CredibleCrowd: Empirical Mechanism Design for Crowdsourced Fact-checking

Mohammad Yaghini
Data Science Lab (DLAB)
EPFL

June 2018

Contents

Chapter 1

Overview

1.1 Introduction

With the advent of new propagation media, like social media and internet news websites, detecting and combating misinformation and fake (or false) news has gained renewed interest. This interest has recently been intensified due to the major role that fake news campaigns played in recent political overturns.

Detecting false news is no easy task. Aside from technical difficulties, even defining what is fake or what is not, is a highly polarizing issue. Interestingly, opinion polarization is both the cause and the effect of fake news. In such a polarized environment, we need systems, rule and regulations to bridge the gap.

When parties stop arguing and agree to disagree, they head to the polling stations to *vote*, with the promise that whatever may be the outcome, it is the "wisdom of the crowd, " and as such, should be respected. The question is why can't we settle our disputes about fake news through such a voting mechanism?

One might argue that holding a vote for every controversial piece of news is economically unfeasible. Plus, a vote to determine the veracity of a claim which can be proved or disproved objectively given enough proof is unnecessary if not outright wrong.

While all these concerns are valid, they are so in an ideal world. One in which we have the resources to investigate every claim. The reality is far from it. We are bombarded with false information on a daily basis, and no person or agency can take up the responsibility of fact checking all of the news we consume. Even if there was such an agency, it could not have assumed a non-political, not partisan agency. Its decisions would have been called an "alternative" fact, and its conclusions would have thus been disregarded.

Automated fact-checking is a promising tool that would alleviate many of the aforementioned issues. Such a system — while potentially expensive to create, setup and maintain — is scalable. It can collect opinion of the masses on a wide variety of issues, and leverage this wisdom of the crowd with transparent rules that enjoy bipartisan support.

An automated fact-checking system can be built to *extract expertise* and *incentivize truthfulness and effortful work*. CredibleCrowd is such a system.

1.2 System Objectives

1. Detecting the fake news pieces
2. Motivating the crowd to detect/report fake news
3. Finding arguments for and against a fact-checking decision and initiating a conversation with added value in the process

1.3 Overview

In Figure (**??**) you can see the overview of the CredibleCrowd. While this is not an exhaustive list of features, it outlines a few key points regarding the design of the system:

1. The system is designed based on cutting-edge research in the domain of mechanism design, cognition theory and behavioral studies and computer science.
2. It is fine-tuned by *iterative experiments* discovering the interplay of different schemes of voting and truthful information solicitation in realizing the objectives of the system.
3. The experiments will be designed to test theories in three main categories:
   (a) How does a mechanism (or a particular implementation of it) lends itself to the problem of fake new detection?
   (b) How is the interplay of Voting and Truthfulness-ensuring mechanisms? Are the two working effectively together?
   (c) Upon success of the last phase of experiments, next we would be interested to see how the addition of annotation, argument mining/discovery and online discussions would change the dynamics of the system. Is it upgrading or degrading the system performance?

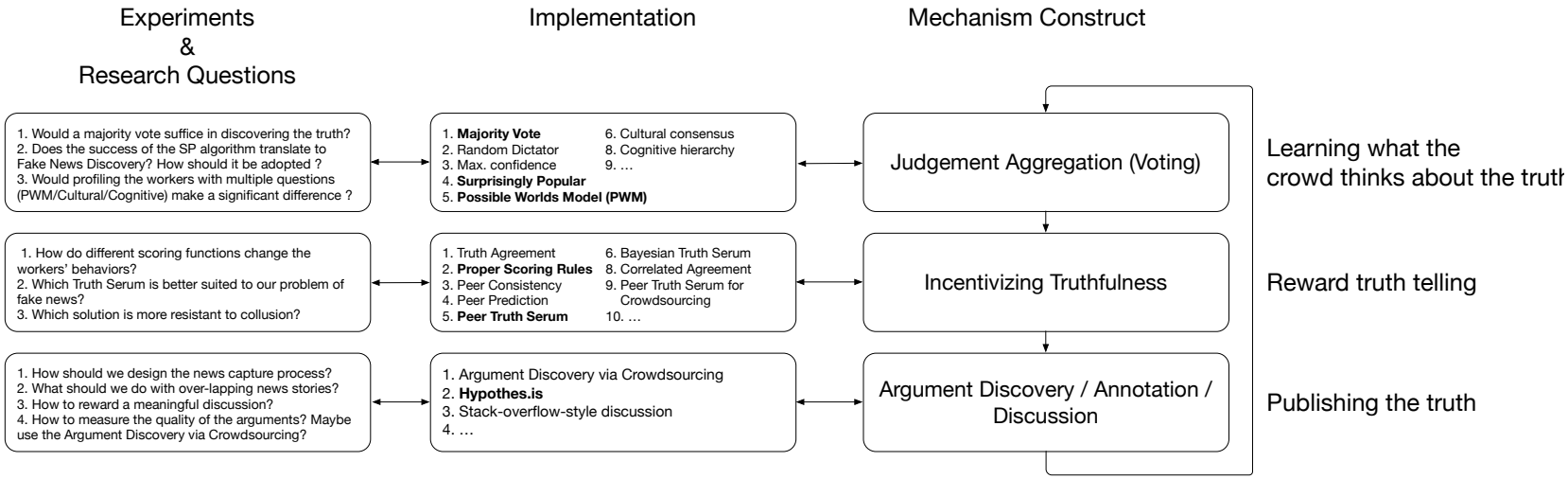


Figure 1.1: Overview of the CredibleCrowd mechanism construct, implementation methods and related experiments and research questions

