Exploring Social Metacognition: the role of confidence in updating estimates of advisor reliability with and without feedback.

Matt Jaquiery

Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Abstract

This *R Markdown* template is for writing an Oxford University thesis. The template is built using Yihui Xie's `bookdown` package, with heavy inspiration from Chester Ismay's `thesisdown`, and the `OxThesis` LaTeX template (most recently adapted by John McManigle).

This template's sample content include illustrations of how to do the various things you need to write a thesis in R Markdown, and largely follow the structure from this R Markdown workshop.

Congratulations for taking a step further into the lands of open, reproducible science, by writing your thesis using a tool that allows you to transparently include tables and dynamically generated plots directly from the underlying data. Hip hooray!

Exploring Social Metacognition: the role of confidence in updating estimates of advisor reliability with and without feedback.

Matt Jaquiery

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

For TBC

Acknowledgements

This is where you will normally thank your advisor, colleagues, family and friends, as well as funding and institutional support. In our case, we will give our praises to the people who developed the ideas and tools that allow us to push open science a little step forward by writing plain-text, transparent, and reproducible theses in R Markdown.

We must be grateful to John Gruber for inventing the original version of Markdown, to John MacFarlane for creating Pandoc (`http://pandoc.org`) which converts Markdown to a large number of output formats, and to Yihui Xie for creating `knitr` which introduced R Markdown as a way of embedding code in Markdown documents, and `bookdown` which added tools for technical and longer-form writing.

Special thanks to Chester Ismay, who created the `thesisdown` package that helped many a PhD student write their theses in R Markdown. And a very special tahnks to John McManigle, whose adaption of Sam Evans' adaptation of Keith Gillow's original maths template for writing an Oxford University DPhil thesis in LaTeX provided the template that I adapted for R Markdown.

Finally, profuse thanks to JJ Allaire, the founder and CEO of RStudio, and Hadley Wickham, the mastermind of the tidyverse without whom we'd all just given up and done data science in Python instead. Thanks for making data science easier, more accessible, and more fun for us all.

Ulrik Lyngs
Linacre College, Oxford
2 December 2018

Abstract

This *R Markdown* template is for writing an Oxford University thesis. The template is built using Yihui Xie's `bookdown` package, with heavy inspiration from Chester Ismay's `thesisdown`, and the `OxThesis` LaTeX template (most recently adapted by John McManigle).

This template's sample content include illustrations of how to do the various things you need to write a thesis in R Markdown, and largely follow the structure from this R Markdown workshop.

Congratulations for taking a step further into the lands of open, reproducible science, by writing your thesis using a tool that allows you to transparently include tables and dynamically generated plots directly from the underlying data. Hip hooray!

Contents

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring in this thesis to spatial dimensions in an image.

Otter . . . . . One of the finest of water mammals.

Hedgehog . . . Quite a nice prickly friend.

Part I

Introduction

1
Introduction

!TODO[Style notes: * Graphs: * Continuous axes which show theoretical limits should have expand = c(0, 0), axes which do not show those limits should have some expansion]

!TODO[update evo models to use this terminology]

$a_{i,t}$ - advice agent i receives at time t

$b_{i,t}$ - bias of agent i at time t

$c_{i,t}^{I/A/F}$ - agent i's confidence in their initial/advisory/final estimate at time t

$e_{i,t}^{I/A/F}$ - agent i's initial/advisory/final estimate at time t

i, j - agent identifiers

$s_{i,t}$ - sensitivity of agent s at time t

u_i - fitness/utility of agent i

v - true value in the world

λ - learning rate

$\omega_{i,t}$ - weighting for agent i's estimate at time t

This thesis attempts to determine whether individual psychological processes in the seeking and taking of advice are sufficient to produce entrenched biases at the network level. The work includes empirical investigation into the individual psychological processes underlying advice seeking and advice taking using behavioural experiments; computational modelling and behavioural experimentation exploring the effects of contextual factors on advice-taking; agent-based computational modelling of the effects interactions between agents with psychological advisor evaluation processes that produce biases in assimilation of information and source selection; and a comparison between the structures of networks produced in these models and those naturally occurring social networks. The organisation is as follows: this introduction establishes the core concepts invoked in this thesis, and describes their treatment in the literature; the following sections each include a short chapter on the specific question addressed and the techniques used, a detailed description of the work conducted, and a short discussion of the conclusions drawn from the work; and a final section offers broader conclusions arising as a consequence of the presented work, alongside some suggestions for related research.

Overview

The first line of behavioural experiments report investigates whether people decide which advisors to consult using internal signals where external feedback is unavailable, extending a previous finding concerning the influence of advice (Pescetelli and Yeung 2018) into the domain of advisor choice. While little support is found for the metacognitive mechanisms underpinning the influence of advice, the frequency of agreement is shown to be a good indicator of advisor choice. Next, the effects of advisor agreement are explored in terms of influence using a different decision-making domain; a date estimation rather than perceptual decision task. Again, people denied objective feedback on the quality of advice are shown to attend to agreement. The work above demonstrates how the influence of advice depends upon the reputation of an advisor built up over time. In a second line of experiments, reputations of advisors are established at the outset and the responses to different pieces of advice are explored. People consider both the properties of advice and the properties of advisors, placing less trust in advice which appears suspicious or which comes from a suspicious source. These behavioural results support implications from evolutionary models which demonstrate that the presence of even a few bad actors in a population can mean that distrusting all advice from all sources a little is adaptive. The evolutionary models are extended to show that this kind of general distrust, known as egocentric discounting, emerges as adaptive even where none of the agents deliberately mislead others, and even where all agents are equally skilled and fully cooperative. Finally, network effects arising from interconnected networks of agents are explored. Agents are based on the empirical results heretofore presented, and the consequences are explored for networks with various starting structures characterised by sparsity and homogeneity, under varying rates and reliabilities of external feedback. !TODO[results of this]. The resulting network characteristics are compared with network structures from real online networks from social media websites. !TODO[results of this].

Advice

Advice is broadly defined as information which comes from a social source. Advice is therefore different from other sources of information in that it is the result of (at least) mental processing of other information. In some cases, it may additionally include discussions among different group members (e.g. advice from the International Advisory Panel on Climate Change). Throughout this thesis, the focus is primarily on advice which comes from a single, stable source, as when we see a post by an acquaintance on social media, or when a stranger provides us with advice.

Advice-taking

Advice occurs in the context of a decision, and forms a part of the information which is integrated during the decision-making process to produce a decision. To the extent that the decision reached differs from the decision that would have occurred had the advice not been presented, the advice has had an effect on the decision; to the extent that this difference changes the decision in a way consistent with the advice, the advice has been 'taken' (as opposed to 'rejected').

It is tacitly implied by many operationalizations of advice-taking that the informational content of the advice determines the extent to which it is taken or rejected !TODO[CITE]. Insofar as the identity of the advisor matters, it matters because it functions as a cue to the informational content of the advice. This is likely a major oversimplification, however, because in many real-world contexts advice-giving and advice-taking form part of a developing social relationship: being consulted for advice and having one's advice followed are inherently rewarding (Hertz and Bahrami 2018; Hertz, Palminteri, et al. 2017); and taking advice can serve as a (sometimes costly) social signal of valuing a relationship with a person or group (Byrne et al. 2016). Furthermore, some authors have argued that people may perceive taking advice as sacrificing their independence or autonomy !TODO[CITATIONNEEDED]. While this thesis follows previous literature in omitting to consider the wider social concerns influencing the taking of advice, it is nevertheless important to remember that the processes investigated herein take place in a variety of social contexts where complex social agents attempt to optimise over numerous goals over numerous timescales.

Three-factor model of trust

The degree to which advice is taken is proportional to the trust placed in the advisor by the decision-maker. Interpersonal trust, or the degree to which one is prepared to place one's fortune in the hands of another (e.g. by relying on their advice), is apportioned by Mayer et al. (1995) onto three properties of the advisor (as judged by the decision-maker): ability, benevolence, and integrity. To these three properties of the advisor we may add the decision-maker's general propensity to trust, as well as situational cues and task cues (e.g. the phenomenon that advice is more readily taken for hard tasks than easy ones, (Gino and Moore 2007)).

Ability

Ability captures the expertise of an advisor: their raw ability to perform the task for which they are giving advice. In some cases this is relatively straightforward, as in the expertise of a General Practitioner in matters of health and disease, and in others more complex, as in the expertise of a hairdresser when deciding on a haircut (when matters of personal taste comingle with aesthetic considerations of facial structure, practical considerations of hair constitution, and social considerations of fashion). The greater the ability of an advisor, the greater the influence of their advice, as demonstrated by experiments showing that participants' decisions are more affected by the advice of advisors who are labelled as more expert in a relevant domain (Sah et al. 2013; Schultze et al. 2017; Sniezek, Schrah, et al. 2004; Sniezek and Van Swol 2001; Soll and Mannes 2011), or are shown to be more expert empirically (Pescetelli 2017; Sah et al. 2013; Ilan Yaniv and Kleinberger 2000).

Benevolence

Benevolence refers to the extent to which the advisor seeks to further the interests of the decision-maker. Where ability represents the absolute limit on the quality of advice, benevolence represents the extent to which the advice approaches this limit. The advice of even a renouned expert may be doubted if there is reason to believe their goal is to mislead, a vital lesson for medieval monarchs with their councils of politicking advisors. Experimental work has shown that psychology students relied more on the advice of their friends than on the advice of labelled experts !TODO[CITATIONNEEDED], and that participants are more inclined to reject advice when uncertainty is attributed to malice rather than ignorance (Schul and Peni 2015)}.

Integrity

Advisors with integrity exhibit adherence to principles which the decision-maker endorses. As with benevolence, integrity acts to determine the extent to which advice approaches the limit imposed by ability. While not mutually exclusive, integrity is typically important where relationships are less personal (e.g. we may place great trust in a General Practitioner because of their expertise in medical matters and their *integrity* in adhering to a set of professional ethical and conduct requirements). !TODO[Some description of the research]

Normative models of advice-taking

Advice-taking can be evaluated formally with reference to a normative model. The simplest and most common of these views the decision-making task as an estimation problem (or combination of estimation problems), and provides an approximately Bayesian variance-weighted integration of independent estimates. To borrow from Galton (1907)}, consider the task of judging the weight of a bullock. We can model any single guess (e) as the true weight (v) plus some error (ϵ):

$$e = v + \epsilon \tag{1.1}$$

The key insight is to observe that the error is drawn from a normal distribution ($\mathcal{N}(\mu = 0, \sigma^2)$)[1]. As the number of samples from this distribution increases, the mean of those samples tends towards the mean of the distribution. Thus, the more estimates are taken, the closer on average the sum of errors will be to 0.

$$\frac{\sum_i^N(e_i)}{N} = \frac{\sum_i^N(v + \mathcal{N}(\mu = 0, \sigma_i^2))}{N} \tag{1.2}$$

$$\frac{\sum_i^N(e_i)}{N} = \frac{\sum_i^N(v)}{N} + \frac{\sum_i^N(\mathcal{N}(\mu = 0, \sigma_i^2))}{N} \tag{1.3}$$

$$\frac{\sum_i^N(e_i)}{N} = \frac{Nv}{N} + \hat{0} \tag{1.4}$$

$$\frac{\sum_i^N(e_i)}{N} \approx v \tag{1.5}$$

Observe that this formulation is true no matter the value of N. On average, it is always better to have more estimates than fewer. This suggests that, even in the

[1]The normal distribution is well-supported by empirical evidence, but note that any symmetrical distribution around 0 will lead to the same conclusion.

situation where there are only two estimates (the decision-maker's and the advisor's), the best policy will be to incorporate both estimates into the final decision.

The variance of the normal distribution from which errors are derived (σ_i^2) is, in the example above, drawn from a normal distribution itself ($\sigma_i^2 \sim \mathcal{N}(\mu = 0, \text{sd}^2)$meaning that it is also cancelled out on average over repeated samples). Where few estimates are taken, weighting those estimates by the variance of the error distributions will increase the accuracy of the estimates in proportion to the difference between the variances:

$$e^F = \frac{\frac{1}{N}\sum_i^N \omega_i e_i^I}{\sum_i^N \omega_i}$$

Where ω_i is $1/\sigma_i^2$.

