $27,000 in tuition debt.** url(http://www.politifact.com/ohio/statements/2012/may/0
brown/sherrod-brown-says-ohio-students-graduate-college-/)

- **Wisconsin women “are paid 81 cents to the dollar of a man doing the same job.”** url(http://www.politifact.com/wisconsin/statements/2012/may/02/kathleen-falk/gubernatorial-hopeful-kathleen-falk-says-women-wis/)

- **“Some billionaires have a tax rate as low as 1 percent.”** url(http://www.politifact.com/truth-o-meter/statements/2011/dec/08/barack-obama/barack-obama-says-some-billionaires-have-tax-rate-/)

- **“President Barack Obama ”added“ $6.5 trillion to the national debt in his first term, more than the $6.3 trillion added by the previous 43 presidents**

combined.” url(http://www.politifact.com/new-jersey/statements/2012/may/04/chain-email/obama-has-added-more-national-debt-previous-43-pre/)

A.3 User Study Data

A.3.1 Control Groups

No Information The data in Table A.1 corresponds to claims that were found in no articles within the study. This means the user was provided no direct evidence about the claims during the study.

No Treatment The data in Table A.2 corresponds to claims found in articles that were presented to the user with no credibility layer at all. This means the user was asked to read an article which cited or referenced the fact checked claim and left to their own devices to reach a verdict without any influence from Truth Goggles.

A.3.2 Treatments

Highlight Mode The data in Table A.3 corresponds to claims found in articles that were presented to the user with the Truth Goggles highlight credibility layer automatically enabled. The highlight layer simply highlights claims that have been fact checked, allowing the user to explore if interested.

Goggles Mode The data in Table A.4 corresponds to claims found in articles that were presented to the user with the Truth Goggles goggles credibility layer automatically enabled. The goggles layer highlights claims that have been fact checked and blurs the remaining text, forcing the user to explore if he or she wishes to continue reading the article.

Safe Mode The data in Table A.5 corresponds to claims found in articles that were presented to the user with the Truth Goggles safe credibility layer automatically enabled. The safe layer censors fact checked claims that have not yet been explored, forcing the user to explore if interested discovering the content.

		95% C.I.			
Metric	Value	Min	Max	Alt. Hyp.	p
Average Verdict	-1.526	-1.605	-1.446	$x \neq 0$	l.t. 2.2e-16
Average Rating (pre)	0.0705	-0.1568	0.2978	$x \neq 0$	0.541
Average Rating (post)	0.0513	-0.1502	0.2528	$x \neq 0$	0.6159
Inaccuracy Saturation (pre)	0.5278	0.4734	0.5822		
Inaccuracy Saturation (post)	0.4872	0.4338	0.5406		
Inaccuracy Saturation Δ		-0.0353	0.1165	$x \neq 0$	0.2934
Inaccuracy Distance (pre)	1.981	1.760	2.201		
Inaccuracy Distance (post)	1.846	1.630	2.063		
Inaccuracy Distance Δ		-0.1731	0.4423	$x \neq 0$	0.39
Over Saturation (pre)	0.6346	0.5582	0.7110		
Over Saturation (post)	0.6538	0.5784	0.7293		
Over Saturation Δ		-0.1262	0.0878	$x \neq 0$	0.7238
Under Saturation (pre)	0.4054	0.2909	0.5199		
Under Saturation (post)	0.2838	0.1786	0.3889		
Under Saturation Δ		-0.0326	0.2758	$x \neq 0$	0.1212
Exact Saturation (pre)	0.1731	0.1131	0.2331		
Exact Saturation (post)	0.2115	0.1467	0.2763		
Exact Saturation Δ		-0.1264	0.0495	$x \neq 0$	0.3904
Total Bias Saturation (pre)	N/A due to claim pool bug				
Total Bias Saturation (post)	N/A due to claim pool bug				
Total Bias Saturation Δ	N/A due to claim pool bug				
Over Bias Saturation (pre)	N/A due to claim pool bug				
Over Bias Saturation (post)	N/A due to claim pool bug				
Over Bias Saturation Δ	N/A due to claim pool bug				
Under Bias Saturation (pre)	N/A due to claim pool bug				
Under Bias Saturation (post)	N/A due to claim pool bug				
Under Bias Saturation Δ	N/A due to claim pool bug				
Inaccuracy Direction (pre)	1.596	1.308	1.88		
Inaccuracy Direction (post)	1.577	1.312	1.84		
Drift	0.0192	-0.1774	0.1389	$x \neq 0$	0.8105
Absolute Drift	0.5833	0.4551	0.7116	$x \neq 0$	8.354e-16
Backfire Drift	N/A due to claim pool bug				
Backfire Drift Saturation	N/A due to claim pool bug				