Chapter 2

Judgment Aggregation By Voting

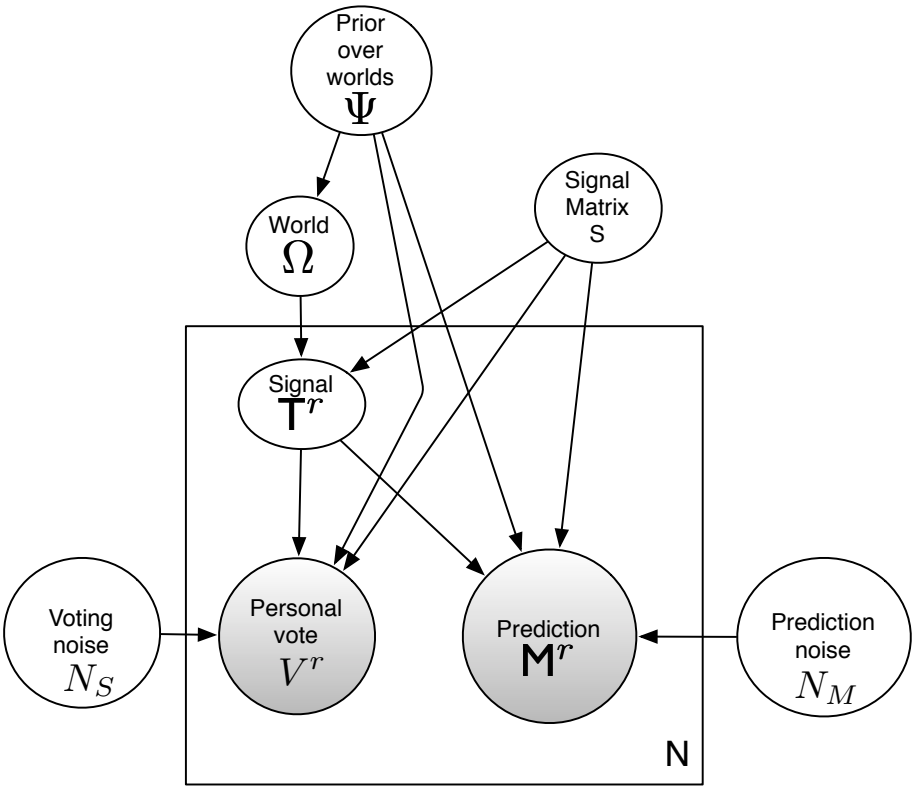
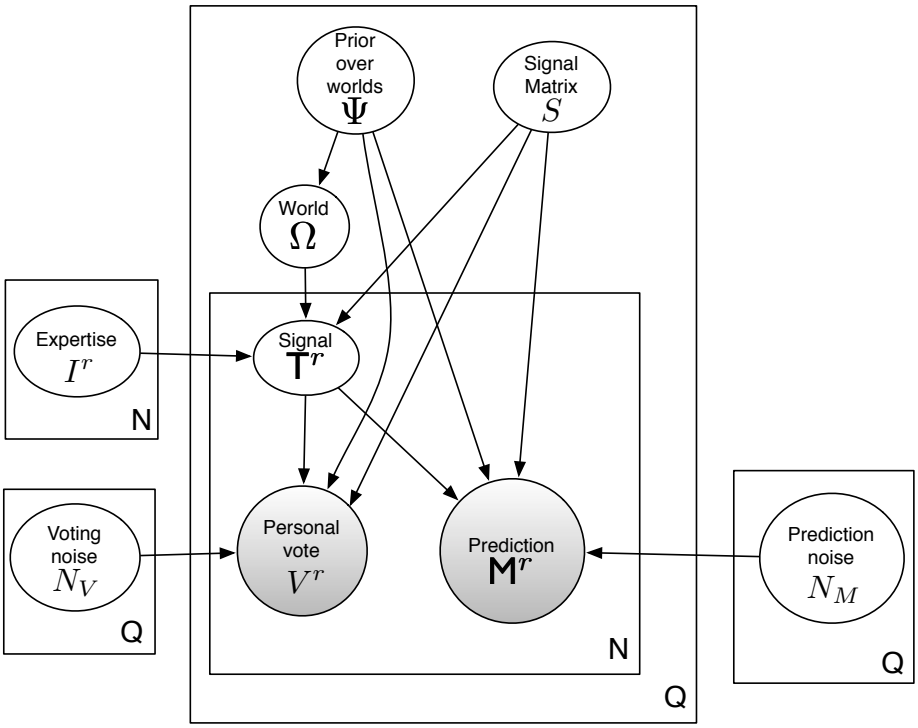


Figure 2.1: The single question possible worlds model (PWM) which is used to infer the underlying world state based on a group's votes and predictions of the votes of others. In keeping with standard graphical model plate notation, nodes are random variables, shaded nodes are observed, an arrow from node X to node Y denotes that Y is conditionally dependent on X, a rectangle around variables indicates that the variables are repeated as many times as indicated in the lower right corner of the rectangle.