Many experimental implementations of this model avoid weighting issues by calibrating decision-makers and advisors to be equally accurate on average ($\sigma^2_{\text{decision-maker}} = \sigma^2_{\text{advisor}}$). The result of this constraint is that the optimal policy is simply to average all estimates together:

$$e^F = \frac{1}{N}\sum_i^N e_i^I \approx v$$

Egocentric-discounting

From the perspective of the normative model above, decision-makers should weigh their own estimate equally with each other estimate they receive in the process of coming to their decision. One of the most robust findings in the literature on advice-taking is that people routinely underweight advisory estimates relative to their own estimates, a phenomenon known as *egocentric discounting*(Dana and Cain 2015; Gino and Moore 2007; Hütter and Ache 2016; Liberman et al. 2012; Minson and Mueller 2012; Rader et al. 2017; Ronayne and Sgroi 2018; See et al. 2011; Soll and Mannes 2011; Trouche et al. 2018; Ilan Yaniv and Kleinberger 2000; Ilan Yaniv and Choshen-Hillel 2012; Ilan Yaniv and Milyavsky 2007). Egocentric discounting occurs in both feedback and no-feedback contexts (Ilan Yaniv and

Kleinberger 2000). Explanations for egocentric discounting are usually framed in terms of personal-level psychology: decision-makers have better access to reasons for their decision (Ilan Yaniv and Kleinberger 2000); overrate their own competence (Sniezek, Schrah, et al. 2004); may have a desire to appear consistent (Ilan Yaniv and Milyavsky 2007); may see opinions as possessions (Soll and Mannes 2011); may be loss-averse to providing a worse final estimate due to advice-taking (Soll and Mannes 2011); or have difficulty avoiding anchoring (Schultze et al. 2017) or repetition bias effects (Trouche et al. 2018). None of these explanations has survived rigorous empirical testing, however, and recently suggestions have widened to include consideration of aggregate-level rather than personal-level causes, with Trouche et al. (2018) arguing that the potential for misaligned incentives between decision-maker and advisor motivate discounting of advice. In the course of this thesis `crossref needed`, I demonstrate that egocentric discounting may be a stable metastrategy which protects against exploitation, carelessness, incompetence, and miscommunication. From this perspective, the normative model pertains to a particular instantiation of a problem with questionable ecological validity given the typical ethology of advice-taking in humans. While such considerations affect the conclusions one draws from egocentric discounting relative to the normative model, they do not detract substantially from the practice of using the normative model as an optimum 'set point' from which to evaluate advice-taking behaviour.

Homophily and echo-chambers

Homophily is the ubiquitous phenomenon that individuals more closely connected to one another within a social network tend to be more similar to one another than would be expected by chance across numerous dimensions, from demographics to attitudes (McPherson et al. 2001). Whether homophily in virtual social networks is responsible for increases in polarisation is debated. Proponents point to increases in polarisation (e.g. in politics: !TODO['Pew Research Center, 2014']), to empirical studies demonstrating homophily in virtual social networks (Cardoso et al. 2017; Colleoni et al. 2014), and to studies examining selective exposure online (Kobayashi and Ikeda 2009), and to echo chambers: egregious examples of highly homophilous networks with pathological polarisation (Sunstein 2002; Sunstein 2018). The empirical components of the argument are contested, with evidence that virtual social networks are less homogenous than offline social networks (and hence depolarising, Barberá 2015), and that selective exposure is a somewhat dubious finding which does not show up clearly online (Garrett 2009a; Garrett 2009b; Nelson and Webster 2017; Sears and Freedman 1967). Modelling work demonstrates, however, that where there is a bias in assimilation of information, homophily exacerbates polarisation (Dandekar et al. 2013). Where polarisation in turn increases homophily, for example through selective exposure or avoidance, a self-reinforcing spiral emerges wherein social connections become increasingly homogenous and attitudes increasingly extreme (H. Song and Boomgaarden 2017).

Source selection and information weighting

Context-dependency of epistemic processes

Part II

Psychology of advice

Contents

2

Psychological mechanisms of advisor evaluation

Where feedback is unavailable, people may use their own sense of certainty as a yardstick for evaluating advice (Pescetelli and Yeung 2018; Pescetelli 2017): advisors who agree when one is confident are perceived as more helpful; while those who disagree when one is confident are perceived as less helpful. Confidence serves as a proxy for objective feedback, and functions well in this role insofar as the judge has high metacognitive resolution (i.e. higher confidence is indicative of a greater probability of being correct).

Use of advice

Normative models of advice taking state that averaging estimates minimises errors. As discussed at length in Section III, the assumptions underlying the normative model do not always hold in the real world. The performance of the normative model can be characterised according to differences between the advisor and the decision-maker on ability and bias !TODO[soll and larick, soll and ? citations; how does PAR relate to what's going on here?]. Recall that the normative model states that advice should contribute to the final decision in proportion to the ability of the advisor compared to the decision-maker.

$$e_i^F = \frac{\omega_i e_i^I + \omega_j e_j^I}{2} \tag{2.1}$$

Where agent i is the decision-maker and agent j is the advisor. This weighting can be simplified to be expressed only in terms of the decision-maker's weighting of the advisor because the two are constrained to sum to 1 by virtue of being relative to one another:

$$\omega_i + \omega_j = 1 \tag{2.2}$$

$$\omega_j = 1 - \omega_i \tag{2.3}$$

$$\therefore e_i^F = \frac{(1-\omega_j) e_i^I + \omega_j a_i}{2} \tag{2.4}$$

Where a_i is the advice received (i.e. agent j's initial estimate - $a_i \equiv e_j^I$). In the normative model, the weighting is equivalent to the ratio of variance of the errors made by each agent:

$$\omega_j = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_i^2} \tag{2.5}$$

The normative model thus represents weighting by relative ability. Precise knowledge of the ability of others relative to oneself is rarely available in the real world, however, and, as discussed in Section III, other assumptions concerning the trustworthiness or interpretability of advice may be violated.

The normative model can be adapted to provide a more psychologically-realistic account of advice usage by substituting the three factor model of trust into the equations in place of the ability variable. We start with the statement within the three factor model that trust (ω) is proportional to ability (a), benevolence (b), and integrity (g).

$$\text{trust} \propto \text{ability} + \text{benvolence} + \text{integrity} \tag{2.6}$$

We can thus replace the measure of accuracy in the normative model with the measure of trust in order to calculate the relative weighting:

$$\omega_j = \frac{\text{trust}_j}{\text{trust}_j + \text{trust}_i} \tag{2.7}$$

At this point we may question whether the variable trust_i is a meaningful property or simply an artefact of mathematical symbol manipulation. Mathematically it provides a fixed point against which trustworthiness of advisors can be measured, allowing for scaling weightings meaningfully across different advisors in different decisions. In real world terms, while it is generally unlikely that benevolence_i and integrity_i will be anything less than maximal, perceptions of one's own ability (ability_i) are likely to allow for others to exceed it. I make no strong claims on the relationship between trust and its component variables other than proportionality, and within this conception it is meaningful to consider weighting as a property of trust in another's judgement relative to one's own, adjusted in some manner for the perception of that other's benevolence and integrity. If the concept of self-trust still appears untenable, note that trust_i can be replaced with a constant without compromising the equations.

We thus return to the normative model of advice-taking, but with an alternative derivation of the weighting between advice and initial estimate:

$$e_i^F = \frac{(1 - \omega_j)e_i^I + \omega_j a_i}{2} \tag{2.8}$$

2.0.1 Critique of the aggregation model

This conception of advice-taking as a weighted aggregation process between an initial estimate and advice underpins both the modelling and the experiments presented in this thesis. It is thus worth taking a little space to highlight areas in which this model is known to depart from reality.

Generality beyond judge-advisor systems

Firstly, the model is an idealised situation approximated by the experimental method: a decision-maker makes an explicit initial estimate, then receives advice, then makes an explicit final decision. Yaniv !TODO[it may have been Rader, or someone else] showed that preventing decision-makers from making initial decisions resulted in very different advice weighting, suggesting that this may be a model of a specific scenario rather than of advice integration per se. The model presented here could in principle explain an integration process where an initial estimate can only be made after the advice is known, but empically performs poorly. At best, it could be argued that pre-exposure to the advice either anchors the initial estimate (thus moving e_i^I systematically closer to a_i), or that having to trust advice because one cannot make one's own decision inflates the weighting of the advisor.

Multiple advisory estimates

Secondly, the model does not perform well when multiple advisors are consulted. The normative model, and the psychological derivative, predicts that a decision-maker's estimate ought to be weighted in conjunction with the other estimates. In other words, as the number of advisory estimates increases, the weight of the initial estimate should decrease. Research !TODO[CITE] shows that this does not happen: the weight of the initial estimate stays relatively constant while the weights of the advisor estimates are reduced. If a decision-maker were to average evenly their initial estimate with an advisor estimate ($\omega_i = .5$; $\omega_j = .5$), adding an extra advisor estimate would result in the weights of the advice being halved while the weight of the initial decision remained constant ($\omega_i = .5$; $\omega_{j \neq i} = .25$).

Individual trial data

The model is supported by patterns in averages. Analysis of individual trials shows that the aggregate patterns of advice-taking appear to be roughly distributed between an averaging strategy and a picking strategy !TODO[CITE someone on pick vs average, maybe soll?]. The model, derived from these patterns, approximates the contribution of an individual trial to the overall average rather than the actual advice-taking strategy on any given trial. ### Justification for use of the aggregation model

The criticisms above are important, but they do not invalidate the model for use in the present project. Here we seek to establish how differences in advice taking manifest according to properties of advisors. These differences are well characterised by the model, especially in the judge-advisor system used for the experiments. All models are wrong !TODO[cite George Box], and inclusion of a more complex model capable of handling the cases outlined above would require greatly increased complexity for relatively little gain in explanatory power. For studying the questions at hand, the psychological model is an appropriate and useful approximation of human behaviour.

Updating advisor weights

The weights assigned to the advisors (relative to the decision-maker themself) are subject to change as the result of experience. This experience can be exogenous or endogenous to the decision-making task. In the exogenous case, advisors may be labelled in a particular way !TODO[cite some experiments manipulating advisor labelling] or have some summary of their performance displayed !TODO[cite some experiments showing advisor performance data as a prompt]. Endogenous experience refers to the information that advice on a given trial carries about the trustworthiness of an advisor. Exogenous experience is relatively straightforward, but endogenous experience requires some explanation.

Endogenous experience of advice means that the weighting of an advisor is in part dependent upon the past advice offered by that advisor. As each piece of advice is evaluated, the overall weighting of the advisor is updated accordingly. For clarity, two simplifiying assumptions are made in the explanation below. Firstly, while it is probable that properties of the advice are used to inform the dimensions of ability, integrity, and benevolence simultaneously, the examples below will deal with ability in isolation. Another project could explore in detail how experience of advice on any given trial updates an advisor's position in 3-dimensional trust space in a Bayesian manner according to the relative certainties about each dimension. This would capture the task of assigning blame for erroneous advice (e.g. was it unintentionally poor - a failure of ability - or deliberately misleading - a failure of benevolence?). Such an undertaking is beyond the scope of this project.

