Evaluating the Effect of Emotion on Confirmation Bias in Online News Consumption

Sarah Halpin

Master of Science
School of Informatics
University of Edinburgh

2018

Abstract

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sarah Halpin*)

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

A tendency to seek out or interpret evidence to confirm one's own beliefs is a commonly-observed phenomenon known as confirmation bias [1]. This tendency can be seen online in many forms, one example of which is the emergence of online communities with polarised political beliefs [2]. These are often referred to as 'echo chambers', and they arise due to a desire to surround oneself with confirming information. This is known as acquisition bias, as it occurs in the process of seeking evidence. A second type of confirmation bias exists called integration in which evidence which threatens or disputes a person's beliefs is dismissed more readily than confirming evidence. This form of confirmation bias is important in the context of online news, as incidences of science denial, conspiracy theories and discrediting of media outlets gain traction [3, 4].

This project is based upon earlier work completed as part of a Master's thesis, [5], in which an experiment was conducted to evaluate confirmation bias in the acquisition and integration of online news. That thesis was inspired by the surprise factor in recent political outcomes, such as the election of Donald Trump. In these events, echo chambers and the polarisation of online communities were often cited as a factor in the results themselves as well as the mainstream media's apparent lack of awareness of public opinion [6]. The project investigated this idea and found a tendency in online users to preferentially seek out evidence confirming their own positions on various news topics.

This thesis aims to measure confirmation bias in the integration of news by comparing subjects' behaviour in an online experiment against a Bayesian model of updating which represents an unbiased agent. Understanding how people are susceptible to biases while reading online news may allow us to make choices to overcome and mitigate bias, leading to a more accurate view of current events and public opinion. This would be a positive move towards reducing the polarisation of communities online and increasing the population's ability or willingness to discern credible information from so-called 'fake news'.

A particular focus of this thesis is on the effect of emotion on confirmation bias. The theory of motivated reasoning suggests that emotion plays a role in many cognitive biases, including confirmation bias [7]. In particular, it suggests that biased reasoning occurs to minimise negative outcomes and maximise positive outcomes for the agent. Fear, anxiety and optimism in particular have been identified as factors in biased reasoning. Understanding the extent to which emotions and traits such as anxiety and fear affect biases is of benefit to the general public when considering how they interpret and interact with online news.

1.2 Project Objectives

This project has two primary objectives:

1. To model confirmation bias in the integration of information obtained online.
2. To identify emotional factors affecting confirmation bias.

To achieve the first, we will design a Bayesian model of updating to apply to data obtained from the experiment completed by Dalgarno [5], as well as data obtained in a new version of the experiment. We will compare shifts in opinion during online media consumption, observed in these experiments, to an unbiased Bayesian update. We will then analyse the degree to which observed data diverges from this unbiased update to determine the effect of confirmation bias.

The second part will be achieved by including measures of emotion in the new version of the experiment. We will examine whether associations with fear or hope lead to more biased reasoning by analysing measures of hope and fear against the confirmation measured in the first task. We will also investigate anxiety, optimism and emotionality as traits and their effect on the bias by measuring these traits using questionnaires and analysing the results against the degree of confirmation observed in the online browsing experiment.

1.3 Outline of Thesis

In section 2 we will provide a review of the background literature upon which this project is based, particularly known sources of confirmation bias, Bayesian models of updating and the effect of emotion on biases. In Section 3, we provide an outline of the experiment, describing the development and adaptation of a previous experiment to accommodate new research objectives, as well as summarizing the implementation . In section 4, we will describe the model we designed, outlining the development of the model, providing a mathematical description, and outlining how the model will be applied to experimental data. In Section 5, we will present the results and analysis of our experimental data, focusing on the two primary objectives outlined in Section 1.2. Finally, Section 6 will contain our conclusions, a summary of limitations of our work and suggestions for further study.

Chapter 2

Background

2.1 Confirmation Bias

Confirmation bias can be described as the tendency to acquire or interpret information in a manner which confirms one's preexisting beliefs. This phenomenon has been observed in a wide range of areas, and has been studied for its effects in economics, psychology, sociology and politics, among other fields. The effect of confirmation bias has also been studied in the context of scientific research, with Mynatt, Doherty, and Tweney [8] noting that researchers tend to avoid testing alternative hypotheses after having observed evidence which confirms their original hypothesis.

A number of studies have noted that confirmation bias can manifest itself at multiple points in the decision-making process. This leads to a breakdown of the bias multiple components including acquisition and integration [9, 10]. Acquisition refers to a bias in the seeking of information, avoiding information which threatens one's current beliefs. Integration refers to bias in the updating of one's views in light of evidence. Confirmation bias in integration results in evidence for one's prior belief having a stronger effect than opposing evidence in later beliefs.

Our experiment is concerned with bias in integration. We will present a series of statements to participants and measure the degree to which they agree or disagree with it. This strength of belief will then be measured again after observing evidence for and against the statement. Our project aims to model how this evidence is assimilated into the participant's current beliefs and whether this process occurs in a biased way.

2.2 Computational Models of Bias

2.2.1 Bayesian Models

Research has shown that many elements of human cognition can be represented by a Bayesian model [11]. These models determine how prior beliefs and new information about a given hypothesis should be combined, and a new belief formed. Here one's beliefs are represented by prior and posterior distributions over all possible outcomes.

Bias can be measured within a Bayesian model in two ways: by comparing observed behaviour in the given task with an optimal unbiased Bayesian updater, or by fitting a Bayesian model to data, by including some parameter which represents the bias of interest. Gerber and Green [12] describes a similar two models of updating: an optimal, unbiased update, and a second model which adds a parameter to determine partisan bias in the belief update. Both of these models will be used to evaluate confirmation bias for this thesis, and they are explained in more detail in Chapter 4.

2.2.1.1 Bayesian Update

The beta distribution is a conjugate prior for binomial distributions, meaning that the resulting posterior is also a beta distribution. As our experiment presents a series of headlines, each with one of two possible outcomes, the resulting likelihood is a binomial distribution, and so our prior should be modelled using a beta distribution. The beta distribution for the prior can be defined as follows:

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \alpha,\beta > 0, \tag{2.1}$$

where $\Gamma(.)$ is the Gamma function. This ratio of Gamma functions serves as a normalisation constant to ensure the distribution integrates to 1.

The binomial likelihood is represented as:

$$P(E|\theta) = \theta^{n_t}(1-\theta)^{N-n_t}, \tag{2.2}$$

where E is the set of evidence observed, n_t is the number of headlines which support the hypothesis being true, and N is the total number of headlines observed.