Table A.1: Data analysis for the “no information” treatment.

		95% C.I.			
Metric	Value	Min	Max	Alt. Hyp.	p
Average Verdict	.0411	-0.2033	0.2855	$x \neq 0$	0.7401
Average Rating (pre)	0	-0.2240	0.2240	$x \neq 0$	1
Average Rating (post)	0.5479452	0.3272	0.7687	$x \neq 0$	2.475e-06
Inaccuracy Saturation (pre)	0.3373	0.2853	0.3893		
Inaccuracy Saturation (post)	0.4909	0.4344	0.5473		
Inaccuracy Saturation Δ		0.08948744	∞	$x > 0$	4.808e-05
Inaccuracy Distance (pre)	1.068	0.9002	1.235		
Inaccuracy Distance (post)	1.575	1.382	1.7682		
Inaccuracy Distance Δ		0.2932	∞	$x > 0$	5.679e-05
Over Saturation (pre)	0.4312	0.3367	0.5257		
Over Saturation (post)	0.6972	0.6100	0.7849		
Over Saturation Δ		0.1587	∞	$x > 0$	3.015e-05
Under Saturation (pre)	0.4386	0.3461	0.5311		
Under Saturation (post)	0.3509	0.2619	0.4398		
Under Saturation Δ		-0.2153	0.0399	$x \neq 0$	0.177
Exact Saturation (pre)	0.3356	0.2581	0.4131		
Exact Saturation (post)	0.2055	0.1392	0.2718		
Exact Saturation Δ		$-\infty$	-0.0450	$x < 0$	0.0061
Total Bias Saturation (pre)	0.5971	0.5381	0.6561		
Total Bias Saturation (post)	0.7213	0.6725	0.7700		
Total Bias Saturation Δ		0.0604	Inf	$x > 0$	0.0008
Over Bias Saturation (pre)	0.6312	0.5473	0.7152		
Over Bias Saturation (post)	0.7588	0.7013	0.8162		
Over Bias Saturation Δ		0.0433	∞	$x > 0$	0.0068
Under Bias Saturation (pre)	0.565	0.4802	0.6498		
Under Bias Saturation (post)	0.65	0.5608	0.7392		
Under Bias Saturation Δ		0.0165	∞	$x < 0$	0.0836
Inaccuracy Direction (pre)	0.0411	-0.2840	0.2018	$x \neq 0$	0.7386
Inaccuracy Direction (post)	0.5068	0.2458	∞	$x > 0$	0.0008
Drift	0.5479	0.3508	∞	$x > 0$	4.535e-06
Absolute Drift	1.096	0.9195	1.272		
Backfire Drift	0.1918	0.1272	0.2564		
Backfire Drift Saturation	0.2917	0.1990	0.3842		

Table A.2: Data analysis for the “no layer” treatment.