2.1 Surprisingly Popular

2.1.1 Example

Imagine a piece of news. The world has a true state $\omega^* \in \Omega$ and $\Omega = \{A, B\}$. There are two options $\{\tilde{A}, \tilde{B}\}$. $\tilde{A}$ and $\tilde{B}$ are opposites: one says that the news is True, while the other says it is False (fake). It is assumed that in state A answer $\tilde{A}$ is the true answer, while in state B, answer $\tilde{B}$ is correct.

The correct answer is not known, or equivalently, people do not know what is the true state ω^* of the world. However, it is assumed that if they knew, they would give a utility of 1 to the true answer and 0 to the other.

A worker w_i is given the piece of news and is asked to fact-check it. He uses the sources of information available to him or her, and receives a signal $s \in a, b$.

Imagine that signal a has probability 0.8 in state A and probability 0.7 in state B and thus be more likely than signal b in both world states. Under this signal distribution, if the actual state was B then the majority would receive signal a, and, assuming a uniform common prior, would vote for state A and so would be incorrect.

If the received signal is a assuming that this signal has a uniform common prior, Then, its estimate of the probability that the world is in state A is $p(\Omega = A|s = a) = p(s = a|\Omega = A)p(\Omega = A)/p(s = a) = (.8)(.5)/((.8)(.5) + (.7)(.5))$ and since this quantity is higher than 0.5, workers receiving signal a vote that

the world is most likely in state A. But if the actual world is state B, workers have a .7 probability of receiving signal a and hence the majority of workers will vote incorrectly.[**mccoy:stat**]

However, using the Surprising Popular voting mechanism, after gathering the both the votes and predictions of the all the workers, we can see whether True or False has been more popular than expected in *both* of the words. This, in turn, is our correct fact checking answer.

2.2 Truthful people assumption

In authors in [**mccoy:stat**] and its preceding work [**prelec:nature**] assume that people are truthful: "people vote for the world state that they believe is most probable." In other words, they believe that eliciting truthful information via modifying the payment function is a separate task in designing voting mechanism. This suggestion is reflected in the separation of Voting and Truthfulness Incentivization submodules. In other words the Voting mechanism assumes truthfulness, and leaves the job of ensuring it to the latter submodule.

Chapter 3

Eliciting Truthful Information

3.1 Taxonomy and Categorization

- **Center (Us):** Who collects the data from workers
- **Data:**
  - **Objective:** every agent observes the same realization of the phenomenon.
  - **Subjective:** each agent observes a possibly different realization of the same phenomenon.
- **Center Goals:**
  - Objective Data =¿ Most Accurate Data
  - Subjective Data =¿ Obtain distribution of values observed by agents.
- **Verifiability:**
  - **Verifiable:** Center will eventually get access to the ground truth.
  - **Unverifiable:** There is no Ground Truth.
- (Worker) Strategies
  - **Heuristic:** reported value does not depend on an observation of the phenomenon.
  - **Cooperative:** the agent invests effort to observe the phenomenon and truthfully reports the observation.
- Truthful Strategies
  - Agents report their belief about the phenomenon truthfully.
  - Cooperative $\subset$ Truthful

3.2 Crowd-sourced Fact Checking

- **Objective Data:** each task (news piece) has a state (True/False) that each worker observes and reports. *Or*
- **Subjective Data:** each worker has a different belief about the veracity of a given task, and we want to correctly poll these opinions.

3.2.1 Why Peer Prediction makes sense in crowd-sourced fact checking?

"The trick is that the common signal that allows the agents to coordinate their strategies is the phenomenon they can all observe."

The true signal (the phenomenon which a piece of news describes) should be the only coordinating signal among two workers.

3.3 What to strive for in a mechanism?

- **Truthfulness:** induces agents to choose a cooperative and truthful strategy;
- **Individually Rational:** where agents can expect a positive utility from participating;
- **Positive Self-selection:** which means only agents that are capable of providing useful data can expect a positive utility from participating.

3.4 Effort: The Need for Compensation

Monetary such as actual money, reputation points, rewards, etc.

Influence or the leverage on the model that the center learns from data

3.5 Belief: The agent's local information

The belief of an information agent is characterized by a prior probability distribution

$$P_i(x) = p_i(x_1), ..., p_i(x_n) \tag{3.1}$$

of what the state X of the phenomenon might be.