Bayes' Theorem can then be used to calculate a posterior distribution. We will ignore the normalisation constant in this step.

$$P(\theta|E) = \frac{P(\theta)P(E|\theta)}{P(E)} \tag{2.3}$$

$$\propto P(\theta)P(E|\theta), \tag{2.4}$$

where $P(E)$ is independent of θ, and thus we treat it as a constant.Then, substituting expressions for the prior and likelihood:

$$P(\theta|E) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{n_t}(1-\theta)^{N-n_t} \tag{2.5}$$

$$\propto \theta^{\alpha+n_t-1}(1-\theta)^{\beta+N-n_t-1} \tag{2.6}$$

This is simply a Beta distribution $Beta(\alpha', \beta')$ with $\alpha' = \alpha + n_t$ and $\beta' = \beta + N - n_t$.

2.2.2 Alternative Models

Doll, Hutchison, and Frank [13] found that certain genes associated with dopamine function are predictive of confirmation-biased learning. This link indicates that confirmation bias may arise through reinforcement, as dopamine plays the role of a reward in reinforcement learning.

A study by Frank, Seeberger, and O'reilly [14] of learning in patients with Parkinson's disease found that dopamine appears to affect learning which favours positive outcomes, rather than avoiding negative outcomes. In confirmation bias, both of these motives are likely to play a role in the update of beliefs, however a reinforcement learning model should focus on positive reinforcement in particular to reflect the role of dopamine in learning.

Using this idea, it may be possible to model confirmation bias in experiments using a reinforcement learning model, in which evidence for a preferred hypothesis comes with reward, while evidence against the preferred hypothesis comes with zero reward.

A typical reinforcement learning model updates beliefs in response to prediction error. In this scheme, a value V represents an expected value of reward from a given action. The discrepancy between the reward received r_i, and the estimate V is then used to update the value:

$$V_{i+1} = V_i + \rho(r_i - V_i), \tag{2.7}$$

where ρ is the learning rate.

However, our formulation does not include prediction error in this form. We wish to estimate a value V that is not a direct estimate of reward. In each update, the degree to which this value is retained is given by parameter α, $0 \leq \alpha \leq 1$. The reward r_i is then added to the updated value with reward sensitivity ρ. After observation i, the value is updated as follows:

$$V_{i+1} = \alpha V_i + \rho r_i, \tag{2.8}$$

where $r_i = 1$ if the evidence favours the preferred hypothesis and 0 otherwise. This is equivalent to Eq. 2.7 if $\alpha = 1 - \rho$. This version allows an independent parameter α to determine memory, similar to the measure of skepticism in the Bayesian model.

The final value can be converted to a probability or belief strength by applying a logistic sigmoid function, hence restraining it to the interval [0,1]. This model differs from a Bayesian model, in which the order of presentation of evidence is not taken into account, while this model allows for a recency bias.

2.3 Factors affecting Confirmation Bias

2.3.1 Emotion

The theory of motivated reasoning [15] suggests that biases driven by emotion exist to mitigate cognitive dissonance and maintain consistency in an individual's world view. Emotion-driven cognition and decision-making can also be referred to as 'hot cognition' [16, 17]. This may lead people to maintain faulty beliefs even after observing strong evidence against them, prioritising consistency and positive outcomes over accuracy. Two emotions emerge as important in goal-oriented reasoning, fear and hope, with fear serving to push perceptions further from undesirable or threatening outcomes, and hope of desirable outcomes strengthening belief further in that direction. It is important to note that these both represent bias towards positive outcomes, though not necessarily a confirmatory bias. However, this may influence confirmation bias, meaning confirmation should be stronger in the case of positive beliefs.

Westen, Blagov, Harenski, Kilts, and Hamann [7] found that decision-making on political issues is influenced by emotion, with fMRI results showing that a reasoning task to choose between political candidates engaged parts of the brain associated with implicit emotion regulation. Other studies have noted a biased effect in political decision making due to partisanship or some existing sentiment towards a candidate or organisation, consistent with motivated reasoning theory. These results are particularly interesting in the context of the proposed experiment of this thesis, which examines behaviours around online news, where content is often political in nature.

2.3.2 Identity

Identification with certain social and political groups and ideologies leads to behaviours which maintain consistency with that group or ideology. This can be explained by the notion of internal consistency in motivated reasoning: if someone doesn't believe in an idea central to a group to which they belong, their beliefs are at odds with their identity. This effect is described in Kahan [18], in which individuals with particular political affiliations were likely to maintain beliefs about climate change which reflected those of their political ideology in spite of contrary evidence.

People also tend to favour outcomes which may have a direct impact on them personally, or when they can identify with the group or individual that it affects. This is described by Hart and Nisbet [19], who suggest that social identity with a victim or victims leads to a stronger support for policies or opinions which positively affect the victim(s). While identification with a group lends itself strongly to confirmation bias, identification with individuals or groups who may be affected by a given outcome leads to both confirmatory and disconfirmatory biases, depending on the individual's prior opinion. This is shown in Dawson, Gilovich, and Regan [20], in which people who were presented with personally threatening outcomes displayed a disconfirmatory bias. We would expect that bias in either case is associated with emotion as described in the previous section, and that emotion is also linked with the personal nature of a given outcome.

2.3.3 Anxiety and Optimism

Beyond the emotions induced by a given story or topic, it is possible that certain personality traits make one more susceptible to biased reasoning and decision making. Despite motivated reasoning suggesting that, in general, belief in favoured outcomes is preferred and maintained, individuals who suffer with anxiety disorders can show the opposite behaviour [21]. This is usually explained as an attentional bias, in which a belief is influenced by the thoughts an individual is experiencing [22]. Patients with anxiety disorders have frequent negative and fearful thoughts [23], and presentation of outcomes which reflect their fears may lead to increased attention to threatening outcomes relative to positive or neutral ones [24]. Thus, while the general population may be more biased towards positive beliefs, high anxiety individuals may display a stronger bias towards negative or threatening beliefs. We wish to test this hypothesis by analysing anxiety against confirmation bias for both positive and negative beliefs, while also establishing whether confirmation bias over all topics is significantly different in high anxiety individuals than the general population.