		95% C.I.			
Metric	Value	Min	Max	Alt. Hyp.	p
Average Verdict	-0.0781	-0.2533	0.0970	$x \neq 0$	0.3806
Average Rating (pre)	-0.2813	-0.4342	-0.1283	$x \neq 0$	0.0004
Average Rating (post)	0.0508	-0.1260	0.2276	$x \neq 0$	0.5722
Inaccuracy Saturation (pre)	0.3379	0.3022	0.3736		
Inaccuracy Saturation (post)	0.2308	0.1946	0.2669		
Inaccuracy Saturation Δ		$-\infty$	-0.0646	$x < 0$	1.943e-05
Inaccuracy Distance (pre)	1.039	0.9264	1.152		
Inaccuracy Distance (post)	0.684	0.5700	0.7972		
Inaccuracy Distance Δ		$-\infty$	-0.2215764	$x < 0$	7.372e-06
Over Saturation (pre)	0.3819	0.3138	0.4500		
Over Saturation (post)	0.3367	0.2705	0.4029		
Over Saturation Δ		-0.1399	0.0495	$x \neq 0$	0.3483
Under Saturation (pre)	0.5098	0.4406	0.5790		
Under Saturation (post)	0.2598	0.1991	0.3205		
Under Saturation Δ		$-\infty$	-0.1731	$x < 0$	7.184e-08
Exact Saturation (pre)	0.2969	0.2405	0.3532		
Exact Saturation (post)	0.5313	0.4697	0.5928		
Exact Saturation Δ		0.1646	∞	$x > 0$	2.546e-08
Total Bias Saturation (pre)	0.5694	0.5288	0.6101		
Total Bias Saturation (post)	0.6035	0.5518	0.6551		
Total Bias Saturation Δ		-0.0314	0.0995	$x \neq 0$	0.3067
Over Bias Saturation (pre)	0.5241	0.4625	0.5858		
Over Bias Saturation (post)	0.5908	0.5272	0.6544		
Over Bias Saturation Δ		-0.0211	0.1544	$x \neq 0$	0.1354
Under Bias Saturation (pre)	0.6026	0.5486	0.6565		
Under Bias Saturation (post)	0.6195	0.5323	0.7067		
Under Bias Saturation Δ		-0.0849	0.1187	$x \neq 0$	0.7419
Inaccuracy Direction (pre)	-0.2031	-0.3719	-0.0343	$x \neq 0$	0.0185
Inaccuracy Direction (post)	0.1289	-0.0117	0.2695	$x \neq 0$	0.0722
Drift	0.3320	0.1495	0.5146	$x \neq 0$	0.0004
Absolute Drift	1.082	0.9509	1.213		
Backfire Drift	0.0547	0.0266	0.0827		
Backfire Drift Saturation	0.0769	0.0378	0.1160		

Table A.3: Data analysis for the “highlight” treatment.

		95% C.I.			
Metric	Value	Min	Max	Alt. Hyp.	p
Average Verdict	-0.3072	-0.5315	0.0830	$x \neq 0$	0.0075
Average Rating (pre)	-0.3253	-0.5328	-0.1178	$x \neq 0$	0.0023
Average Rating (post)	0.0508	-0.3473	0.1064	$x \neq 0$	0.2959
Inaccuracy Saturation (pre)	0.3358	0.2866	0.3851		
Inaccuracy Saturation (post)	0.2701	0.2230	0.3171		
Inaccuracy Saturation Δ		$-\infty$	-0.0089	$x < 0$	0.0288
Inaccuracy Distance (pre)	1.042	0.8909	1.193		
Inaccuracy Distance (post)	0.7771	0.6471	0.9071		
Inaccuracy Distance Δ		$-\infty$	-0.0984	$x < 0$	0.0046
Over Saturation (pre)	0.4074	0.3235	0.4914		
Over Saturation (post)	0.3778	0.2949	0.4606		
Over Saturation Δ		-0.1470	0.0878	$x \neq 0$	0.6197
Under Saturation (pre)	0.4576	0.3664	0.5488		
Under Saturation (post)	0.3475	0.2603	0.4346		
Under Saturation Δ		$-\infty$	-0.0050	$x < 0$	0.0426
Exact Saturation (pre)	0.3434	0.2704	0.4164		
Exact Saturation (post)	0.4458	0.3694	0.5222		
Exact Saturation Δ		0.0141	∞	$x > 0$	0.02826
Total Bias Saturation (pre)	0.6154	0.5589	0.6720		
Total Bias Saturation (post)	0.6033	0.5403	0.6662		
Total Bias Saturation Δ		-0.0963	0.0719	$x \neq 0$	0.7755
Over Bias Saturation (pre)	0.5530	0.4728	0.6333		
Over Bias Saturation (post)	0.5703	0.4877	0.6528		
Over Bias Saturation Δ		-0.0965	0.1310	$x \neq 0$	0.7645
Under Bias Saturation (pre)	0.6790	0.6003	0.7577		
Under Bias Saturation (post)	0.6443	0.5448	0.7438		
Under Bias Saturation Δ		-0.1599	0.0905	$x \neq 0$	0.5829
Inaccuracy Direction (pre)	-0.0181	-0.2384	0.2022	$x \neq 0$	0.8715
Inaccuracy Direction (post)	0.1867	0.0125	0.3609	$x \neq 0$	0.0358
Drift	0.2048	-0.0335	0.4431	$x \neq 0$	0.0916
Absolute Drift	1.096	0.9250	1.268		
Backfire Drift	0.0542	0.0194	0.0890		
Backfire Drift Saturation	0.0882	0.0322	0.1442		