Secondly, it is assumed that advice is judged on its own merit as an estimate rather than on its usefulness as advice. The former means that advice is assessed in terms of the optimality of the decision recommended by the advice itself. The latter assesses advice based on the optimality of the decision based on advice relative to the optimality of the decision which *would have been made had the advice not been received.* There is some evidence that people alter their advice-giving behaviour in anticipation of discounting on the part of the decision-maker !TODO[CITE; for a case in human-machine teaming see @azariaStrategicAdviceProvision2016], somewhat akin to starting negotiations with a higher demand than one is hoping to settle for. There is no evidence as yet as to whether decision-makers anticipate and adjust for this adjustment on the part of the advisor. For the questions considered here, conclusions obtained under these simplifying assumptions are likely to hold even when the additional complexity is restored. The effects in the real world of interactivity between trust dimensions and game theoretic adjustments in the giving and interpretation of advice are likely to be small in comparison to general effects of advisor updating. ### Evaluation of advice {-}

A single piece of advice can be evaluated using its own properties and the properties of the advisor giving the advice. Furthermore, that evaluation can serve to update the properties of the advisor. A piece of advice's own properties will include its plausibility (e.g. participants in estimation tasks discount advice which is distant from their own initial estimates more heavily (I. Yaniv 2004)), while the properties of the advisor will include the advisor's trustworthiness (see above). The updating of trust following experience of advice is likely to be largely in the domain of ability, although other domains may be affected where the advice is particularly egregious. ## Updating advisor evaluations {-}

While a single piece of advice must be taken on its own terms, people can construct relatively accurate estimates of advisors' advice when provided with feedback on the decisions they use the advice to make (Pescetelli 2017; Sah et al. 2013; Ilan Yaniv and Kleinberger 2000). This likely happens as an analogue of reinforcement learning, where feedback allows an error signal to be used to update the estimate of the advisor's ability ($\hat{s}^a$) rather than one's own beliefs about the world, according to some learning rate (λ).

$$\text{this is wrong; need to check RL models for an analogue} \tag{2.9}$$

$$\hat{s}_{t+1} = \hat{s}_t + |e^a_t - v| \cdot \lambda \tag{2.10}$$

2.0.2 Criticism of the advisor evalutation model

Ecological validity

While many experiments have established the existence reinforcement learning in humans and other animals, it is unclear whether reinforcement learning operates in the social domain in which advising takes place. It is not obvious that there are many situations in the course of everyday relationships which can be characterised by the rapid advice-feedback cycle required to learn about advisor ability in the manner modelled above. !TODO[who wrote this paper]? argued in a review that a wide variety of social phenomena could be explained via reinforcement learning processes. Additionally, Heyes and colleagues have argued that social learning is wholly explicable in terms of general reinforcement learning processes !TODO CITE. Reinforcement learning in the social domain operates on the basis of rapid feedback, just as in the non-social domain. Below, the advisor evaluation model is extended to cases where objective feedback is not available by substituting the decision-maker's confidence for objective feedback. While not foolproof, the method allows better-than-average approximation of the quality of advisors.

Advisor evaluation without feedback

Where feedback is not available, participants in experiments continue to demonstrate an ability to respond rationally to differences in advisor quality (Pescetelli 2017). This is evidently not done through access to the correct real-world values, because feedback providing those values is unavailable, and, were participants aware of those values themselves, it stands to reason they would have provided those values (and thus not require advice!). Pescetelli and Yeung (2017) suggest the mechanism for this ability to discriminate between advisors in the absence of feedback is performing updates based on confidence-weighted agreement.

Agreement

Consider first the non-weighted agreement case, where the advisor's estimate at time t (e^a_t) and the decision-maker's estimates (e^d) are binary ($\in 0, 1$). The estimate of the advisor's ability ($\hat{s}^a$) is updated positively if the advisor and decision-maker agree, and negatively otherwise, according to the learning rate λ.

$$\hat{s}^a_{t+1} = \hat{s}^a_t + (-2\ |e^d_t - e^a_t| + 1) \cdot \lambda \tag{2.11}$$

Confidence-weighted agreement

The updating of advice contingent on agreement may be weighted by confidence in the initial decision (c^d), such that agreement and disagreement are considered more informative about the quality of the advice when the decision with which they agree or disagree is more certain.

$$\hat{s}^a_{t+1} = \hat{s}^a_t + (-2\ |e^d_t - e^a_t| + 1) \cdot c^d_t \cdot \lambda \tag{2.12}$$

Continuous estimate case

2.1 Measuring advice-taking

Judge-advisor system

Table of experiments

Advisors	Choice	Task	Feedback	Result
AiC vs AiU	Yes	Perceptual, binary (MATLAB)	No	Suggestive
AiC vs AiU	Yes	Perceptual, binary	No	Inconclusive
In/accurate	Yes	Perceptual, binary	No	Accuracy selected more often
Low/High agreement	Yes	Perceptual, binary	No	Agreement selected more often
Accurate vs Agreement	No	Estimation, continuous	Yes and No	Accurate preferred given feedback, agreement preferred without feedback

3
Behavioural experiment method

The behavioural experiments reported in this thesis share a common structure. This structure is detailed here to reduce repetition elsewhere in the thesis. Individual experiments have truncated methods sections in which the specific deviations from the general method are noted. ## General method {#gm}

The experiments take place using a judge-advisor system. Participants give an **initial estimate** for a decision-making task, receive **advice**, and then provide a **final decision**. The advice is always computer-generated, although the specifics of the generating procedure vary between experiments. ### Participants {#m-participants}

Recruitment

Human participants were recruited from the online experiment participation platform Prolific (`https://prolific.co`). Participants were prevented from taking the study if they had participated in one of the other studies in the thesis, or if they had an overall approval rating on Prolific of less than 95/100.

Payment

Participants were paid approximately GBP10-15/hour pro rata. Experiments took the average participant between 10 and 30 minutes to complete.

Some participants encountered technical problems, prompting them to contact me via the Prolific platform. These participants were thanked, and additional information about the problem sought if necessary. These participants were paid on an ad hoc basis depending upon the time taken before the errors emerged and the detail of the error reports.

Later studies introduced attention checks which terminated the study as a consequence for failure. It is not clear whether this technique constitutes best practice on Prolific because automatic termination means participants may return the study rather than having their participation explicitly rejected (and thus affecting their Prolific participation rating). Participants who returned the studies were not paid. Participants who attempted to complete the study with an invalid code after having their participation terminated for failing attention checks were also not paid, and their completion attempt was rejected on the Prolific platform. There is an ongoing ethical debate concerning non-payment of participants who fail attention checks in online studies. The studies in this thesis used a mixture of paying for and not paying for participation attempts with failed attention checks. In online studies, where low-effort participation is a serious and enduring concern, platforms such as Prolific make clear to participants and researchers that payment is only expected for responses which are given with satisfactory effort. Participants are thus fully aware of and consenting to the process of screening results for adequate effort prior to payment. It is important to note the difference between low effort responses, for which payment may ethically be withheld, and atypical responses, which may represent genuine engagement with the task. It is unethical, in my view, to withhold payment from participants for atypical responses within an experiment (including low or high response times, accuracy, etc).[1] Participants should only be denied

[1]Data may be excluded from *analysis* for these reasons, but the participants should still be paid.

payment for failing to provide adequate responses to explicit attention checks.

Demographics

Demographic information on participants, such as age and gender, was not collected. While there is a robust case for collecting these data and conducting sex-disaggregated analyses (**criadoperezInvisibleWomenExposing2019**), initial concerns over General Data Protection Regulation resulted in a cautious approach to the collection of data concerning protected characteristics of participants. !TODO[move to discussion]Gender differences, whether due to socialisation, biological factors, or their interactions, may well alter advice-taking and expressed confidence in decisions. I suspect, although I can offer no evidence, that gender differences in the results presented in this thesis will at most show overlapping distributions. I do not think it highly plausible that different strategies are wholly the preserve of any particular gender, or that egocentric discounting is markedly stronger in any particular gender.

Participants were at least 18 years of age, confirmed by the requirements for possessing an account on the Prolific platform and by explicit confirmation when giving informed consent.

3.0.1 Ethics

Ethical approval for the studies in the thesis was granted by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (References: R55382/RE001; R55382/RE002).

3.0.2 Procedure

Participants visited the Uniform Resource Locator for the study by following a link from Prolific using their own device (Figure 3.1). Early studies only supported computers, but later studies included support for tablets and smartphones. After viewing an information sheet describing the study and giving their consent to participate, participants began the study proper. The study introduced the software

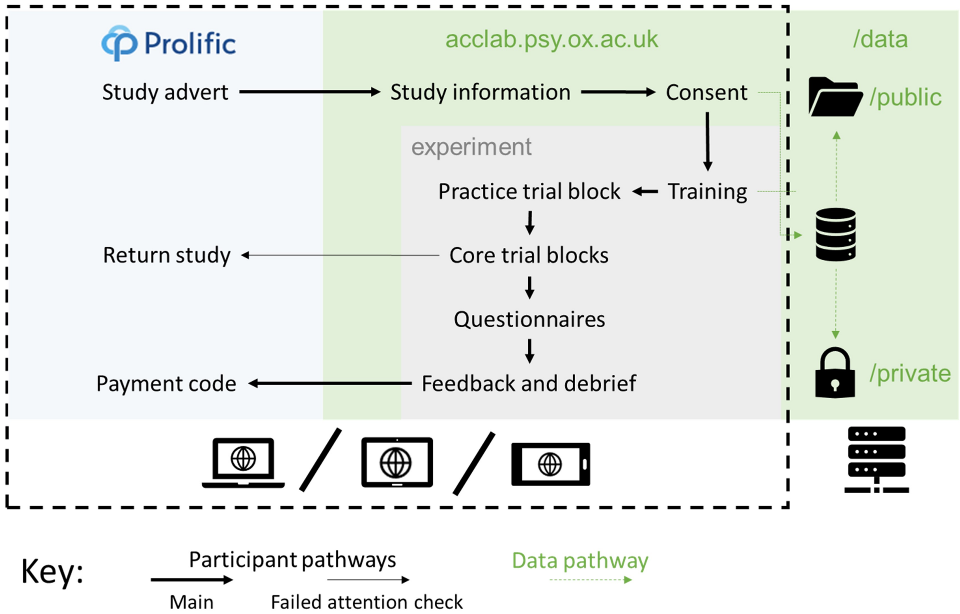

Figure 3.1: Participant pathway through the studies.
Participants used their own devices to complete the study, which was presented on a website written in HTML, CSS, and JavaScript. The data were saved on the server using PHP.

to the participant interactively, demonstrating the decision-making task and how responses could be made. Next, participants were given a block of practice trials to familiarise them with the decision-making task. Participants were then introduced to advice, and given a block of practice trials in which they received advice. The core experimental blocks followed the practice with advice. Finally, debrief questions were presented and feedback provided concerning the participant's performance, including a stable link to the feedback and a payment code. The participant entered the payment code into the Prolific platform and their participantion was at an end.

On each trial, participants were faced with a decision-making task for which they offered an initial estimate. They then received advice (on some trials they were able to choose which of two advisors would provide this advice). They then made a final decision. On feedback trials, they received feedback on their final decision. The schematic for this trial structure is shown for the Dots Task in Figure 3.2.

!TODO[Check feedback duration and style in image caption]

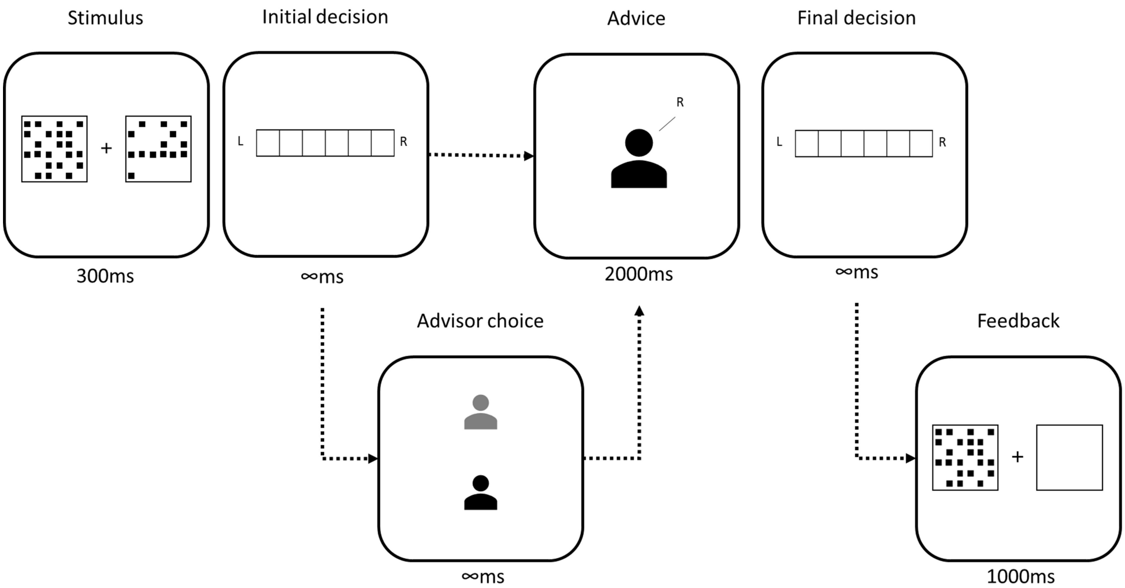

Figure 3.2: Trial structure of the Dots Task.
In the initial estimate phase, participants saw two boxes of dots presented simultaneously for 300ms. Participants then reported whether there were more dots on in the left or the right box, and how confident they were in this decision. Participants then received advice, sometimes being offered the choice of which advisor would provide the advice. The advice was displayed for 2000ms before participants could submit a final decision, again reporting which box they believe contained more dots and their confidence in their decision. On feedback trials, feedback was presented by redisplaying the correct box while showing the other box as empty.