Optimists are also susceptible to biased learning. As explained in section 2.2.1, there is a natural tendency for behaviour to reflect a preference towards positive outcomes. This effect is linked to the optimism of the individual, with optimists having a stronger tendency towards positive outcomes or events [25]. This phenomenon has been studied in the context of the stock market in particular, to understand how an optimistic confirmation biased trader fuels the creation of ‘bubbles’, while pessimistic bias pushes prices down [26]. A similar study by Park, Konana, Gu, Kumar, and Raghunathan [27] found a correlation between confirmation bias and overconfidence or optimism of stock traders’ performances. We wish to examine whether an individual’s optimism correlates with confirmation bias for positive prior beliefs, consistent with these findings.

Dopamine in the brain has been linked to both anxiety [28, 29] and positive affect and extraversion, which are linked to optimism [30]. This suggests if confirmation bias occurs through some reinforcement mechanism using dopamine as a reward system, optimism and anxiety may be correlated with bias. Dopamine has been linked in particular with optimism bias and motivated reasoning, which would lead confirmation to be stronger for positive prior beliefs [31]. We will assess differences between positive and negative updates under a reinforcement learning model of bias in order to test this hypothesis.

2.4 Online News

The claim that online news has led to the rise in echo-chambers and filter bubbles on social media is only partially true. While research suggests that online news has led to a polarisation of views, it also indicates that individuals now have more exposure to opposing views online than in print media [32]. Therefore, a natural browsing environment is likely to present a variety of positions on current events. Furthermore, Tewksbury [33] found that online news provides more autonomy to the reader, however this leads to a lower engagement with public affairs stories compared with print media. We therefore aim to present a simulated online browsing environment, which provides a range of views on each topic presented, and allows readers to choose the news sources they browse.

2.5 Summary of Research Questions

We will outline three models of confirmation bias and apply these models to experimental data in order to deduce whether confirmation bias exists in the online environment that our experiment simulates. We will carry out a model comparison to assess each model's suitability in measuring confirmation bias.

When confirmation bias has been measured, we wish to investigate underlying dimensions of the bias. We will carry analysis out to determine if there exist correlations between confirmation bias and:

- the emotions of fear and hope.
- the relevance of a statement to the individual.
- the personality traits of optimism and anxiety.

Since prior research has noted a distinction in how people process positive and negative information, we also wish to carry out separate analysis for anxiety and optimism in which prior beliefs are segregated into positive and negative.

Chapter 3

Experimental Methods

3.1 Previous Experiment

An experiment was conducted by Dalgarno [5] to collect data in order to measure confirmation bias in acquisition and integration for an online news browsing task. The experimental setup was as follows. There were two parts to the experiment: an initial browsing task and a series of surveys on personality traits.

3.1.1 Part 1

In this section, participants were presented with ten separate trials, in which they were asked to indicate their degree of belief in a statement related to news or current events, followed by a screen in which they were allowed to freely browse news headlines by three news sources related to the statement, until they reached a quota of fifteen headlines. The news sources were designed such that one source, Alpha News presented headlines which were always in agreement with the subjects reported belief, while Premier News always presented opposing headlines. A third source, First News, displayed headlines for both positions.

This was followed by a second statement screen where participants again their belief in the statement after observing evidence. [Appendix B will contain figures which show the user interface for this section.-doesn't yet]. Belief strength was rated on a slider scale. After all ten trials, subjects were asked to rate the perceived bias of each news source.

3.1.2 Part 2

Part 2 of this experiment consisted of four personality surveys, measuring 4 different personality traits which were deemed likely to have an interaction with confirmation bias:

- Depression: The Personality Health Questionnaire 9 (PHQ-9) [34].
- Schizotypy: The Schizotypal Personality Questionnaire (SPQ-B) [35].
- Mania: The Altman Self-Rating Mania Scale (ASRM) [36].
- Optimism: The Life Orientation Test - Revised (LOT-R) [37].

3.1.3 Analysis

Analysis of the data from this experiment aimed first to investigate acquisition bias by comparing the engagement for each of the three news sources. For this, the design choice was made to randomise the order of news sources presented to eliminate bias due to ordering. It was found that there was a slight but significant preference towards Alpha News, consistent with the hypothesis that information is sought to confirm existing beliefs.

Secondly, a simple scheme was used to identify one particular case of confirmatory updates, in which an individual's strength of belief in their preferred hypothesis increased in spite of a majority of observed evidence opposing this hypothesis. The tendency to update in this confirmatory manner was compared with personality traits, although no significant relationships were found.

3.1.4 Limitations

The experiment was still well suited to the principle aims of this project. However a number of potential issues were considered in deciding how to adapt the experiment.

The existing experiment had no measurement of confidence to estimate the variance of the prior.

Due to the analysis of acquisition bias, it was required to shuffle the order of news sources to eliminate order-effect bias [38]. However, this may have caused some confusion to participants, and slowed the development of associations between news sources and headlines.

In the case where a person' opinion had not changed at all, the continuous nature of a slider meant that a small change may occur due to human error, while the intention was to provide the same response.

3.2 Experimental Development

3.2.1 Changes to Methodology

We used the experiment outlined in Section 3.1 as a template for our new experiment. However, we wished to address some of the limitations listed in Section 3.1.4.

We introduced two questions to aid with estimating the width of the prior distribution:

- 'How confident do you feel about the strength of your belief?'
- 'How much knowledge do you already have on the topic of this statement?'

It is important to note that these questions will not absolutely determine variance for the prior, but are indicative of relative width.

As acquisition bias is not the focus of this project, the news sources were not shuffled in the new experiment, to simplify the interface and increase the chance of observing associations between headlines and news sources. Headlines were displayed throughout the browsing period to create a more realistic online browsing experience. We also removed the question about perceived bias of each news source as it was not relevant to our research question.

Finally, the number of statements was increased from ten to sixteen, due to a less limited time constraint on the new experiment. The statements were chosen to have approximate balance between positive and negative sentiment, and all statements relate to either important current events (for example Donald Trump), as well as more long-term political issues (U.S. healthcare system). The statements were applicable globally, but many were centered on political issues particularly relevant in the United States. Existing research associates confirmation bias with political partisanship, and an effort was made to include a number of issues with strong partisan gaps, including global warming, healthcare, education and the class divide [39].

3.2.2 Additions for New Research Objectives

Furthermore, a number of changes were made in order to collect information related to emotion, anxiety and optimism, to allow for the new objectives of this project.