Table A.4: Data analysis for the “goggles” treatment.

		95% C.I.			
Metric	Value	Min	Max	Alt. Hyp.	p
Average Verdict	-0.0629	-0.2842	0.1584	$x \neq 0$	0.5754
Average Rating (pre)	0.1447	-0.3485	0.0592	$x \neq 0$	0.1631
Average Rating (post)	0.0692	-0.1600	0.2984	$x \neq 0$	0.5519
Inaccuracy Saturation (pre)	0.3690	0.3219	0.4160		
Inaccuracy Saturation (post)	0.3014	0.2530	0.3497		
Inaccuracy Saturation Δ		$-\infty$	-0.0112	$x < 0$	0.02437
Inaccuracy Distance (pre)	1.164	1.012	1.315		
Inaccuracy Distance (post)	0.8868	0.7447	1.029		
Inaccuracy Distance Δ		$-\infty$	-0.1033	$x < 0$	0.0045
Over Saturation (pre)	0.4553	0.3660	0.5445		
Over Saturation (post)	0.4065	0.3185	0.4945		
Over Saturation Δ		-0.1735	0.0760	$x \neq 0$	0.4419
Under Saturation (pre)	0.4697	0.3834	0.5560		
Under Saturation (post)	0.3636	0.2805	0.4468		
Under Saturation Δ		$-\infty$	-0.0061	x¡0	0.0405
Exact Saturation (pre)	0.2579	0.1891	0.3266		
Exact Saturation (post)	0.3836	0.3072	0.4601		
Exact Saturation Δ		0.0399	∞	$x > 0$	0.0081
Total Bias Saturation (pre)	0.6271	0.5739	0.6803		
Total Bias Saturation (post)	0.6454	0.5839	0.7069		
Total Bias Saturation Δ		-0.0626	0.0991	$x \neq 0$	0.656
Over Bias Saturation (pre)	0.5685	0.4894	0.6475		
Over Bias Saturation (post)	0.6733	0.5903	0.7563		
Over Bias Saturation Δ		-0.0084	0.2181	$x \neq 0$	0.0692
Under Bias Saturation (pre)	0.6801	0.6089	0.7513		
Under Bias Saturation (post)	0.6163	0.5228	0.7098		
Under Bias Saturation Δ		-0.1800	0.0524	$x \neq 0$	0.2787
Inaccuracy Direction (pre)	-0.0818	-0.3188	0.1553	$x \neq 0$	0.4968
Inaccuracy Direction (post)	0.1321	-0.0658	0.3300	$x \neq 0$	0.1894
Drift	0.2139	-0.0362	0.4639	$x \neq 0$	0.0932
Absolute Drift	1.195	1.026	1.364		
Backfire Drift	0.0629	0.0247	0.1010		
Backfire Drift Saturation	0.0901	0.0360	0.1442		

Table A.5: Data analysis for the “safe” treatment.