Perceptual decision (Dots Task)

Stimuli in the Dots Task consisted of two boxes arranged to the left and right of a fixation cross (Figure 3.3). These boxes were briefly and simultaneously filled with an array of non-overlapping dots, and the participant was instructed to indentify the box with the most dots. The number of dots was exactly determined by the difficulty of the trial: the box with the !TODO[check this description] least dots had 200 - the difficulty, while the box with the most had 200 + the difficulty. The dots did not move during the presentation of the stimulus. There was thus an objectively correct answer to the question which, given enough time, could be precisely determined from the stimulus.

The Dots Task stimuli can be customised to make the discrimination easier or more difficult. This means that the stimuli can be adjusted to maintain a

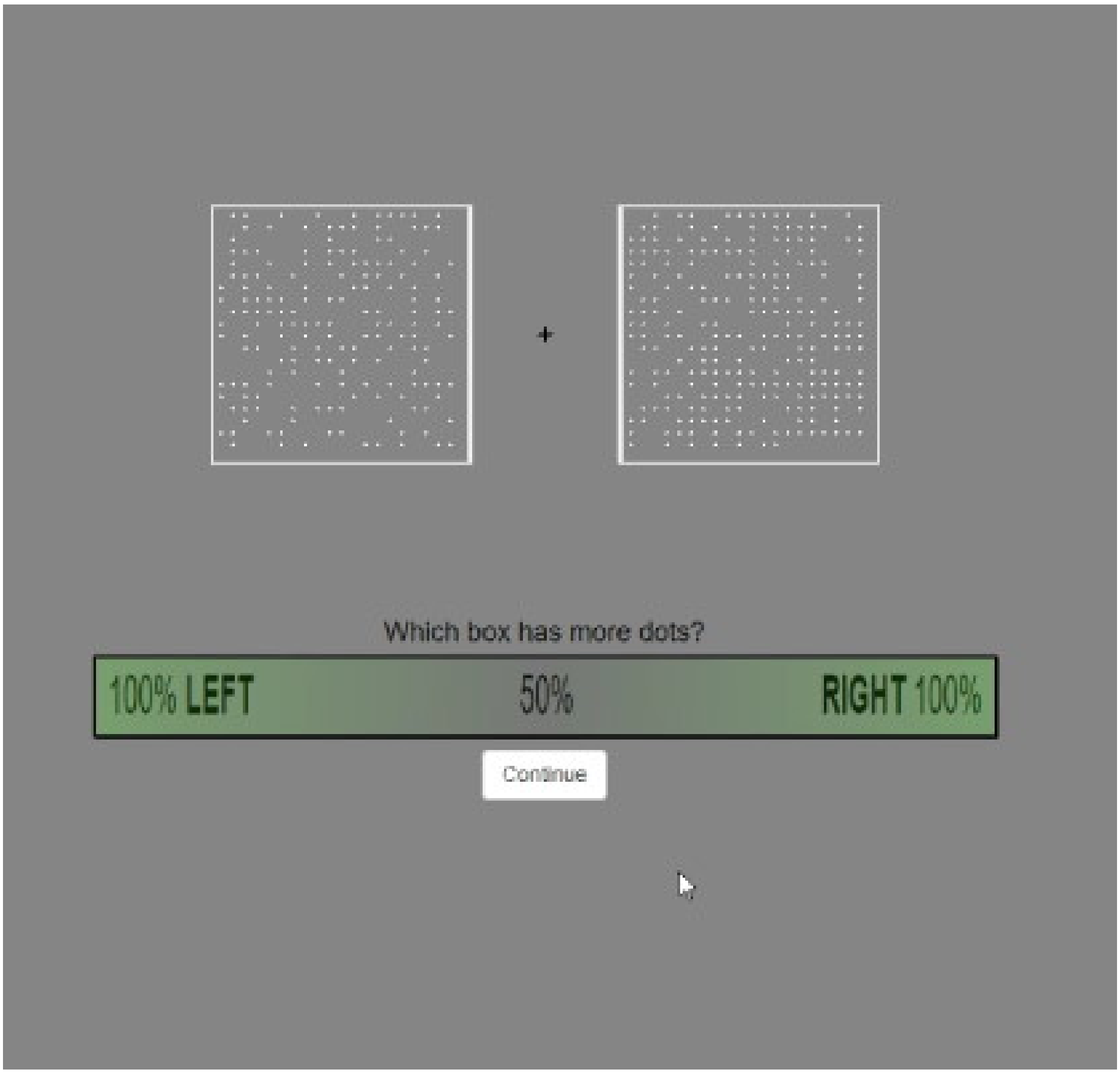

Figure 3.3: Dots Task stimulus.

specific accuracy for each individual participant, allowing confidence to be examined in the absence of confounds with the probability of being correct. Stimuli were continually adjusted throughout the experiment to maintain an initial estimate accuracy of around 72% using a 2-down-1-up staircase procedure. There were a substantial number !TODO[how many?] of trials in the practice block so that participants could eliminate practice effects and thus experience a more stable objective difficulty during the core trial blocks. After each block participants were told what percentage of the final decisions they had provided were correct and allowed to take a short, self-paced break.

Figure 3.4: Dates Task with continuous responses.

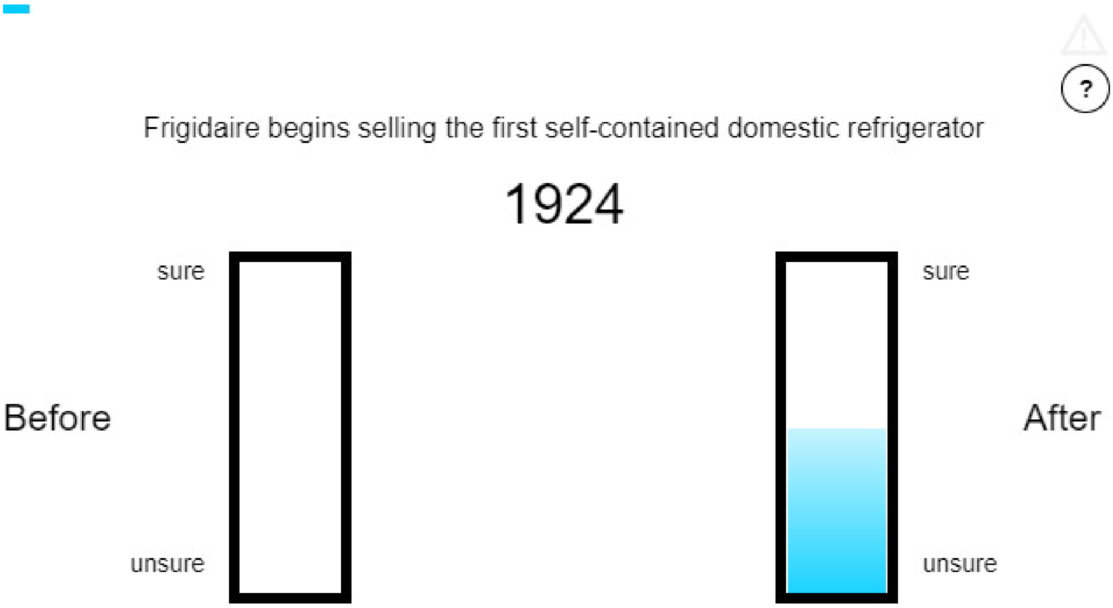

Figure 3.5: Dates Task with binary responses.

Estimation (Dates Task)

Rationale

Continuous

Table 3.1: Table : Experiment 1 advisor advice profiles

	Initial decision confidence	Probability of agreement (%)	
		Bias Sharing	Anti Bias
Participant correct	High (top 30%)	80	60
	Medium (middle 40%)	70	70
	Low (bottom 30%)	60	80
Participant incorrect	Any	30	30
Total agreement	Participant correct	70	70
	Participant incorrect	30	30

Binary

3.0.3 Advisor advice profiles

The advisers are virtual agents whose probability of agreeing with the participant's decision varies as a function of the participant's confidence and correctness in the initial decision phase. Table 3.1 illustrates how this relationship functions, and shows that the overall correctness and agreement rates of the advisers is equivalent overall. Importantly, on largest minority of trials, the middle 40%, the advisers are exactly equivalent, meaning these trials can be compared directly without confounds arising from agreement rate and initial confidence.

3.0.4 Analysis

Dependent variables

Advisor choice

Weight on Advice

Statistics

Bayesian statistics

Frequentist statistics

Software Data analysis was performed using R (R Core Team 2018). For a full list of packages and software environment information, see !TODO[figure out where to include this stuff. Appendix? Also link to a containerized version of this.]

Bayes Factors (BF) are presented alongside p values and test statistics. A BF < 0.33 indicates decisive evidence in favour of the null hypothesis over the alternative hypothesis (with lower values being increasingly clear), BF > 3 indicates decisive evidence of the alternative over the null (with higher values being increasingly clear), and 0.33 <= BF <= 3 indicates there is insufficient evidence to reach a conclusion.

3.0.5 Capped influence

Influence, the dependant variable in some analyses, is calculated as the extent to which the judge's initial decision is revised in the direction of the advisor's advice. The initial (C_1) and final (C_2) decisions are made on a scale stretching from -55 to +55 with zero excluded, where values <0 indicate a 'left' decision and values >0 indicate a 'right' decision, and greater magnitudes indicate increased confidence. Influence (I) is given for agreement trials by the shift towards the advice:

$$I|\text{agree} = f(C_1)\begin{cases} C_2 - C_1 & C_1 > 0 \\ -C_2 + C_1 & C_1 < 0 \end{cases} \tag{3.1}$$

And by the inverse of this for disagreement trials:

$$I|\text{disagree} = -I|\text{agree} \tag{3.2}$$

The confidence scale excludes 0, and thus the final decision can always be more extreme when moving against the direction of the initial answer than when moving further in the direction of the initial answer. A capped measure of influence was used to minimise biases arising from the natural asymmetry of the scale. This measure was calculated by truncating absolute influence values which were greater than the maximum influence which could have obtained had the final decision been a maximal response in the direction of the initial answer (Figure 3.6).

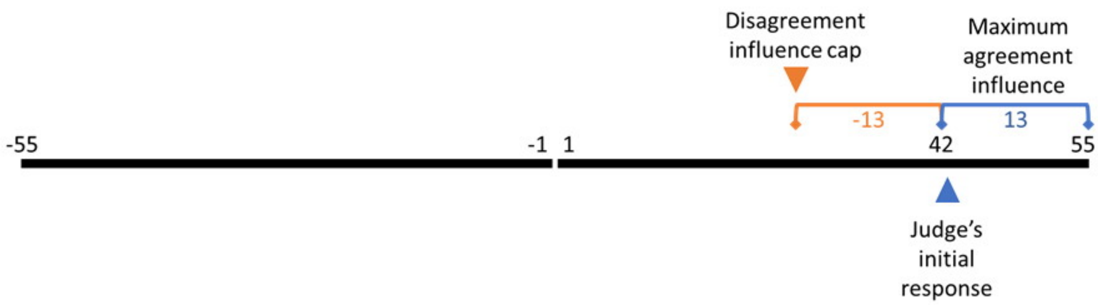

Figure 3.6: Capping influence to avoid scale bias.
In this example the judge's initial response is 42, meaning that their final decision could be up to 13 points more confident or up to 97 points less confident. Any final decision which is more than 13 points less confident is therefore capped at 13 points less confident.

The capped influence measure I_{capped} is obtained by:

$$I_{\text{capped}} = f(C_1)\begin{cases}\min(I, 2C_1 - 55) & C_1 > 0\\ \max(I, 2C_1 + 55) & C_1 < 0\end{cases} \tag{3.3}$$

The explicit measure of trust is obtained using questionnaires. The questionnaires are delivered at 4 time points, and consist of 4 questions per advisor which are answered on a 1-100 scale.