In part 1 of the experiment a number of questions were posed to determine the degree to which participants identified and related emotionally to statements:

- 'To what extent does this topic make you feel emotional?'
- 'To what extent is the topic in this statement relevant or important to you?'
- 'When considering this statement, to what extent did you feel hopeful?'
- 'When considering this statement, to what extent did you feel fearful?'

In part 2 of this experiment, we wished to measure trait anxiety and optimism levels, and decided on a third measure to assess moods more generally. Thus, Part 2 was changed to include only three psychological questionnaires:

- Anxiety: Generalized Anxiety Disorder-7 (GAD-7) [40].
- Optimism: Life Orientation Test - Revised (LOT-R) [37].
- Mood/Emotionality: Positive and Negative Affect Schedule (PANAS) [41].

3.2.3 Pilot Experiment

A pilot experiment was conducted with four participating students from the University of Edinburgh. Feedback was collected after completion of the experiment. Data from the pilot experiment was not used in any analysis.

A number of students noted that other headlines disappeared after clicking on a given news source to see more headlines. This felt unrealistic, and meant that they clicked on blank news sources at random, and not based on headlines. A design change was made to maintain all headlines throughout the browsing stage for each statement.

One participant made an association between 'First News' and the existing news site 'Fox News', due to a similarity in names and logos. The three logos from the 2017

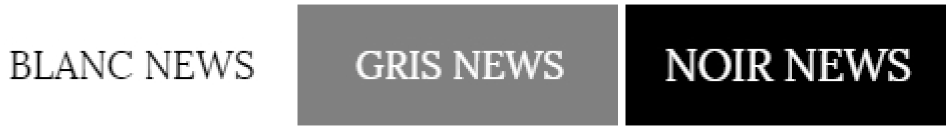


Figure 3.1: News Logos.

experiment and pilot are shown in Appendix B. The decision was made to simplify the logos and neutralise the colours to minimise associations with existing news sources. The names were chosen to match the colours, in order to limit sources of bias. The new names are also more easily distinguished than ‘First’, ‘Premier’ and ‘Alpha’, which are synonymous. The new logos are shown in Figure 3.1.

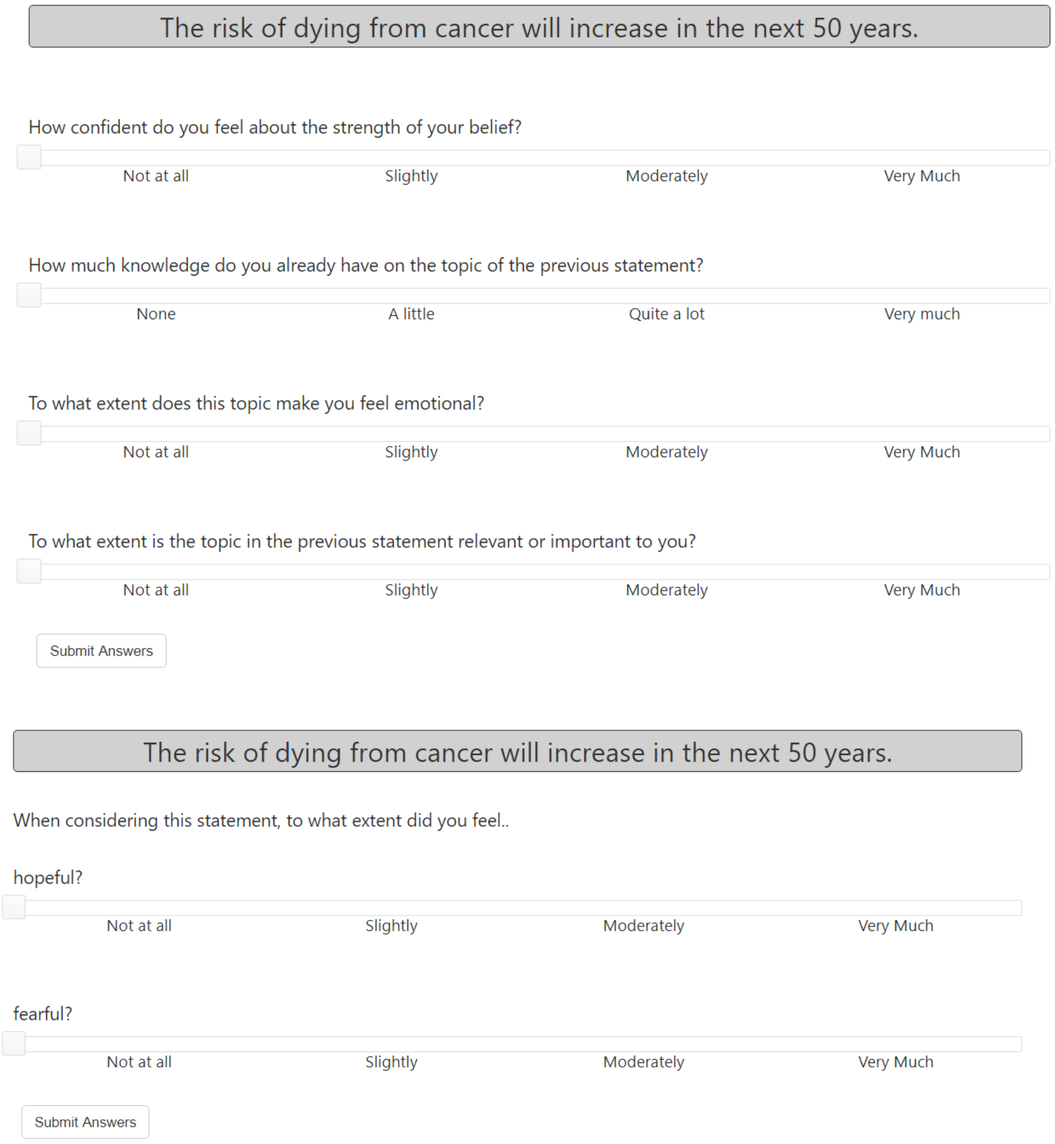
Finally, most participants noted a general tendency to ignore or not notice the news sources. In addition to simplifying the logos and names, we added an instruction on the headline page (‘Take some time to browse the following news stories’), to encourage greater concentration during browsing.

3.3 Experiment Description

Below I will outline the experiment as it was distributed to participants via Amazon Mechanical Turk. Before commencing this experiment, a School of Informatics Ethics form was completed to Level 1. A consent form was displayed to participants prior to starting the experiment; this is included in Appendix A.