Following an observation, based on the received signal s_i it will update its prior

$$P_i(X = x|S_i = s) = P_i(x|s) \tag{3.2}$$

to a posterior distribution.

Table 3.1: Verifiability $\iff$ Objectivity $\iff$ Ground Truth

Verifiability	Objectivity	Ground Truth	Example
Verifiable	Objective	Always exists for the true value	Temperature Measurements
Unverifiable	Objective Subjective	Doesn't exist for individual data (worker reports). Exists for distribution of the reports. (Every worker samples from the same distribution.)	Restaurant Reviews

3.5.1 Belief Update

Belief update using Bayes Law For **Objective** data, belief update should be self-dominating:

Definition 1. An agent's belief update is self-dominating if and only if the observed o has the highest probability among all possible values x:

$$q(o|o) > q(x|o), \forall x \neq o. \tag{3.3}$$

Ex.$\delta = 1/t$ would compute the moving average if o is the t'th observation.

For Subjective data, it is enough for belief update to be self-predicting:

Definition 2. An agent's belief update is self-predicting if and only if the observed o has the highest relative increase in probability among among all possible values:

$$q(o|o)/p(o) > q(x|o)/p(x), \forall x \neq o. \tag{3.4}$$

Example 1.

$$\hat{q} = \begin{cases} (1-\delta)p(x) + \delta & \text{for } x = o \\ (1-\delta)p(x) & \text{for } x \neq o \end{cases} = (1-\delta)p(x) + \delta.\mathbb{1}_{x=o} \tag{3.5}$$

3.6 Summary

See table (**??**) for a summary of the relation between verifiability, objectivity the state of the Ground Truth and a couple of examples.

3.7 Why Peer prediction wouldn't work as expected

- The main (very high-level) idea behind previous peer-prediction mechanisms can be understood as a "clever majority vote"—every agent is paid according to a specific similarity between her and her peer. Thus, they

point out that in the peer-grading example, coordinating on just checking the grammar can guarantee good agreement with other agents, but with substantially reduced effort.[**kong:noverify**]

- Gao et al point out that things are likely even worse than this. If the cheap signals correlate more than the expensive signals, then the peer-prediction techniques incentivize agents to not report the true answer, but instead focus on cheap signals! For example, in the essay grading above, it is likely that assessments of grammatical correctness will agree more than assessments of overall essay quality. Because of this, peer-prediction mechanisms will pay agents more overall for lower-quality information.[**kong:noverify**]

- One approach to counter coordinated low-quality strategies is to use trusted agents that provide the correct answers for some randomly selected subset of tasks. In a hybrid mechanism, agents' reports will be either compared to other agents, or to such trusted reports. If the probability of having a trusted agents as a peer is sufficiently high, other low-quality equilibria can be broken. However, It has recently been shown that if the coordinated low-quality strategy provides higher payoffs than the cooperative strategy (for example, because it involves no measurement noise), it may be better to use simple truth agreement rather than a combination with peer consistency mechanisms as a complement to the truthful reports.[**boi:book**, **gao:peer**]

- An issue related to precision is the risk of agents coordinating on a signal other than the phenomenon itself, called low-quality signal in Gao et al.[**gao:peer**]

Chapter 4

Experiments

Since we are dealing with people and how they deal with spread of misinformation, we must base our mechanism on observations from real crowds. Therefore, we have designed and conducted experiments to gauge the performance of different models at each phase of the system.

The current work only deals with the first phase, "Judgment Aggregation." Nevertheless, in the future work's section, we outline the design and requirements of experiments for second and third phases.

4.1 Experiment 1

In this experiment, we want to answer these research questions:

1. Would a majority vote suffice in discovering the truth?
2. Does the success of the SP algorithm translate to Fake News Discovery? How should it be adopted ?
3. Would profiling the workers with multiple questions (PWM/Cultural/Cognitive) make a significant difference ?

For this first experiment, we used the crowd-sourcing platform, Amazon Mechanical Turk (AMT). In AMT, a **requester** can post Human Intelligence Tasks (**HITs**) on a ledger, and MTurk **workers** with (optional) qualifications can accept the HIT and start working on this.

AMT is frequently used for tasks that are still onerous for machines to perform well, including labeling datasets etc. However, many researchers in varied fields from Human Cognition and Neuroscience to Psychology and Behavioral Studies, have been using AMT as a platform to perform controlled-experiments. In fact, these have become so frequent that separate platform have appeared to ease the process of conducting such experiments, such as *psiTurk*.

4.1.1 Experiment Setting

For this experiment, to avoid an unbalanced dataset of answers, we designed a web-page HIT, where we asked our respondents to answer three questions about 11 news pieces that we have selected from a labeled dataset of fake news [**allcott:stanford**]:

1. Is this news piece real or fake?
2. How confident are you in your answer?
3. What percentage of other American MechanicalTurk respondents do you think would answer "Real" to the first question?

The respondents had to provide a percentage above 50% for question 2, and a percentage between 0 and 100% for question 3. Moreover, in order to ensure that respondents have read the article, and are not just filling out the form randomly, we ask a forth question about a *honeypot* embedded in the text.