3.1 Open science approach

3.1.1 Open science

Nullius in verba ("take nobody's word for it") is written in stone above the entrance to the Royal Society's library. This fundamental principle of science, that it proceeds on evidence rather than assertion, has frequently been forgotten in practice. Concerns about sloppy, self-deluding, or outright fradulent science have existed since at least the time of Bacon. The modern open science movement in psychology dates from the early 2010s. Simmons et al. demonstrated how easily false positive results could emerge from unconstrained researcher degrees of freedom in analysis (**simmonsFalsePositivePsychologyUndisclosed2011**), Nosek and colleagues published a roadmap for improving the structure and function of academic research and publishing (**nosekScientificUtopiaOpening2012**;

nosekScientificUtopiaII2012), and the Open Science Collaboration began (**collaborationEstima**
In the years following, a deluge of papers, movements, and practical changes have emerged. The meaning of open science varies within each sub-discipline, and this section outlines how the experiments comprising this thesis have been conducted in a reproducible and transparent manner.

3.1.2 Badges

Following the Center for Open Science (`https://cos.io`), this thesis uses a series of badges to indicate adherence to particular aspects of open science. Three badges, *preregistration*, *open materials*, and *open data*, are adopted directly from the Centre and used according to the Centre's rules (`https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/`). Studies which qualify for a badge will have the badge displayed immediately below their title. Each badge will contain a link to online resources which provide the content for which the badge is awarded.

Preregistration

Preregistration of a study means that information about the study has been solidified prior to the analysis of the data. This means that hypotheses cannot be changed to represent unanticipated or overly-specific findings as a priori predicted (**kerrHARKingHypothesizingResults1998**). In practice in this thesis, preregistration means describing in detail the design and analysis plan for an experiment and depositing the description with a reputable organisation prior to data being collected. The links which accompany the preregistration badge will point to the preregistration document. These measures help to prevent presenting a highly selected and biased interpretation of the data as the result of a natural analytical process.

The preregistration badge also appears within results sections to designate those statistical investigations which were included in the preregistration. Some analyses are exploratory. These exploratory analyses are not included in the preregistration, because they are inspired by the data themselves. They are reported after the preregistered analyses, or are clearly designated as exploratory in the text.

Open materials

A foundational principle of science is that findings can be reproduced by other people. Open materials facilitate reproduction by making it easier to rerun an experiment. Open materials also increase the likelihood that errors can be identified. In the case of the behavioural experiments reported here, the open materials include computer code necessary to run the experiment. The links accompanying the open materials badge points to this code.

Open data

Theories are the output of science as a whole, but data are the output of any individual study. Sharing data directly allows other scientists to check and extend the data analysis conducted, to reuse the data in meta-analyses, and to repurpose the data for other investigations. This increases the robustness of the results, and increases the efficiency of science as a whole. The links accompanying the open data badge point to online storage where the data can be obtained for a study, along with appropriate metadata.

3.1.3 Thesis workflow

This thesis is written in RMarkdown, with the data fetched and analysed at the time the document is produced using the publically available pipeline - the entire document can be reproduced locally using the source code in an appropriate environment. A Docker environment copying the environment used to produce this document is available at !TODO[the containerisation thing]

4
Psychology of source selection

Contents

The model of advisor evaluation described earlier requires empirical support. The model of advisor evaluation with feedback is well supported by data which indicate that, given objective feedback, people can use the feedback to learn about the trustworthiness of advisors. The extent to which advice is taken is commonly used as a measure of a participant's trust in an advisor, on the argument that the participant seeks to maximise task performance and task performance is maximised by taking more advice from more trustworthy advisors. !TODO[Mini lit review - advisor accuracy training with feedback (Yaniv?), Niccolo and Nick's feedback stuff]. Overall, this means that people prefer more accurate advice over less accurate advice.

When objective feedback is unavailable, people can still demonstrate a greater dependence upon advice from more as opposed to less accurate advisors. !TODO[Mini lit review, Niccolo's stuff, maybe prior stuff - check Niccolo's thesis for references]. This is a consequence of agreement: where the base probability of being correct is greater than chance, the independent estimates of people who are more accurate will agree more often (leading to 100% agreement on the correct answer for two independent decision-makers of perfect accuracy). In the absence of feedback, therefore, agreement can be used as a proxy for accuracy, as formalised in the model.

The role of agreement is demonstrated clearly in experiments where the objective accuracy of advisors is balanced, but the agreement rates of the advisors is varied. !TODO[cite Niccolo CITE]. Pescetelli and Yeung demonstrated that advice is more influential from advisors who tend to agree with a participant more frequently when objective feedback is not provided. This is despite the fact that advice is more influential when it disagrees with the participant's initial estimate. [1] These data suggest that people may be using agreement as a proxy for accuracy, although they may simply prefer agreement over disagreement when there is no accuracy cost to be paid. I report the results of an experiment in which an agreeing advisor was compared with an accurate advisor under conditions of feedback or no feedback. Results indicated that, as predicted by the models, participants preferred the accurate advisor when feedback was provided and the agreeing advisor when feedback was withheld.

[1] This is partly due to the nature of the judge-advisor system: there is always room for disagreement to be more extreme than agreement, because agreement is lower-bounded by the participant's initial estimate.

Pescetelli and Yeung developed a more sophisticated model of advisor evaluation in which the increase in trust gained when an advisor agreed with the decision-maker was contingent upon the confidence of the decision-maker's initial estimate. Intuitively, if I am highly certain that I am correct on a given question, an advisor who disagrees with me is likely to be incorrect, whereas one who agrees with me is likely to be correct. Provided confidence is indicative of the objective probability of being correct, as confidence in the initial decision increases it more closely approximates objective feedback for the purposes of evaluating advice. !TODO[detailed account of Niccolo's evidence for the confidence-weighted model]. I report the results of experiments designed to extend Pescetelli and Yeung's results to the domain of advisor influence.

!TODO[move this section to a discussion somewhere?] An alternative explanation for the data is that disagreeing advice is more surpising when it comes from an advisor who usually agrees, and that it is its increased salience increases its influence rather than increased trust in the advisor. If that were the case then the same ought to be true for agreeing advice from a disagreeing advisor, and this is not the case \mccorrect{!TODO[Do we have/can we produce evidence for this? Perhaps look at our participants where advisors disagree with them by chance more than e.g. 70% of the time and check for increased influence on agreement trials vs advisors who disagree less than 70% of the time]}.

4.1 Accuracy?

!TODO[check Niccolo covered this] {#ac-acc}

Pescetelli and Yeung !TODO[cite new paper] demonstrated that more accurate advisors are more influential (regardless of the presence of feedback) in a lab-based perceptual decision-making task. We attempted to extend this finding to the domain of advisor selection in two online tasks: a 'Dots Task' requiring similar perceptual decision-making to the task used by Pescetelli and Yeung, and an estimation-based 'Dates Task'.

The ability to distinguish between accurate advisors in these experiments is important because they relate directly to the phenomenon we are attempting to explain: rational advice-seeking behaviour in the absence of feedback.

4.1.1 Dots Task

Open scholarship practices

https://osf.io/u5hgj

!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html

Unanalysed data Several early versions of this experiment were run where bugs in the experiment code made the results unreliable. The earliest versions contained a bug where advisors instructed to agree with a participant instead provided advice identifying the correct answer. Other versions had a bug in the staircasing code used to titrate the difficulty of the task was converging on too high a value (74% initial decision accuracy as opposed to 71%). Once the staircasing bug was fixed, two more experiments were run, one with 60 practice trials in which participants did not quite reach the desired accuracy before the beginning of the main experiment, and one with 120 practice trials which constitutes the data analysed below. Overall, participants' data was collected in these excluded versions and not included in analysis. While not useful for the hypothesis of this experiment, the excluded data can be used for analysing responses to advice, provided care is taken with the very early versions to ensure advice is interpreted correctly.

Method

!TODO[clarify any methodological differences from the main methods chapter]

Table 4.1: Table : Advisor advice profiles for Dots task Agreement experiment

	Probability of agreement (%)	
	High accuracy	Low accuracy
Participant correct	.800	.600
Participant incorrect	.200	.400
Total agreement	.626	.542

Advisor profiles The two advisor profiles (Table 4.1) used in the experiment were High accuracy and Low accuracy. The High accuracy advisor was correct 80% of the time while the Low accuracy advisor was correct 20% of the time. The advisor profiles were not balanced for overall agreement rates.

```
## Warning: ‘is.tibble()‘ is deprecated as of tibble 2.0.0.
## Please use ‘is_tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
## Warning: ‘as.tibble()‘ is deprecated as of tibble 2.0.0.
## Please use ‘as_tibble()‘ instead.
## The signature and semantics have changed, see ‘?as_tibble‘.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

Results

Exclusions

Task performance

Mainpulation checks

Hypothesis test

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
## Picking joint bandwidth of 0.0409
```

$t(58) = 3.84$, $p < .001$, $d = 0.50$, $\mathrm{BF} = 79.12$; $M = 0.57$ $[0.54, 0.61]$, $\mu = 0.5$

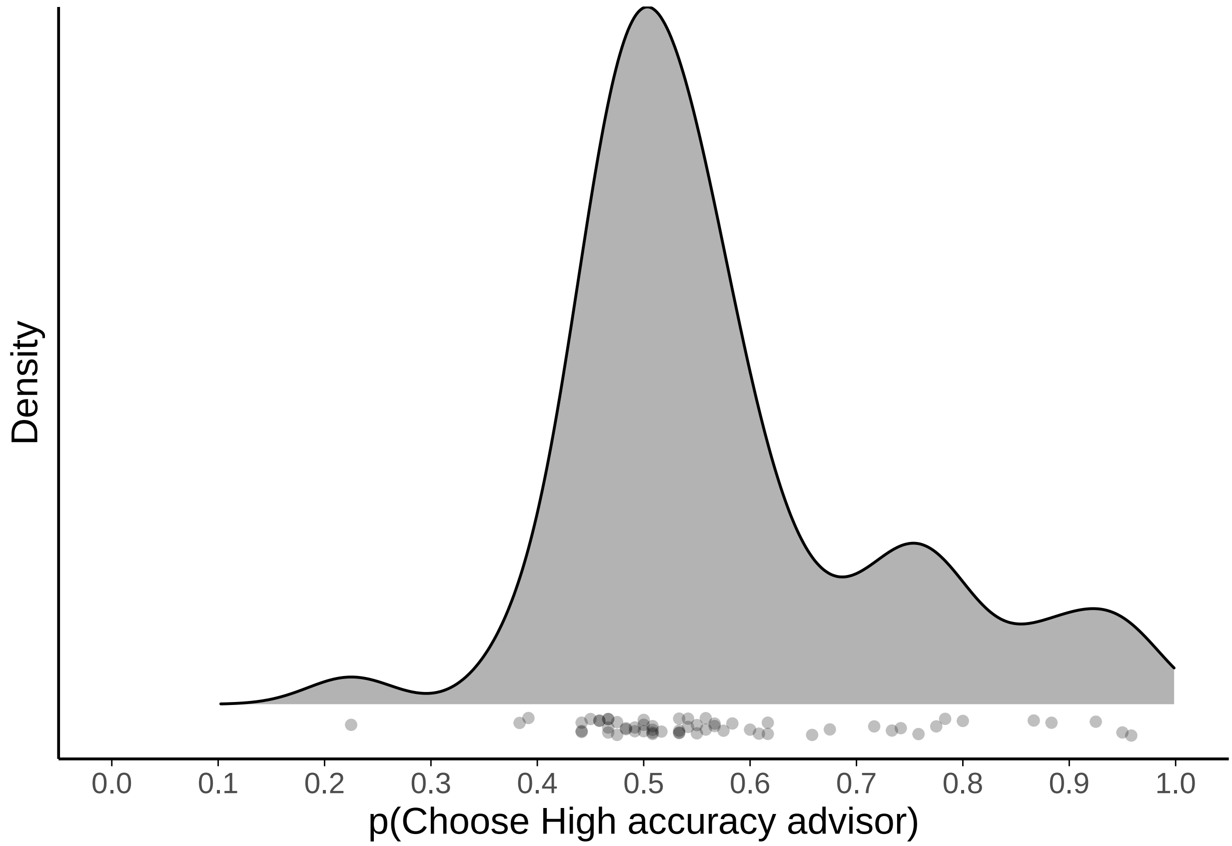

Figure 4.1: Participants' pick rate for the advisors in the Choice phase of the experiment. The shaded area shows a density plot of the individual participants' pick rates, shown by semi-transparent dots.