A.4 Licenses

At the time of this document's publication, the source code behind the Truth Goggles project was released to the world using the GPL. It can be accessed online through the Truth Goggles Git repository [64].

Appendix B

Bibliography

[1] 4chan. Bel-air (fresh prince) — know your meme. http://knowyourmeme.com/memes/bel-air-fresh-prince, 2009. accessed 17-May-2012.

[2] 4chan. Milhouse is not a meme — know your meme. http://knowyourmeme.com/memes/milhouse-is-not-a-meme, 2009. accessed 17-May-2012.

[3] Bill Adair. Politifact — about politifact. http://www.politifact.com/. accessed 17-May-2012.

[4] Bill Adair. Politifact — sorting out the truth in politics. http://www.politifact.com/. accessed 17-May-2012.

[5] Adblock. Adblock plus for chrome for annoyance-free web surfing. http://adblockplus.org/. accessed 17-May-2012.

[6] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2007. ACM.

[7] Adobe. Html editor software, web design software — adobe dreamweaver cs6. http://www.adobe.com/products/dreamweaver.html. accessed 17-May-2012.

[8] Jostin Asuncion. X is not your personal army — know your meme. http://knowyourmeme.com/memes/x-is-not-your-personal-army, 2009. accessed 17-May-2012.

[9] Roy F. Baumeister and Leonard S. Newman. Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin*, 20(1):3–19, 1994.

[10] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press, 2006.

[11] Shea Bennett. Fox news obama bin laden dead typo causes twitter backlash. http://www.mediabistro.com/alltwitter/fox-news-obama-bin-laden_b7943, 2011. accessed 17-May-2012.

[12] Roger E. Bohn and James E. Short. How much information? 2009 report on american consumers. http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf, 2009. accessed 17-May-2012.

[13] Arthur S. Brisbane. Should the times be a truth vigilante? http://publiceditor.blogs.nytimes.com/2012/01/12/should-the-times-be-a-truth-vigilante/, 2012. accessed 17-May-2012.

[14] Arthur S. Brisbane. Update to my previous post on truth vigilantes. http://publiceditor.blogs.nytimes.com/2012/01/12/update-to-my-previous-post-on-truth-vigilantes/, 2012. accessed 17-May-2012.

[15] John G. Bullock. PhD thesis.

[16] Julie G. Bush, Hollyn M. Johnson, and Colleen M. Seifert. The implications of corrections: Then why did you mention it? 1994.

[17] Hadley Cantril. *The Invasion From Mars: A Study In The Psychology Of Panic.* Transaction Publishers, 1940.

[18] Pew Research Center. Press widely criticized, but trusted more than other information sources. http://www.people-press.org/files/legacy-pdf/9-22-2011\%20Media\%20Attitudes\%20Release.pdf, 2011. accessed 17-May-2012.

[19] Michael D. Cobb, Brendan Nyhan, and Jason Reifler. Beliefs Dont Always Persevere: How political figures are punished when positive information about them is discredited. http://www.dartmouth.edu/~nyhan/positive-misinformation.pdf.

[20] John Cook and Stephan Lewandowsky. The debunking handbook. http://www.skepticalscience.com/docs/Debunking_Handbook.pdf, 2011.

[21] Michael Dobbs. The rise of political fact-checking - how reagan inspired a journalistic movement: A reporters eye view. 2012.

[22] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 341–350. ACM, 2010.

[23] Fabrice Florin. Your guide to good journalism - newstrust. http://www.newstrust.net. accessed 17-May-2012.

[24] Media Matters for America. Media matters for america. http://mediamatters.org/. accessed 17-May-2012.

[25] Mark Frauenfelder. Homophobic news site changes athlete tyson gay to tyson homosexual - boing boing. http://boingboing.net/2008/06/30/homophobic-news-site.html, 2008. accessed 17-May-2012.