Honeypot is a relatively easy riddle to be solved by respondents for validating their engaged (effortful) answers before continuing to the next question. It is embedded into the news text, as an out-of-context sentence about a semantically unrelated thing like a fruit. Honeypots are hard to recognize by a syntax- or style-minded parser, but are obvious to a semantic-minded one.

Remuneration and Bonuses To replicate the experiment carried out by Prelec et. al in [**prelec:nature**], we decided to use an almost identical remuneration scheme to theirs. Therefore, the participants were paid a flat participation fee of $2 for their participation in the 10-minute study regardless of their answers. Their expertise was awarded by separate bonuses on their accuracy of answers and estimation of others, with separate $2 bonuses to the top 20%.

Question 2/11 Show Instructions

FEDERAL COURT: ASSAULT RIFLES ARE 'WEAPONS OF WAR' NOT PROTECTED BY SECOND AMENDMENT

A federal court in Maryland ruled Tuesday that assault rifles and other so-called "weapons of war" are not protected under the Second Amendment.
The U.S. Court of Appeals upheld a previous decision that created the original assault rifles ban in Maryland.
"We are convinced that the banned assault weapons and large-capacity magazines are among those arms that are 'like' M-16 rifles -- 'weapons that are most useful in military service' -- which the Heller Court singled out as being beyond the Second Amendment's reach," Judge Robert King wrote for the 10-4 decision.
"Put simply, we have no power to extend Second Amendment protection to the weapons of war that the Heller decision explicitly excluded from such coverage."
Citing the Heller case, King wrote that assault rifles are "devastating weapons of war whose only legitimate purpose is to lay waste to a battlefield full of combatants." But its sweet mango was coveted by neighbors.
"The majority concludes that the semiautomatic rifles banned by Maryland law are most useful in military service, even though they are not in regular use by any military force, including the United States Army," the decision said.
It noted that such weapons have also been used for recent mass shootings in Aurora, Colo., San Bernardino, Calif., and Orlando, Fla. -- making those cities "synonymous with the slaughters that occurred there."

1. Is this news piece real or fake?

- ○ Real
- ○ Fake

2. How confident are you in your answer?

[ ] %

3. What percentage of other American MechanicalTurk respondents do you think would answer "Real" to the first question?

[ ] % of respondents would say that this news article is real.

4. Show us that you have read the news.

Show us by typing the name of the **fruit** that you encountered in an out-of-context sentence, like "Lemon is good for a cold."

[ ]

Next

Figure 4.1: Sample question from Experiment 1

Is this news **real** or **fake** ?

Please read the news pieces that follow, and answer these questions:

1. If you think the news piece is **objectively fake** (it contains **mainly misinformation**) or not,
2. Your confidence in your response,
3. Your estimation of others' answers to the first question.
   (Note: The responders are all American MTurk workers.)

You can **double** or even **triple** your earnings with bonuses.

Things to know

- This experiment is not optimized for phones. If possible, please use a **desktop computer**.
- News pieces vary in length but the whole experiment is designed to take roughly 10 minutes. Do not rush; you are given 6 hours to finish the task.
- Your answers will be accepted (in 1 hour) and you will be paid $2 for your participation regardless of your answers, so tell us your honest opinion.
- Your expertise/knowledge will be awarded in the form of additional bonuses.

Bonuses

There are two bonuses that you can collect in addition to your flat participation fee.

Accuracy Bonus

Based on your answer to Questions 1 and 2 we calculate an *accuracy* score for you according to the following table, and calculate a sum of scores. The top 20% of MTurk workers who have the highest sum of scores will be awarded a bonus of $2.

Your confidence	Your score if you are correct	Your score if you are wrong
50%	0	0
60%	9	-11
70%	16	-24
80%	21	-39
90%	24	-56
100%	25	-75

- The more certain you claim to be, the more points you can win.
- As you approach 100%, the penalty for being incorrect climbs much faster than the gains for being correct.
- In general, you will score the most points if the numbers correspond to your true levels of confidence. *Expressing too much confidence is a common mistake in this game.*

Estimation Bonus

You will be rewarded an additional $2 bonus if you can make a good estimation of others' responses (Question 3). Note that the experiment is only open to American workers on MTurk. The bonus goes to the top 20% of MTurk workers with the best estimation.