3.3.1 Part 1: Statements and Responses

This part consisted of 16 separate exercises. In each, the participant was presented with a statements and asked to rate the degree to which they agreed with the statement on a slider scale from ‘Completely Disagree’ to ‘Completely Agree’. Next, the participant was asked a series of questions to determine confidence in their belief, emotional response, and prior knowledge of the topic. The screens including these questions are shown in Figure 3.2. This was followed by a screen in which participants were asked to browse headlines at their leisure up to a quota of 15 headlines. This screen is shown in Figure ??.



The risk of dying from cancer will increase in the next 50 years.

How confident do you feel about the strength of your belief?

Not at all   Slightly   Moderately   Very Much

How much knowledge do you already have on the topic of the previous statement?

None   A little   Quite a lot   Very much

To what extent does this topic make you feel emotional?

Not at all   Slightly   Moderately   Very Much

To what extent is the topic in the previous statement relevant or important to you?

Not at all   Slightly   Moderately   Very Much

The risk of dying from cancer will increase in the next 50 years.

When considering this statement, to what extent did you feel..

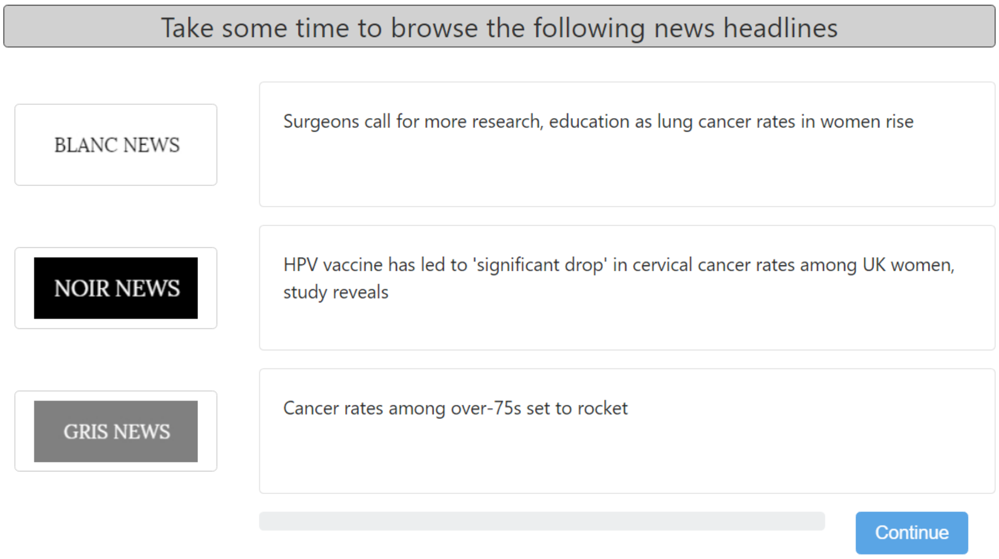
hopeful?

Not at all   Slightly   Moderately   Very Much

fearful?

Not at all   Slightly   Moderately   Very Much

Figure 3.2: Screens measuring personal and emotional response to statement.



Take some time to browse the following news headlines

Surgeons call for more research, education as lung cancer rates in women rise

HPV vaccine has led to 'significant drop' in cervical cancer rates among UK women, study reveals

Cancer rates among over-75s set to rocket

Figure 3.3: Illustrative example of the headline browsing screen.

3.3.2 Part 2: Evaluating Personality Traits

The participants were finally presented with three separate questionnaires, as specified in Section 3.2.2: GAD-7, LOT-R and PANAS. These were all assessed on a Likert scale rating system, with designated scoring systems as indicated by their accompanying literature [40, 37, 41].

3.4 Implementation

The 2017 experiment used the JSPsych library to create the experiment in JavaScript (JS) for use online. This experiment built upon the existing JS code. Headlines were sourced manually from a variety of news sources using Google News[1]. [I will direct to appendix with list of news sources used with URLs, as a greater variety was used than in Alasdair's project]

The link[2] to the experiment was made available on Amazon's Mechanical Turk[3]. MTurk allows requesters to impose specific criteria on participants. Bearing in mind that the statements were US-centric, it was chosen to restrict participation to residents

[1] https://news.google.com/

[2] https://groups.inf.ed.ac.uk/online_questions/

[3] https://www.mturk.com/

of the United States. Each worker on MTurk is also given an approval rating (0-100%), and requesters may filter using this criteria. While there is no recommendation for a minimum approval rating, some studies use 95% as a cut-off [42, 43], thus we chose a 95% minimum approval rating.

Chapter 4

Model Descriptions

4.1 Model Development

In this chapter I will outline the three models that will be used to measure the effect of confirmation bias. I will formally describe each model, as well as any assumptions made when applying the given model to experimental data. I will briefly list the limitations of each model. The first model represents ideal behaviour, to which experimental data will be compared, while Models 2 and 3 will be fit directly to experimental data, with particular model parameters measuring confirmation bias.

4.1.1 Model 1: Unbiased Bayesian Model

This model aims to represent an unbiased Bayesian updater. The experimental data will consist of an initial belief strength P_i, a set of observed evidence E and a final belief strength P_f for each statement.

This model will fit a prior distribution of belief strengths using the data for P_i. Then an update will be calculated by applying Bayes Theorem to this prior as outlined in Section 2.2.1, giving an 'ideal' posterior distribution. Confirmation bias will be measured by comparing observed experimental data P_f with this unbiased posterior estimate, to test whether participant's beliefs conform to the update rule defined by this model. I will outline below how the prior distribution is estimated and how the posterior is calculated from observed evidence.

The Beta distribution is represented by two parameters (α, β). In data from the experiment conducted in 2017 [5] our prior belief is currently represented solely by a single belief strength. We will interpret this as the mode or highest point of the distribution. In order to estimate the parameters to define a distribution, we must make some assumptions about the variance of the prior distribution. For the older experimental data, we will choose the variance such that near $P_i = 0.5$, the distribution is completely flat, and condition the variance on the distance from a neutral prior, with the assumption that more strongly held beliefs come with greater confidence in that belief strength.

The experimental data from the new experiment includes ratings of confidence in one's belief strength and the amount of prior knowledge one has on the given topic. Thus, these quantities will be used instead of belief strength to determine the relative width of the prior in new data, with greater confidence and knowledge contributing to a narrow prior. We still wish to ensure a flat prior for belief strengths near $P_i = 0.5$.