4.1.2 Dates Task

Open scholarship practices

`https://osf.io/5xpvq`

!TODO[OSFify data for these studies]

`https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html`

Unanalysed data Early versions of this experiment (v0-0-1, v0-0-2) included a bug which prevented feedback from being shown during the familiarisation phase even to participants in the Feedback condition. The 13 participants whose data was collected in these versions is not included in analysis. These participants could theoretically be included in the No feedback condition regardless of their condition label in the data, but this is not done here.

Method

This study used the Binary version of the Dates Task.

Results

Exclusions

Task performance

Mainpulation checks

Hypothesis test

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## Picking joint bandwidth of 0.137
```

$t(33) = 3.41$, $p = .002$, $d = 0.58$, $\mathrm{BF} = 19.70$; $M_{Feedback} = 0.67$ [0.57, 0.78], $\mu = 0.5$

$t(27)$ = -0.93, p = .363, d = 0.18, BF = 0.30; $M_{Nofeedback}$ = 0.45 [0.33, 0.57], μ = 0.5

4.1.3 Discussion

Where feedback is provided on advisors' performance, people seem to prefer high accuracy advisors to low accuracy advisors. Where feedback is not provided, people may need substantially more experience to learn that some advisors are more accurate than others, because this happens in the Dots task but not in the Dates task.

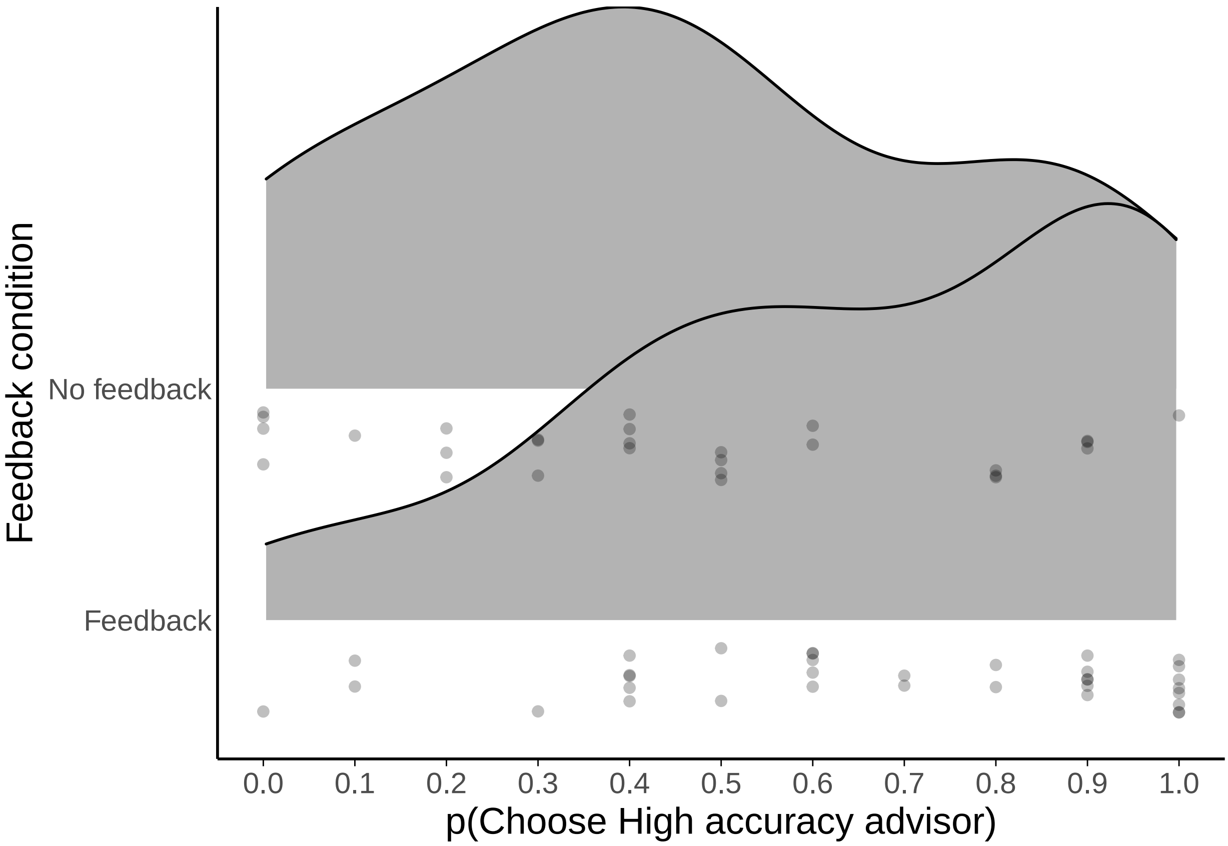

Figure 4.2: Participants' pick rate for the advisors in the Choice phase of the experiment. The shaded area shows a density plot of the individual participants' pick rates, shown by semi-transparent dots. Participants in the Feedback condition received feedback during the Familiarization phase, but not during the Choice phase.

4.2 Agreement !TODO[check Niccolo covered this]

Pescetelli and Yeung !TODO[cite new paper] demonstrated that advisors who agree !TODO[somewhere we need to talk about what we mean by agreement, how Niccolo defined it, how we define it (varies between binary/continuous tasks), etc.] more frequently are more influential (regardless of the presence of feedback) in a lab-based perceptual decision-making task. We attempted to extend this finding to the domain of advisor selection in two online tasks: a 'Dots Task' requiring similar perceptual decision-making to the task used by Pescetelli and Yeung, and an estimation-based 'Dates Task'.

The ability to distinguish between accurate advisors in these experiments is important because they relate directly to the phenomenon we are attempting to

Table 4.2: Table : Advisor advice profiles for Dots task Agreement experiment

	Probability of agreement (%)	
	High agreement	Low agreement
Participant correct	.840	.660
Participant incorrect	.610	.170
Total agreement	.773	.518

explain: rational advice-seeking behaviour in the absence of feedback.

4.2.1 Dots Task

Open scholarship practices

NA

!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html

Unanalysed data There were no unanalysed data for this experiment.

Method

!TODO[clarify any methodological differences from the main methods chapter]

Advisor profiles The two advisor profiles (Table 4.2) used in the experiment were High accuracy and Low accuracy. The High accuracy advisor was correct 70.95% of the time while the Low accuracy advisor was correct 70.93% of the time. The advisor profiles were not balanced for overall agreement rates.

Results

Exclusions

Task performance

Mainpulation checks

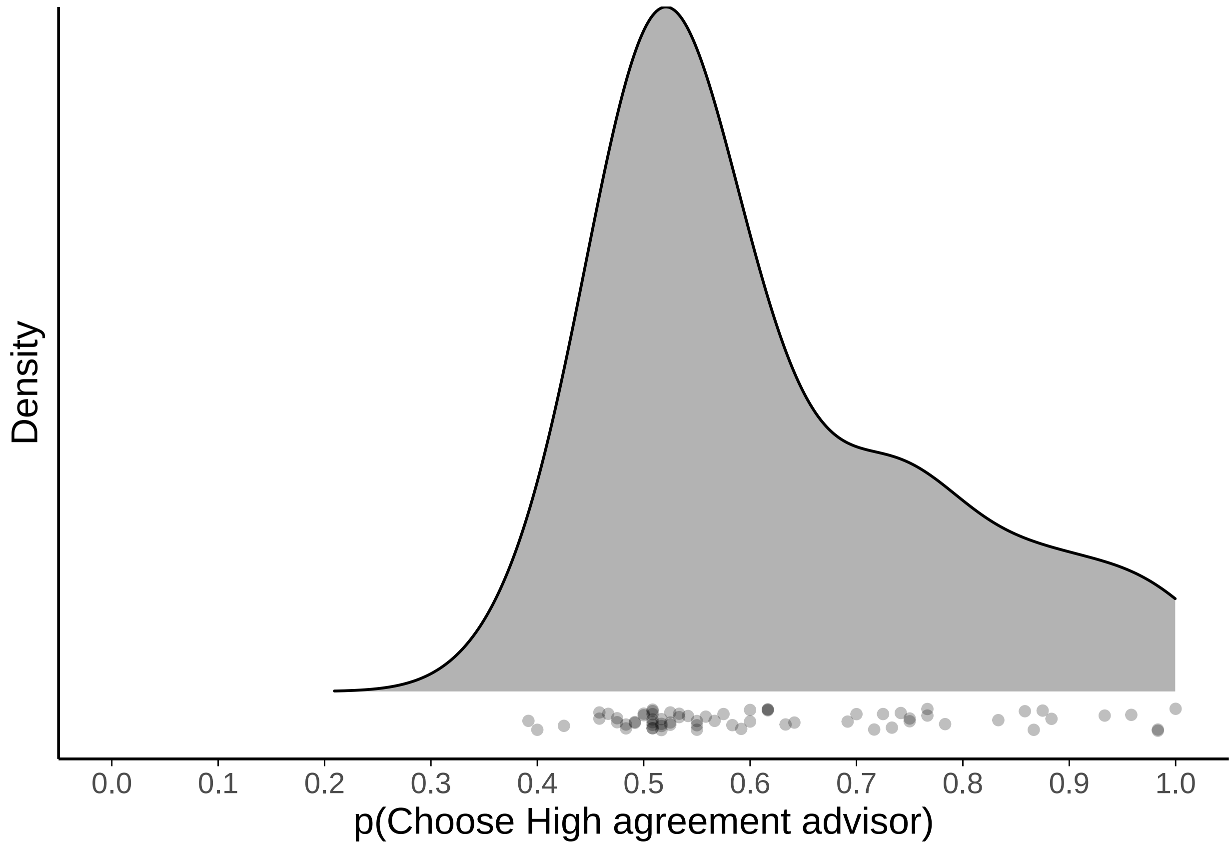

Figure 4.3: Participants' pick rate for the advisors in the Choice phase of the experiment. The shaded area shows a density plot of the individual participants' pick rates, shown by semi-transparent dots.

Hypothesis test

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
## Picking joint bandwidth of 0.0608
```

$t(67) = 6.15$, $p < .001$, $d = 0.75$, $\mathrm{BF} = 265393.82$; $M = 0.62$ $[0.58, 0.66]$, $\mu = 0.5$

4.2.2 Dates Task

!TODO[This needs to be about AGREEMENT, not accuracy!]

Open scholarship practices

https://osf.io/8d7vg

!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html

Unanalysed data Early versions of this experiment (v0-0-1, v0-0-2) included a bug which prevented feedback from being shown during the familiarisation phase even to participants in the Feedback condition. The 13 participants whose data was collected in these versions is not included in analysis. These participants could theoretically be included in the No feedback condition regardless of their condition label in the data, but this is not done here.

Method

This study used the Binary version of the Dates Task.