Start the survey

Figure 4.2: Instructions Page for Experiment 1

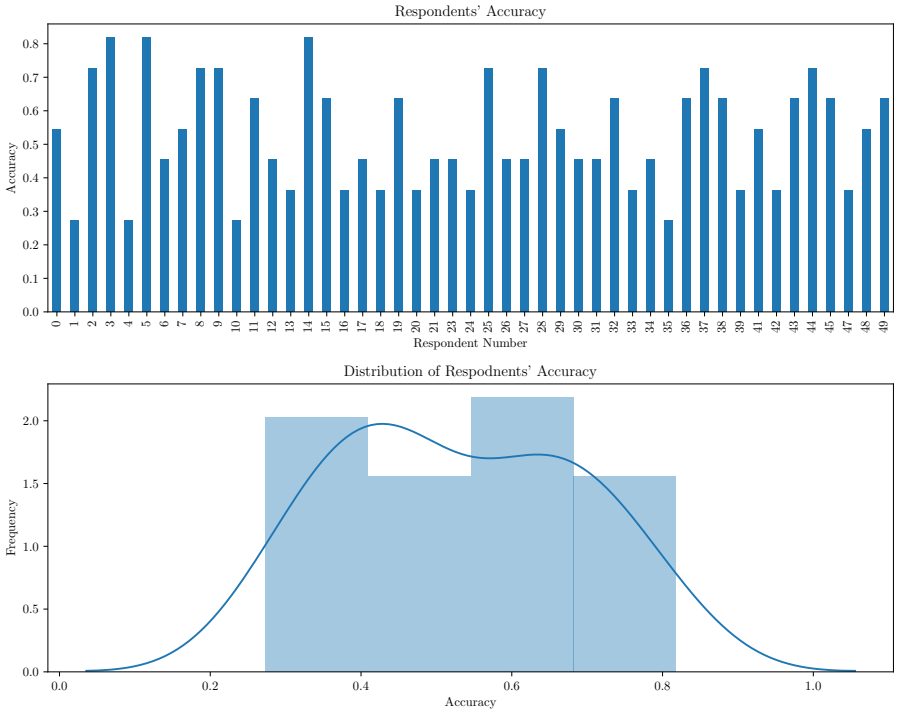


Figure 4.3: AMT Respondents are not very accurate. The mean accuracy is just 53%.

4.1.2 Results

50 participants took part in the experiment. 3 of the respondents submitted the experiment prematurely and as such their data was removed from the dataset.

4.1.3 Accuracy

The experiment contained 8 fake news articles and 3 real ones. Respondent's are not accurate on average (Figure **??**). The majority of answers

4.1.4 Confidence

Figure **??** shows the histogram of confidences expressed by respondents in each trial. We observe a wide range of confidences in most trials. For most of the trials, the reported confidence seems very uninformative. Indeed, as we see in section **??**, weighting the majority vote with the confidences expressed makes no difference in the performance of the voting scheme.