We can thus use estimates of the mode and variance of the distribution in order to determine values for α and β to represent the prior distribution:

$$Mode = \frac{\alpha - 1}{\alpha + \beta - 2}, \qquad \alpha, \beta > 1$$

$$Var = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

From these definitions it is possible to calculate α and β analytically, however it is more convenient to do so numerically.

We have now estimated a Beta prior from the experimental data, and the likelihood is calculated simply from the number of headlines observed in support of the hypothesis, n_t and the total number of headlines observes=d, which in our experiment is constant at 15. The resulting posterior distribution is:

$$P(\theta|E) = Beta(\alpha + n_t, \beta + N - n_t).$$

Maximum a Posteriori estimation chooses the probability at which this posterior is maximal, i.e. the mode, so we use the following to represent the response of an ideal, unbiased agent after viewing evidence:

$$\text{Final Response} = \frac{\alpha + n_t - 1}{\alpha + \beta + N - 2}.$$

Finally, the degree to which a response was confirmatory was approximated by the difference between the response observed and the response predicted by the unbiased model. If we denote the initial statement response as R_i, the final response as R_f, and the model's estimated final response as R_M, then confirmation C is given by:

$$C = \begin{cases} (R_f - R_M), & \text{if } R_i > 0.5. \\ (R_M - R_f), & \text{if } R_i < 0.5. \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

The quantity C will then be used to analyse the degree to which individuals display a confirmatory bias in integration.

4.1.1.1 Advantages and Limitations

This model is simple to fit and apply to experimental data, and gives an approximate indication of how biased an individual or population is. It also has the advantage that it can be independently applied to each question for each individual, where later models will only be able to present measures of bias at the level of the individual (representing all questions) or the question (representing all individuals).

However, this model suffers from a number of limitations. The first is that, for data from the 2017 experiment, the variance of each prior is unknown, and is estimated based on the assumption that stronger beliefs always lead to a narrow prior while less strong beliefs are represented by a flat prior. However, the experiment conducted for this year's experiment also collected information on the person's prior knowledge and confidence, in order to obtain a more accurate relative estimate of variance. The assumption of a flat neutral prior remains necessary to obtain absolute measures of variance.

Secondly, the likelihood in this model is defined by Equation 2.2, and this model updates each prior depending solely on the balance of confirming and opposing headlines. However, headlines may not be seen as absolute evidence, and each individual will have their own degree of skepticism when considering headlines as evidence. This skepticism is not taken into account here, so the estimated ideal response may represent an agent who puts too much weight on the value of headlines. The model described below aims to address this issue by fitting updates directly to the experimental data.

4.1.2 Model 2: Bayesian Model with Confirmation Bias

A second Bayesian model aims to directly fit experimental data by including parameters which represent biases in integration. In this model, we use the fact that headlines are generally interpreted as weak evidence that a hypothesis is true or false. In order to incorporate this into a Bayesian model similar to that described in the previous section, weights will be applied to evidence for and against the favoured hypothesis, which will then be fit using the experimental data. A third weight representing the overall strength of all evidence will also be fit.

This scheme is formally defined as follows. The prior distribution is estimated in the same way as described in Section 4.1.1, numerically fitting α and β from the equations for mode and variance of the Beta distribution.

We then wish to change how evidence is weighted in the likelihood. In Eq. 2.2, we had the following definition for the likelihood:

$$P(E|\theta) = \theta^{n_t}(1-\theta)^{N-n_t}, \tag{4.2}$$

with n_t and N directly representing the number of headlines indicating a true hypothesis and total number of headlines respectively. In a weighted model, we wish to create a representative strength of evidence for and against the hypothesis, which means scaling these quantities by weights yet to be fitted.

We wish to include two factors in these weights: the overall strength of evidence K, as well as two separate weights which indicate the relative strength given to confirmatory and disconfirmatory evidence. As the likelihood is framed in terms of headlines supporting and opposing the given hypothesis, confirmatory weights W_C and disconfirmatory weights W_D must be applied in two distinct cases, depending on the prior response R_i:

$$\begin{cases} P(E|\theta) = \theta^{KW_C n_t}(1-\theta)^{KW_D(N-n_t)} & \text{if } R_i > 0.5. \\ P(E|\theta) = \theta^{KW_D n_t}(1-\theta)^{KW_C(N-n_t)} & \text{if } R_i < 0.5. \end{cases} \tag{4.3}$$

The case of $R_i = 0.5$ is simply omitted from these updates, as completely neutral beliefs by definition cannot lead to a confirmatory or disconfirmatory bias.

The posterior mode is then:

$$\begin{cases} \frac{\alpha+KW_C n_t-1}{\alpha+\beta+K(W_C n_t+W_D(N-n_t))-2} & \text{if } R_i > 0.5. \\ \frac{\alpha+KW_D n_t-1}{\alpha+\beta+K(W_D n_t+W_C(N-n_t))-2} & \text{if } R_i < 0.5. \end{cases} \tag{4.4}$$

Confirmation bias can then be calculated by comparing the weights on confirmatory and disconfirmatory headlines, with significantly heavier weights on confirmatory headlines indicating the presence of confirmation bias in the population.

Advantages and Limitations

This model has the advantage of having included a parameter K which indicates skepticism or low strength of belief in headlines, leading to a more complete view of belief updates as a combination of unbiased skepticism and biased preference for confirmatory headlines.

However, the estimation of the variance of the prior is still based on a number of assumptions that may not hold true in general. The model does not take into account the temporal order of headlines, which may have some effect on the overall impression an individual takes from a set of headlines. A reinforcement learning model, described in Section 4.1.3, aims to address these limitations.

4.1.3 Model 3: Reinforcement Learning Model

A third model is based on the idea that bias towards confirmatory evidence is reward-based, and thus there is a reinforcement learning effect. In this model, we assume that the initial response corresponds to some value of truth of the hypothesis. As described in Section 2.2.2, this value can be updated after the i^{th} headline as follows:

$$V_{i+1} = \alpha V_i + \rho r_i, \tag{4.5}$$

This value is not a probability, but the probability or belief strength can be obtained from the final value V_f using a logistic sigmoid function:

$$p = \frac{1}{1+e^{\beta(V_f-t)}}, \tag{4.6}$$

where β is an inverse temperature parameter and t is a threshold. Temperature determines decisiveness of beliefs, with low temperatures leading to very strong beliefs in the preferred hypothesis, while high temperatures lead to belief strength which changes slowly with changes in V.