Results

Exclusions

Task performance

Mainpulation checks

Hypothesis test

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## Picking joint bandwidth of 0.114
```

$t(38) = 0.46$, $p = .648$, $d = 0.07$, BF $= 0.19$; $M_{Feedback} = 0.52$ [0.42, 0.62], $\mu = 0.5$

$t(34) = 2.62$, $p = .013$, $d = 0.44$, BF $= 3.39$; $M_{Nofeedback} = 0.63$ [0.53, 0.73], $\mu = 0.5$

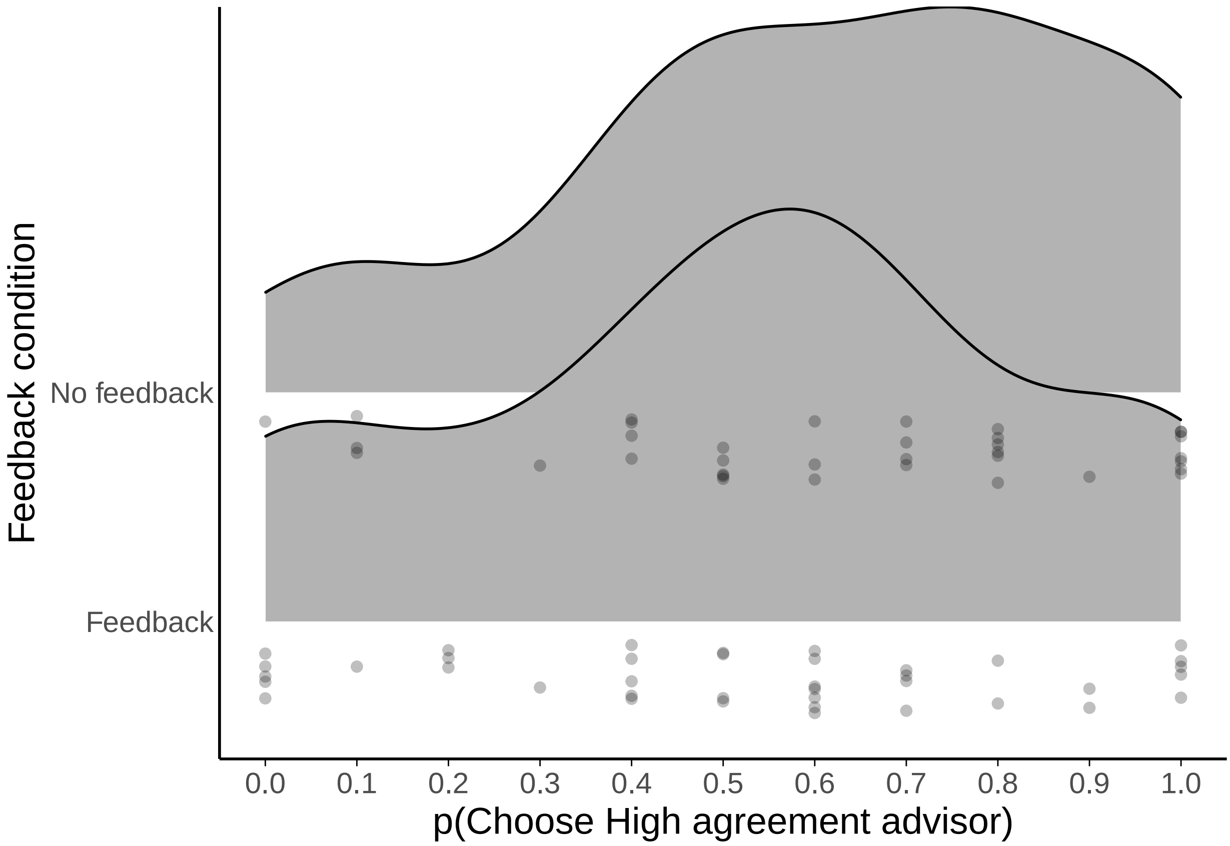

Figure 4.4: Participants' pick rate for the advisors in the Choice phase of the experiment. The shaded area shows a density plot of the individual participants' pick rates, shown by semi-transparent dots. Participants in the Feedback condition received feedback during the Familiarization phase, but not during the Choice phase.

4.2.3 Discussion

Where feedback is provided on advisors' performance, participants prefer high agreement advisors to low agreement advisors.

4.3 Accuracy vs.

agreement (Date estimation) {#ac-ava}

4.3.1 Method

4.3.2 Result

Dots Task

Dates Task

4.3.3 Discussion

!TODO[This is really the critical test - can we show the same crossover pattern with feedback that we saw in Taking Advice]

4.4 Confidence-contingent advice

4.4.1 Online study

Method

Result

Dots Task

Dates Task

4.4.2 Lab study

`https://aspredicted.org/ze3tn.pdf`

`https://github.com/mjaquiery/nofeedback_trust`

!TODO[Use a sensible archive format for this study data, archive on OSF, and produce data dictionary]

Pescetelli et al. (2018) showed that, in the absence of objective feedback, advice was more influential coming from an advisor who agrees with a participant when that participant is confident (*Bias Sharing*) than coming from an advisor who agrees with a participant when that participant is unconfident (*Anti Bias*). This provides evidence of a metacognitive sensitivity in the tracking of advice and the updating of advisor utility. Here we investigate whether these effects show up in the domain of advisor selection.

The literature on information exposure and evaluation indicates that people evaluate more favourably information which agrees with their currently-held opinion !TODO[REF], and preferentially seek out information sources which are likely to provide information which agrees with their currently-held opinion (Garrett 2009a; Sears and Freedman 1967). If this holds true in the context of the judge-advisor system, advice from *Bias Sharing* advisers ought to be evaluated more favourably (influence should increase) and should be sought more frequently. Given the evidence in favour of the first of these propositions, we here investigate the latter: given a choice, will judges prefer to receive advice from a *Bias Sharing* advisor over receiving advice from an advisor who does not share the judge's bias?

Pescetelli et al. (2018) used a judge-advisor system to demonstrate that judges are influenced to a greater extent by advisers who share their biases. Participants played the role of judge in a judge-advisor system, while the advisers were virtual agents whose advice-giving was dependent upon the confidence and correctness of the judges' initial decisions. The advisers were balanced for overall agreement with the judge and objective correctness of advice. We place participants in a similar paradigm in which they are given a choice between advisers, and hypothesise that they will more frequently seek advice from the *Bias Sharing* advisor than from the *Anti Bias* advisor.

Method

Participants 24 participants ($M_{age} = 22 \pm SD\ 4.7$, 5 male, 19 female, 0 other) recruited from the University of Oxford participant recruitment platforms took part in the experiment. An additional 2 participants attended experimental sessions but their data were not analysed. Participants were compensated for their time with either course credit for a psychology degree, or 10GBP. ##### Procedure {#ex1-m-procedure}

The experiment consisted of a judge-advisor system with a perceptual decision task (Figure 4.5). Participants played the role of the judge, and the advisers were played by virtual agents whose answers depended upon the confidence with which the judge reported the initial decision. In the majority of trials (92%), participants were offered advice from virtual advisers. In one third of these trials ('choice trials'), participants chose which advisor to receive advice from by clicking on their respective portraits appearing at the top and bottom of the screen. On the remaining two thirds of trials ('forced trials'), participants were forced to take advice from one of the two advisers (equiprobably). On these trials, the forced advisor's portrait appeared at the top or bottom of the screen, with a red cross appearing in the other location, which was not selectable. On the remaining 8% of trials, participants received no advice and were given no opportunity to revise their initial decision. These 'catch trials' were included to encourage participants to attend to the initial decisions.

Each participant completed 363 trials (51 practice trials over 2 blocks + 12 x 26-trial experimental blocks) in which they had to identify the box with the most dots (Figure 4.5). The difficulty of the task was continually adjusted throughout the experiment using a 2-down, 1-up staircase procedure to keep the participant's initial decision accuracy at 72%. At the end of each block, participants were notified as to their (final decision) accuracy in the block and given the opportunity to rest for as long as they wished. Throughout the experiment a progress bar provided a graphical indication of the number of trials remaining in the experiment. After each block participants were told what percentage of the (final) answers they had provided were correct and allowed to take a short, self-paced break. Prior to the first experimental block, after the final experimental block, and after the 4th and 8th experimental blocks, participants were presented with a questionnaire (Figure 4.6). The questionnaire contained 4 questions for each advisor. The questions asked for the judge's assessment of the advisor's likeability, trustworthiness, influence, and ability to do the task. The questions presented before the first experimental block were worded prospectively (e.g. 'How much are you going to like this person?' as opposed to 'How much do you like this person?'). Answers were provided by

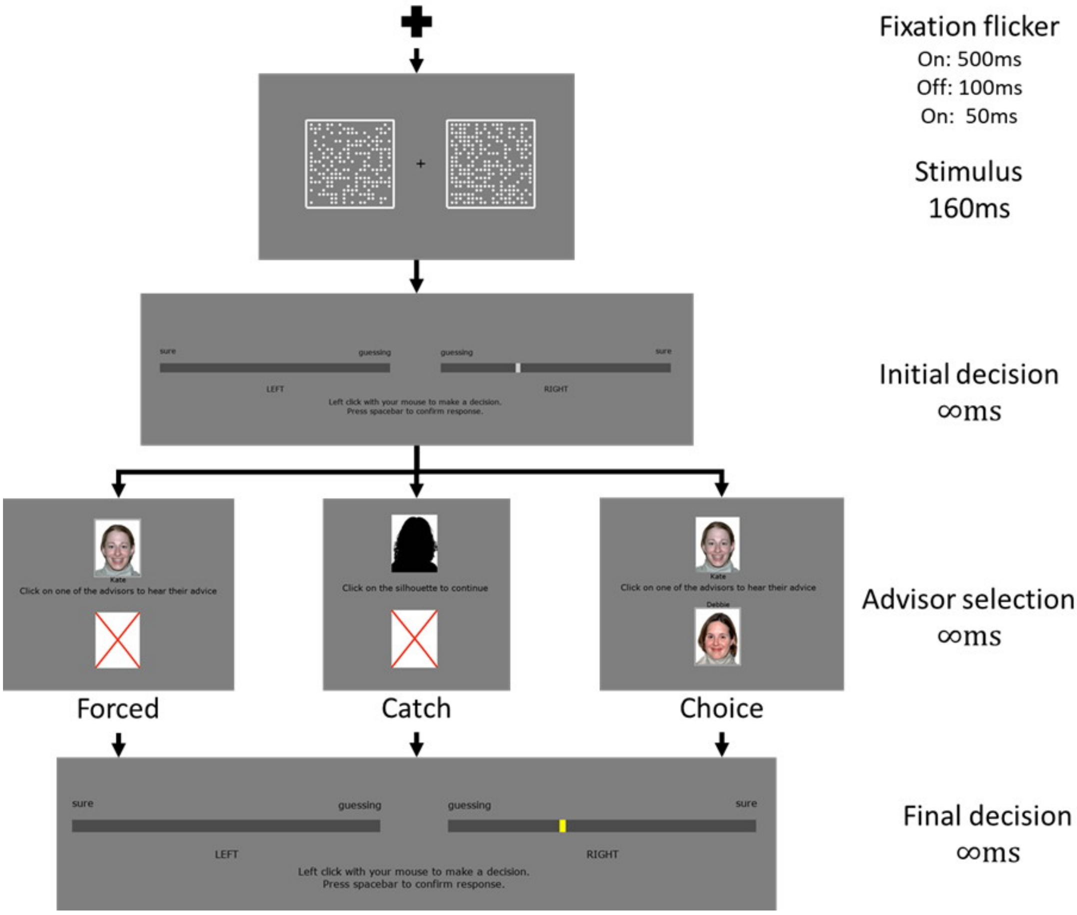

Figure 4.5: Experiment 1 procedure.
The task began with a blank screen containing only a fixation cross and progress bar. Momentarily prior to the onset of the stimuli the fixation cross flickered. The stimuli, two rectangles containing approximately 200 dots each, appeared for 0.16s, one on either side of the fixation cross. Once the stimuli disappeared, a response-collection screen appeared and prompted the participant to indicate their initial decision and its confidence by selecting a point within one of two regions. The left region indicated a decision that the target was on the left, and increasingly-leftwards points within that region indicated increasing confidence in that decision. The right region indicated a decision that the target was on the right, and increasingly-rightwards points within that region indicated increasing confidence in that decision.
Next, the participant was presented with a choice screen. The choice screen displayed two images, one at the top of the screen and one at the bottom. The images were one of the following: an advisor portrait, a silhouette, or a red cross. The red cross was not selectable, forcing participants to choose the other option. The silhouette offered no advice, and was only ever offered as a forced choice. Selecting an advisor image provided the participant with the opinion of that advisor on the trial.
Having heard the advice, the participant was again presented with the response-collection screen, with a yellow indicator marking their original response. A second (final) judgement was collected using this screen (except on catch trials), and the trial concluded.

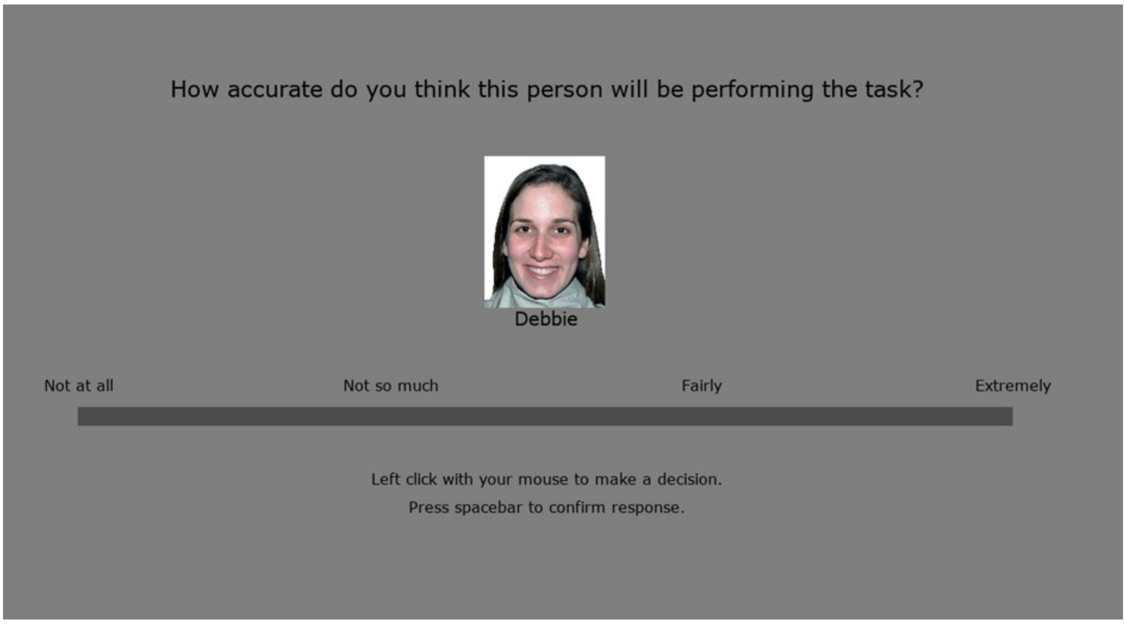

Figure 4.6: Experiment 1 advisor questionnaire.
Participants rated advisors on a number of different dimensions.

moving a sliding scale below the advisor's portrait towards the right for more favourable responses (marked 'extremely') or towards the left for less favourable responses (marked 'not at all').

Each participant attended the experiment individually, was welcomed and briefed on the experimental procedure, and had their informed consent recorded, before the experiment began. They were seated a comfortable distance in front of a 24' (1440x900 resolution) computer screen in a small, quiet, and dimly-lit room. The experiment took place wholly on the computer, and lasted around 45 minutes.

The experiment was programmed in MATLAB R2017b (*MATLAB* 2017) using the Psychtoolbox-3 package (Kleiner et al. 2007).

Advisor advice profiles The advisers are virtual agents whose probability of agreeing with the participant's decision varies as a function of the participant's confidence and correctness in the initial decision phase. Table 4.3 illustrates how this relationship functions, and shows that the overall correctness and agreement rates of the advisers is equivalent overall. Importantly, on largest minority of trials, the middle 40%, the advisers are exactly equivalent, meaning these trials

Table 4.3: Table : Experiment 1 advisor advice profiles

	Initial decision confidence	Probability of agreement (%)	
		Bias Sharing	Anti Bias
Participant correct	High (top 30%)	80	60
	Medium (middle 40%)	70	70
	Low (bottom 30%)	60	80
Participant incorrect	Any	30	30
Total agreement	Participant correct	70	70
	Participant incorrect	30	30

can be compared directly without confounds arising from agreement rate and initial confidence.

Analysis Data analysis was performed using R (R Core Team 2018). For a full list of packages and software environment information, see !TODO[figure out where to include this stuff. Appendix? Also link to a containerized version of this.]

Bayes Factors (BF) are presented alongside p values and test statistics. A BF < 0.33 indicates decisive evidence in favour of the null hypothesis over the alternative hypothesis (with lower values being increasingly clear), BF > 3 indicates decisive evidence of the alternative over the null (with higher values being increasingly clear), and 0.33 <= BF <= 3 indicates there is insufficient evidence to reach a conclusion.

Capped influence Influence, the dependant variable in some analyses, is calculated as the extent to which the judge's initial decision is revised in the direction of the advisor's advice. The initial (C_1) and final (C_2) decisions are made on a scale stretching from -55 to +55 with zero excluded, where values <0 indicate a 'left' decision and values >0 indicate a 'right' decision, and greater magnitudes indicate increased confidence. Influence (I) is given for agreement trials by the shift towards the advice:

$$I|\text{agree} = f(C_1) \begin{cases} C_2 - C_1 & C_1 > 0 \\ -C_2 + C_1 & C_1 < 0 \end{cases} \tag{4.1}$$

And by the inverse of this for disagreement trials:

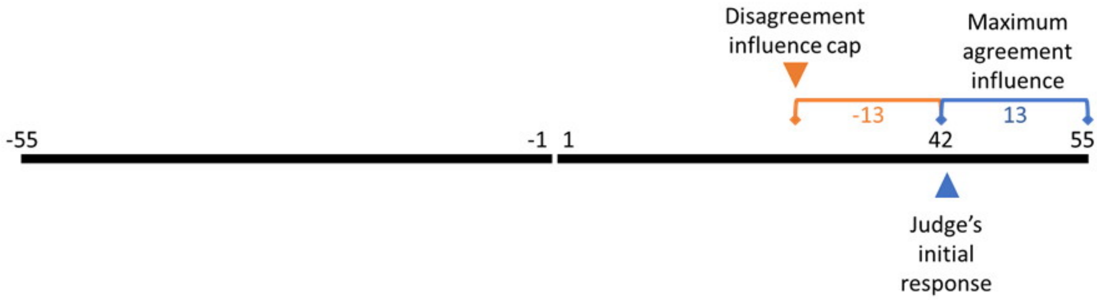

Figure 4.7: Capping influence to avoid scale bias.
In this example the judge's initial response is 42, meaning that their final decision could be up to 13 points more confident or up to 97 points less confident. Any final decision which is more than 13 points less confident is therefore capped at 13 points less confident.

$$I|\text{disagree} = -I|\text{agree} \tag{4.2}$$

The confidence scale excludes 0, and thus the final decision can always be more extreme when moving against the direction of the initial answer than when moving further in the direction of the initial answer. A capped measure of influence was used to minimise biases arising from the natural asymmetry of the scale. This measure was calculated by truncating absolute influence values which were greater than the maximum influence which could have obtained had the final decision been a maximal response in the direction of the initial answer (Figure 4.7).

The capped influence measure I_{capped} is obtained by:

$$I_{\text{capped}} = f(C_1)\begin{cases}\min(I, 2C_1 - 55) & C_1 > 0\\ \max(I, 2C_1 + 55) & C_1 < 0\end{cases} \tag{4.3}$$

The explicit measure of trust is obtained using questionnaires. The questionnaires are delivered at 4 time points, and consist of 4 questions per advisor which are answered on a 1-100 scale.

Result

Descriptive statistics 26 participants took part in the study. One participant was unable to complete the experiment due to technical difficulties. Pre-registration of the study analysis stated that data would be collected from 24 participants, so the final (overbooked) subject tested was excluded from analysis. Descriptive statistics for the 24 participants included in the analysis are presented in Table 4.4.