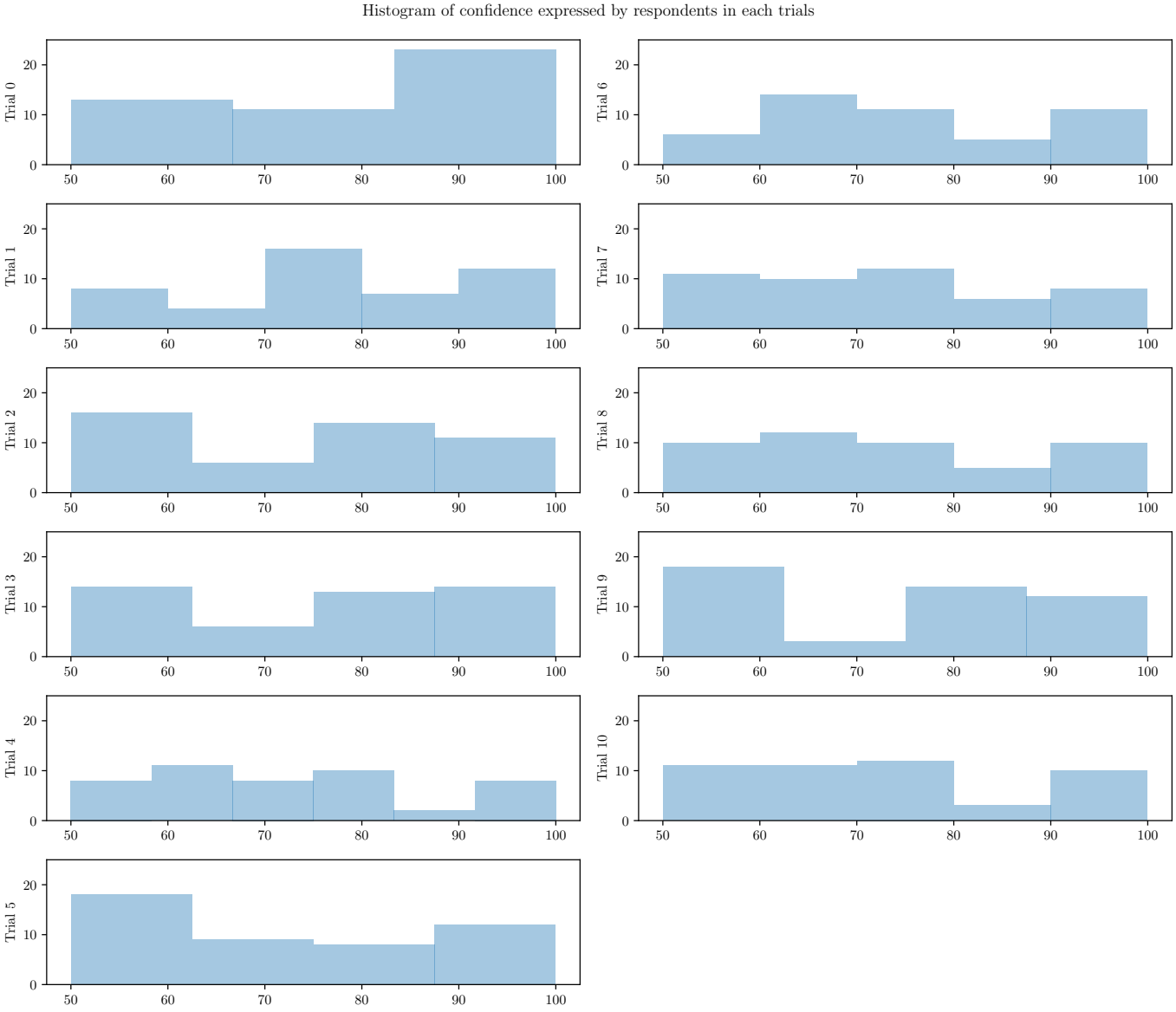


Figure 4.4: Histogram of confidences in the experiment. Respondents report a wide range of confidence in most questions.

4.1.5 Estimation Scores

We ask the respondents to give us an estimation of others responses to the first question. Specifically, we ask them how much of the respondents, they think, would consider the news piece real.

For the Surprisingly Popular (SP) voting scheme, we need to compare the average estimation for the each of two candidate answer to the actual votes they've gathered. Whichever has a higher share of the votes than was estimated (on average), is the surprisingly popular scheme.

Figure **??** shows the histogram of estimations, divided into two partitions — based on the actual vote of participants.

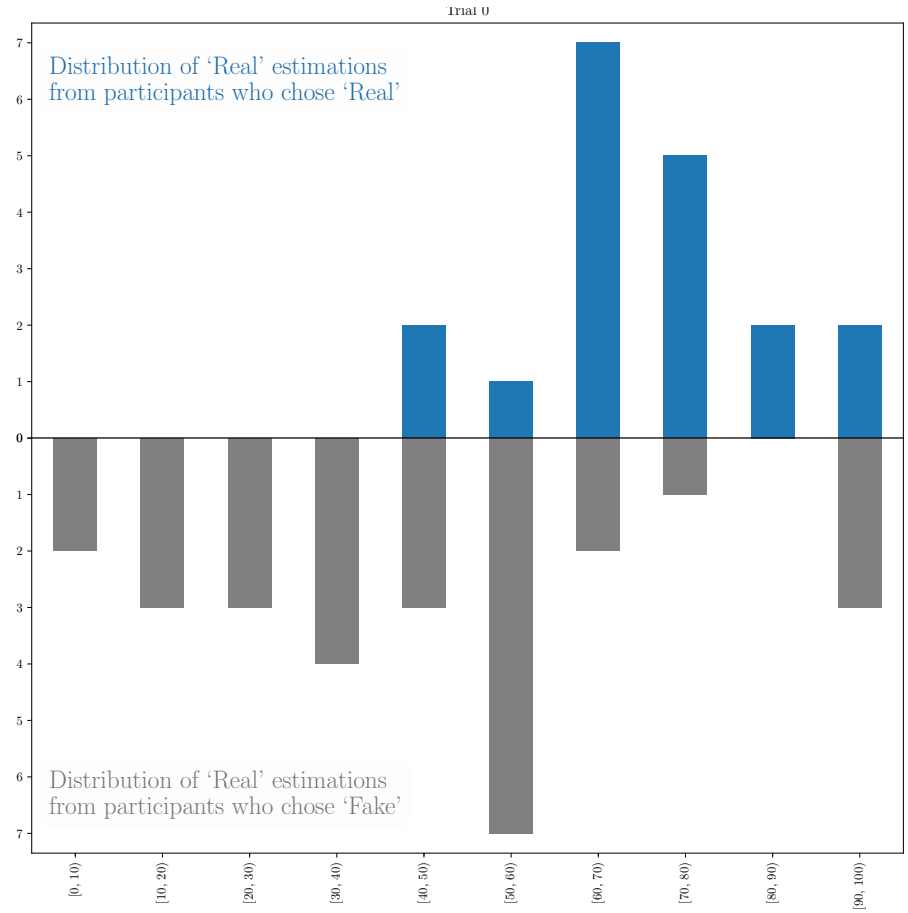


Figure 4.5: Estimation histogram for Trial 0. The histograms has been divided into two partitions, depending on the actual vote of participants.

4.1.6 Surprisingly Popular and Estimation Scores

The intuition behind the SP voting mechanism is that either the 'experts' (those who know the true answer to the 'real'/'fake' question) are in the majority, or if they are a minority, they know it themselves– in the sense that they estimate that a majority of the population would choose the opposite of what they have chosen.