Confirmation bias is then indicated by a reward parameter ρ which significantly exceeds that expected by unbiased belief updates. The memory parameter α describes the degree to which old information is ignored in the update. Unbiased updates can be approximated by a reward parameter which causes the value to move equally in the direction of belief or disbelief when provided with a headline for or against a hypothesis.

Advantages and Limitations

This model includes the temporal ordering of headlines, therefore allowing for a memory mechanism which weights recently observed information more heavily in updates. Furthermore, this model does not require any assumptions about one's confidence in belief strength.

However, the reward parameter is not a direct measure of confirmation bias. In reality, confirmation bias may arise from a reward-based learning system, but likely has many influencing factors. This model may oversimplify the influence of these factors by reducing biased learning to a reward parameter.

4.1.4 Fitting the Models

Model 1 does not need to be fit to experimental data. Confirmation is measured by direct comparison of the data to the posterior estimate predicted by this model. A confirmation significantly greater than 0 on a 1 sample t-test indicates a significant confirmation bias effect in the group of participants.

The parameters of Model 2 and Model 3 will be fit to data using a least-square fit. The Python module `scipy.optimise` allows one to define the equations in terms of unknown parameters for the posterior belief strength defined in Eq. 4.4 for Model 2 and 4.6 for Model 3. The `least_square` function can fit these unknown parameters directly to the data for final belief strength by minimising the square error between the model estimate and the data, subject to any upper and lower bounds which may be imposed. [I will include more specific detail of what parameters were fit and what bounds were used but I need to finish some work on the RL model]

4.1.5 Model Comparison

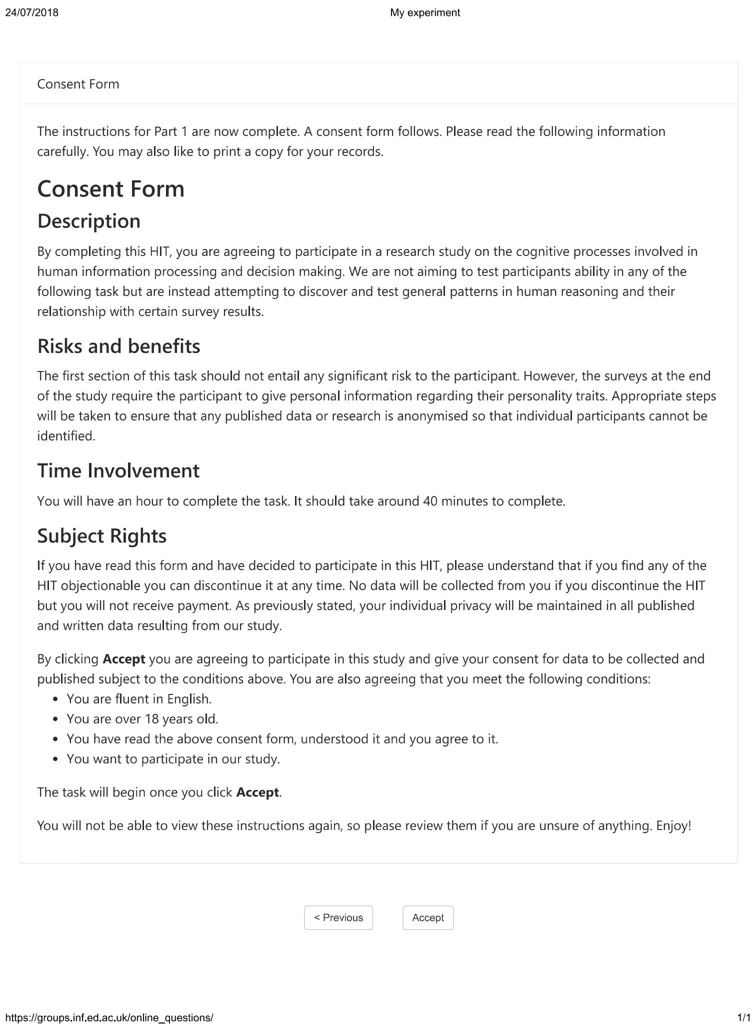
Beyond the qualitative limitations and advantages outlined in the model descriptions, it is necessary to carry out quantitative evaluations of each model.

First, the confirmation bias should have strong correlation between models. That is, the confirmation bias as measured for each individual under one model should correlate significantly with the same measure under each of the other two models. If this is not the case, then one or more measures are insufficient to define confirmation bias.

For Models 2 and 3, model selection tools such as Akaike's Information Criterion and Bayesian Information Criterion [44, 45] may be used to assess how well each model represents the experimental data. These statistical methods assess the likelihood of existing data under the given model, including penalty terms for the number of parameters, in order to balance model simplicity and accuracy.

Appendix A

Consent Form



24/07/2018 My experiment

Consent Form

The instructions for Part 1 are now complete. A consent form follows. Please read the following information carefully. You may also like to print a copy for your records.

Consent Form

Description

By completing this HIT, you are agreeing to participate in a research study on the cognitive processes involved in human information processing and decision making. We are not aiming to test participants ability in any of the following task but are instead attempting to discover and test general patterns in human reasoning and their relationship with certain survey results.

Risks and benefits

The first section of this task should not entail any significant risk to the participant. However, the surveys at the end of the study require the participant to give personal information regarding their personality traits. Appropriate steps will be taken to ensure that any published data or research is anonymised so that individual participants cannot be identified.

Time Involvement

You will have an hour to complete the task. It should take around 40 minutes to complete.

Subject Rights

If you have read this form and have decided to participate in this HIT, please understand that if you find any of the HIT objectionable you can discontinue it at any time. No data will be collected from you if you discontinue the HIT but you will not receive payment. As previously stated, your individual privacy will be maintained in all published and written data resulting from our study.

By clicking **Accept** you are agreeing to participate in this study and give your consent for data to be collected and published subject to the conditions above. You are also agreeing that you meet the following conditions:

- You are fluent in English.
- You are over 18 years old.
- You have read the above consent form, understood it and you agree to it.
- You want to participate in our study.

The task will begin once you click **Accept**.

You will not be able to view these instructions again, so please review them if you are unsure of anything. Enjoy!

https://groups.inf.ed.ac.uk/online_questions/ 1/1

Figure A.1: Consent Form

Appendix B

Supplementary Figures: 2017 Experiment

Bibliography

[1] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[2] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific reports*, 7:40391, 2017.

[3] Stephan Lewandowsky, Gilles E Gignac, and Klaus Oberauer. The role of conspiracist ideation and worldviews in predicting rejection of science. *PloS one*, 8 (10):e75637, 2013.