[26] Futurama. Hypnotoad — know your meme. http://knowyourmeme.com/memes/hypnotoad, 2009. accessed 17-May-2012.

[27] Matthew Gerke and John Cabral. Lazytruth. http://www.skeptive.com/. accessed 17-May-2012.

[28] Albert C. Gunther and Kathleen Schmitt. Mapping boundaries of the hostile media effect. *Journal of Communication*, 54(1):55–70, 2004.

[29] Matt Hickey. Charlie sheen-censoring browser plug-in is here. http://news.cnet.com/8301-17938_105-20040766-1.html. accessed 17-May-2012.

[30] Twitter Inc. Twitter / home. https://twitter.com/. accessed 17-May-2012.

[31] The Internet. This looks shopped — know your meme. http://knowyourmeme.com/memes/this-looks-shopped, 2010. accessed 17-May-2012.

[32] The Internet. Cats — know your meme. http://knowyourmeme.com/memes/subcultures/cats, 2011. accessed 17-May-2012.

[33] Brooks Jackson. Factcheck.org — a project of the annenberg public policy center. http://www.factcheck.org/. accessed 17-May-2012.

[34] Brooks Jackson. Factcheck.org : About us. http://www.factcheck.org/. accessed 17-May-2012.

[35] Eliana Johnson. Reuters pulls 920 pictures by discredited photographer. http://www.nysun.com/foreign/reuters-pulls-920-pictures-by-discredited/37474/, 2006. accessed 17-May-2012.

[36] jQuery Project. jquery: The write less, do more, javascript library. http://jquery.com/. accessed 17-May-2012.

[37] Glenn Kessler. About the fact checker. http://www.washingtonpost.com/blogs/fact-checker/post/about-the-fact-checker/2011/12/05/gIQAa0FBYO_blog.html. accessed 17-May-2012.

[38] Zombie Kid. I like turtles — know your meme. http://knowyourmeme.com/memes/i-like-turtles, 2009. accessed 17-May-2012.

[39] Spee Kosloff, Jeff Greenberg, Toni Schmader, Mark Dechesne, and David Weise. Smearing the opposition: Implicit and explicit stigmatization of the 2008 u.s. presidential candidates and the current u.s. president. *Journal of Experimental Psychology: General*, 139(3):383–398, 2010.

[40] Arie W. Kruglanski and Donna M. Webster. Motivated closing of the mind: Seizing and freezing. *Psychological Review*, 103(2):263–283, 1996.

[41] Aparna Kumar. Third voice trails off... http://www.wired.com/techbiz/media/news/2001/04/42803, 2001. accessed 17-May-2012.

[42] The Yes Men. Dow does the wrong thing — the yes men. `http://theyesmen.org/hijinks/bhopalpressrelease`. accessed 17-May-2012.

[43] Barbara Mikkelson and David Mikkelson. snopes.com: About the people behind snopes.com. `http://snopes.com/info/aboutus.asp`. accessed 17-May-2012.

[44] Chris Mooney. Can Geeks Defeat Lies? Thoughts on a Fresh New Approach to Dealing With Online Errors, Misrepresentations, and Quackery. `http://www.desmogblog.com/can-geeks-defeat-lies-thoughts-fresh-new-approach-dealing-online-errors-misrepresentations-and-quackery`, 2012. accessed 17-May-2012.

[45] Jonathan S Morris. Slanted objectivity? perceived media bias, cable news exposure, and political attitudes. *Social Science Quarterly*, 88(3):707–728, 2007.

[46] Randall Munroe. Citogenesis. `http://xkcd.com/978/`, 2011. accessed 17-May-2012.

[47] mySQL Project. Mysql :: The world's most popular open source database. `http://www.mysql.com`. accessed 17-May-2012.

[48] NewsBusters. Newsbusters.org — exposing liberal media bias. `http://newsbusters.org/`. accessed 17-May-2012.