If neither of these conditions are met, the mechanism fails to detect the true answer. The SP tries to assign as much weight as possible to the *experts* in the population, even when they are in a minority. As such, the failure of SP signals the lack of expert opinion in the population.

Figure **??** shows histograms for all trials. SP fails in trials in trials 0, 1, 3, 5 and 8. In Trial 0, the 'experts' are in minority but they are seemingly unaware of this, since they still estimate that a majority will favor their position. Similarly in Trial 1, the experts (those who have chosen 'fake') estimate that most of the population would also think that the news is fake. The analysis for trials 5 and 8 are similar.

In Trial 3, the experts seem to be in a majority, but their majority is not strong enough to compensate for the fact they too are over-estimating the population. However the difference is very small and with some minor filtering of the noisy data, such a situation can be avoided.

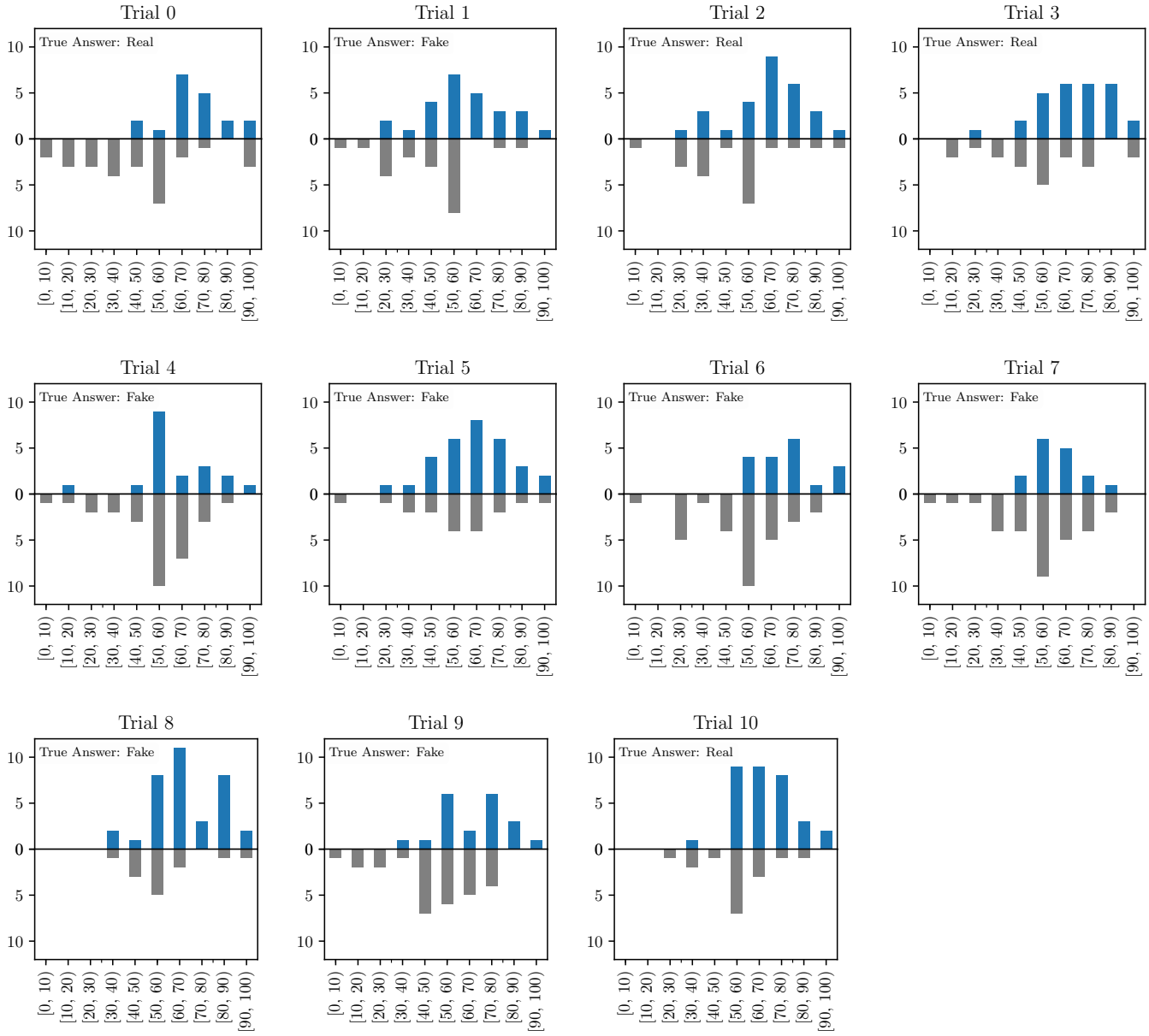


Figure 4.6: Estimation histograms for all trials. For the SP mechanism to work, 'those who know' (who choose the right answer) should either be in the majority themselves or else estimate that 'those who don't know are in the majority'. Without any filtering of the data, SP fails in trials 0, 1, 3, 5 and 8 because neither of these conditions hold. The data is too noisy, as respondents have not taken the time to assess others' potential response.