[4] Jason Turcotte, Chance York, Jacob Irving, Rosanne M Scholl, and Raymond J Pingree. News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, 20(5):520–535, 2015.

[5] Alasdair Dalgarno. Measuring the confirmation bias using a novel online experiment. Master's thesis, The University of Edinburgh, 2017.

[6] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[7] Drew Westen, Pavel S. Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann. Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 u.s. presidential election. *Journal of Cognitive Neuroscience*, 18(11):1947–1958, 2006. doi: 10.1162/jocn.2006.18.11.1947. URL `https://doi.org/10.1162/jocn.2006.18.11.1947`.

[8] Clifford R Mynatt, Michael E Doherty, and Ryan D Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The quarterly journal of experimental psychology*, 29(1):85–95, 1977.

[9] Martin Jones and Robert Sugden. Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1):59–99, 2001.

[10] Steve D Charman. The forensic confirmation bias: A problem of evidence integration, not just evidence evaluation. 2013.

[11] Robert A Jacobs and John K Kruschke. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011.

[12] Alan Gerber and Donald Green. Misperceptions about perceptual bias. *Annual Review of Political Science*, 2(1):189–210, 1999. doi: 10.1146/annurev.polisci.2.1.189. URL `https://doi.org/10.1146/annurev.polisci.2.1.189`.

[13] Bradley B Doll, Kent E Hutchison, and Michael J Frank. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *Journal of Neuroscience*, 31(16):6188–6198, 2011.

[14] Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, 306(5703): 1940–1943, 2004.

[15] Ziva Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3): 480, 1990.

[16] David P Redlawsk. Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, 64(4): 1021–1044, 2002.

[17] Milton Lodge and Charles S Taber. The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, 26(3):455–482, 2005.

[18] Dan M Kahan. Ideology, motivated reasoning, and cognitive reflection: An experimental study. 2012.

[19] P. Sol Hart and Erik C. Nisbet. Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, 39(6):701–723, 2012. doi: 10.1177/0093650211416646. URL `https://doi.org/10.1177/0093650211416646`.

[20] Erica Dawson, Thomas Gilovich, and Dennis T Regan. Motivated reasoning and performance on the was on selection task. *Personality and Social Psychology Bulletin*, 28(10):1379–1387, 2002.

[21] Joachim Stöber. Trait anxiety and pessimistic appraisal of risk and chance. *Personality and Individual Differences*, 22(4):465–476, 1997.

[22] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J Bakermans-Kranenburg, and Marinus H Van Ijzendoorn. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological bulletin*, 133(1):1, 2007.

[23] Peter J Lang and Lisa M McTeague. The anxiety disorder spectrum: Fear imagery, physiological reactivity, and differential diagnosis. *Anxiety, Stress, & Coping*, 22(1):5–25, 2009.

[24] Jenny Yiend and Andrew Mathews. Anxiety and attention to threatening pictures. *The Quarterly Journal of Experimental Psychology Section A*, 54(3):665–681, 2001. doi: 10.1080/713755991. URL `https://doi.org/10.1080/713755991`.

[25] Suzanne C Segerstrom. Optimism and attentional bias for negative and positive stimuli. *Personality and Social Psychology Bulletin*, 27(10):1334–1343, 2001.

[26] Sebastien Pouget, Julien Sauvagnat, and Stephane Villeneuve. A mind is a terrible thing to change: confirmatory bias in financial markets. *The Review of Financial Studies*, 30(6):2066–2109, 2017.

[27] JaeHong Park, Prabhudev Konana, Bin Gu, Alok Kumar, and Rajagopal Raghunathan. Confirmation bias, overconfidence, and investment performance: Evidence from stock message boards. 2010.

[28] Miguel Pérez de la Mora, Andrea Gallegos-Cari, Yexel Arizmendi-García, Daniel Marcellino, and Kjell Fuxe. Role of dopamine receptor mechanisms in the

amygdaloid modulation of fear and anxiety: structural and functional analysis. *Progress in neurobiology*, 90(2):198–216, 2010.

[29] Susanne Nikolaus, Christina Antke, Markus Beu, and Hans-Wilhelm Müller. Cortical gaba, striatal dopamine and midbrain serotonin as the key players in compulsive and anxiety disorders-results from in vivo imaging studies. *Reviews in the Neurosciences*, 21(2):119–140, 2010.

[30] Richard A Depue and Paul F Collins. Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and brain sciences*, 22(3):491–517, 1999.

[31] Tali Sharot, Marc Guitart-Masip, Christoph W Korn, Rumana Chowdhury, and Raymond J Dolan. How dopamine enhances an optimism bias in humans. *Current Biology*, 22(16):1477–1481, 2012.

[32] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

[33] David Tewksbury. What do americans really want to know? tracking the behavior of news readers on the internet. *Journal of communication*, 53(4):694–710, 2003.

[34] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9): 606–613, 2001.

[35] Alex S Cohen, Russell A Matthews, Gina M Najolia, and Laura A Brown. Toward a more psychometrically sound brief measure of schizotypal traits: introducing the spq-brief revised. *Journal of personality disorders*, 24(4):516–537, 2010.

[36] Edward G Altman, Donald Hedeker, James L Peterson, and John M Davis. The altman self-rating mania scale. *Biological psychiatry*, 42(10):948–955, 1997.

[37] Michael F Scheier, Charles S Carver, and Michael W Bridges. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation test. *Journal of personality and social psychology*, 67(6):1063, 1994.

[38] William D Perreault. Controlling order-effect bias. *The Public Opinion Quarterly*, 39(4):544–551, 1975.

[39] Frank Newport and Andrew Dugan. Partisan differences growing on a number of issues. *Gallup.com*, August 2017. https://news.gallup.com/opinion/polling-matters/215210/partisan-differences-growing-number-issues.aspx, Accessed 30/07/2018.

[40] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.

[41] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[42] David J Hauser and Norbert Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1):400–407, 2016.

[43] Carolyn Chen, Lee White, Timothy Kowalewski, Rajesh Aggarwal, Chris Lintott, Bryan Comstock, Katie Kuksenok, Cecilia Aragon, Daniel Holst, and Thomas Lendvay. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research*, 187(1):65–71, 2014.

[44] Hirotugu Akaike. Factor analysis and aic. In *Selected Papers of Hirotugu Akaike*, pages 371–386. Springer, 1987.

[45] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2): 261–304, 2004.