[49] Brendan Nyhan and Jason Reifler. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, 32(2):303–330, June 2010.

[50] Brendan Nyhan and Jason Reifler. Opening the political mind? the effects of self-affirmation and graphical information on factual misperception. `http://www.dartmouth.edu/~nyhan/opening-political-mind.pdf`, 2011.

[51] Brendan Nyhan and Jason Reifler. Misinformation and fact-checking: Research findings from social science. 2012.

[52] Eli Pariser. The filter bubble. `http://www.thefilterbubble.com/ted-talk`, 2011. accessed 17-May-2012.

[53] Andrew Phelps. How to peek through dan schultzs truth goggles, the b.s. detection software, right now. http://www.niemanlab.org/2012/05/how-to-peek-through-dan-schultzs-truth-goggles-the-b-s-detection-software-right-now/. accessed 17-May-2012.

[54] Luminoso Project. Luminoso. http://lumino.so/, note = accessed 17-May-2012.

[55] MediaCloud Project. Media cloud. http://www.mediacloud.org, note = accessed 17-May-2012.

[56] NLTK Project. Natural language toolkit nltk 2.0 documentation. http://www.nltk.org/. accessed 17-May-2012.

[57] OAuth Project. Oauth 2.0 oauth. http://oauth.net/2/, note = accessed 17-May-2012.

[58] PHP Project. Php: Hypertext preprocessor. http://php.net/. accessed 17-May-2012.

[59] Python Project. Python programming language official website. http://www.python.org/. accessed 17-May-2012.

[60] Fairness Accuracy In Reporting. Fairness accuracy in reporting (fair). http://www.fair.org/. accessed 17-May-2012.

[61] Matt Richardson. Make — enough already: The arduino solution to overexposed celebs. http://blog.makezine.com/2011/08/16/enough-already-the-arduino-solution-to-overexposed-celebs. accessed 17-May-2012.

[62] Alex Rodriguez. Restful web services: The basics. https://www.ibm.com/developerworks/webservices/library/ws-restful/. accessed 17-May-2012.

[63] RTMark. Rtmark: An answer to the wef — reamweaver — past projects. http://www.rtmark.com/wef.html. accessed 17-May-2012.

[64] Daniel Schultz. slifty/truth-goggles. https://github.com/slifty/truth-goggles. accessed 17-May-2012.

[65] Daniel Schultz and Sasha Costanza-Chock. newsjack.in. `http://www.newsjack.in`. accessed 17-May-2012.

[66] Craig Silverman. *Regret the Error: How Media Mistakes Pollute the Press and Imperil Free Speec.* Sterling Publishing Company, 2007.

[67] Ian Skurnik, Carolyn Yoon, Denise C. Park, and Norbert Schwarz. How warnings about false claims become recommendations. *Journal of Consumer Research*, 31:713–724, 2005.

[68] Matt Stempeck. Lazytruth. `http://www.lazytruth.com/`, 2012. accessed 17-May-2012.

[69] Cass R. Sunstein. *Republic.com 2.0.* Princeton University Press, 2007.

[70] Charles S. Taber and Milton Lodge. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3):755769, 2006.

[71] Yariv Tsfati and Joseph N. Cappella. Do People Watch what they Do Not Trust?Exploring the Association between News Media Skepticism and Exposure. *Communication Research*, 30(5):504529, 2003.

[72] Press Pass TV. Press pass tv — media that moves -. `http://presspasstv.org/`. accessed 17-May-2012.

[73] R. P. Vallone, L. Ross, and M. R. Lepper. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of personality and social psychology*, 49(3):577–585, September 1985.

[74] Dan Whaley. Hypothes.is — the internet, peer reviewed. `http://hypothes.is/`. accessed 17-May-2012.

[75] Wikipedia. Wikipedia. `http://www.wikipedia.org/`. accessed 17-May-2012.

[76] WikiTrust. Wikitrust. `http://www.wikitrust.net/`. accessed 17-May-2012.