```
## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'cor2' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'step' (override with `.groups` argument)
## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'step' (override with `.groups` argument)
## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

The descriptive statistics demonstrate the contingent agreement of the advisers, with the Bias Sharing and Anti Bias advisers both agreeing at close to the target rate for most participants. While the ranges overall patterns are as designed, the variation in individual experience means some participants may have experienced by chance an advisor profile which contradicted the generative pattern (e.g. a Bias Sharing advisor who agreed on fewer high-confidence trials than mid-confidence trials). This is especially likely to be true of contingencies with fewer trials per participant,

Table 4.4: Table : Descriptive statistics for Experiment 1

			Target	Mean	95% CI Low	95% CI High
Participant proportion correct	**Initial estimate**	Bias Sharing	.71	.68	.67	.70
		Anti Bias	.71	.71	.69	.72
		Both	**.71**	**.70**	**.68**	**.71**
	Final decision	Bias Sharing	-	.70	.69	.71
		Anti Bias	-	.73	.72	.75
		Both	**-**	**.72**	**.71**	**.72**
Advisor-participant agreement rate	**Bias Sharing advisor**	Low confidence	.80	.60	.57	.63
		Medium confidence	.70	.70	.68	.72
		High confidence	.60	.79	.76	.82
		Initial wrong	.70	.30	.29	.32
		Initial correct	.30	.69	.68	.71
		All	**-**	**.57**	**.56**	**.58**
	Anti Bias advisor	Low confidence	.60	.79	.75	.82
		Medium confidence	.70	.70	.67	.74
		High confidence	.80	.64	.60	.67
		Initial wrong	.70	.32	.29	.34
		Initial wrong	.30	.71	.69	.73
		All	**-**	**.59**	**.57**	**.61**
Mean initial confidence	**Initial judgement**	Wrong	-	19.1	15.4	22.7
		Correct	-	22.8	19.1	26.5
		Both	**-**	**21.7**	**18.0**	**25.4**
Mean final confidence	**Final decision**	Wrong	-	18.2	14.8	21.6
		Correct	-	23.6	20.0	27.1
		Both	**-**	**22.0**	**18.6**	**25.5**
	Advisor agrees	Bias Sharing	-	26.1	22.6	29.6
		Anti Bias	-	24.6	21.2	28.1
		Both	**-**	**22.0**	**18.6**	**25.5**
	Advisor disagrees	Bias Sharing	-	16.6	12.9	20.2
		Anti Bias	-	18.3	14.4	22.3
		Both	**-**	**22.0**	**18.6**	**25.5**

such as the incorrect trials. Overall, however, both the mean and 95% confidence intervals suggest the pattern was as desired for most participants most of the time.

Participants' revisions to their confidence were mostly in the direction of the advice, symmetrical across left and right responses, and usually relatively small, especially in agreement trials (Figure 4.8).

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

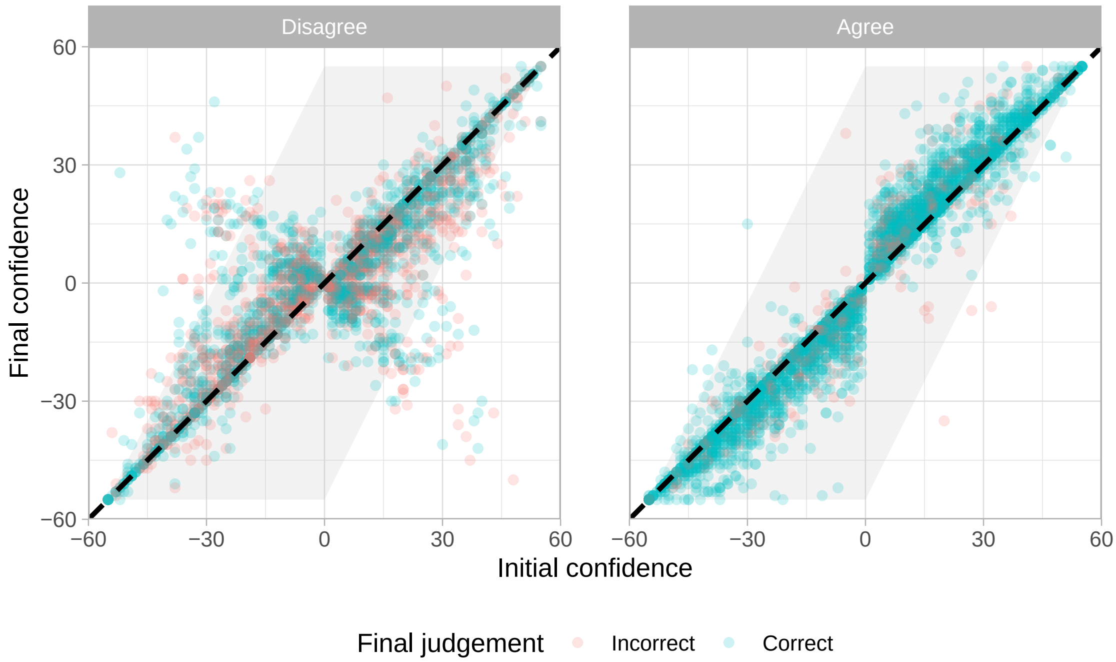

Figure 4.8: Initial vs final confidence.
Influence of the advisors is evident in the deviation from the dashed y = x line. Points lying below the line indicate a more leftward response from initial to final judgement. Points above the line indicate a more rightward response in the final judgement. The further away from the y = x line, the greater the change from initial to final judgement. Separate plots show agreement vs disagreement trials (between the advisor and judge), and separate colours indicate whether the judge's final decision was correct or incorrect. The shaded area indicates the boundary for the symmetrical influence measure. Points outside this area are truncated by moving them vertically until they meet the grey area.

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

Advisor selection We hypothesised that the participants would display different pick rates for the Bias Sharing advisor versus the Anti Bias advisor. This hypothesis was evaluated by calculating the proportion of choice trials on which each participant picked the Bias Sharing advisor, and then testing these values as a one-sample t-test against the null hypothesis that the pick rates would be 0.5. No support was found for this hypothesis (; Figure 4.9), although the Bayesian test indicated that the data were not sufficient to conclude that no effect was present.

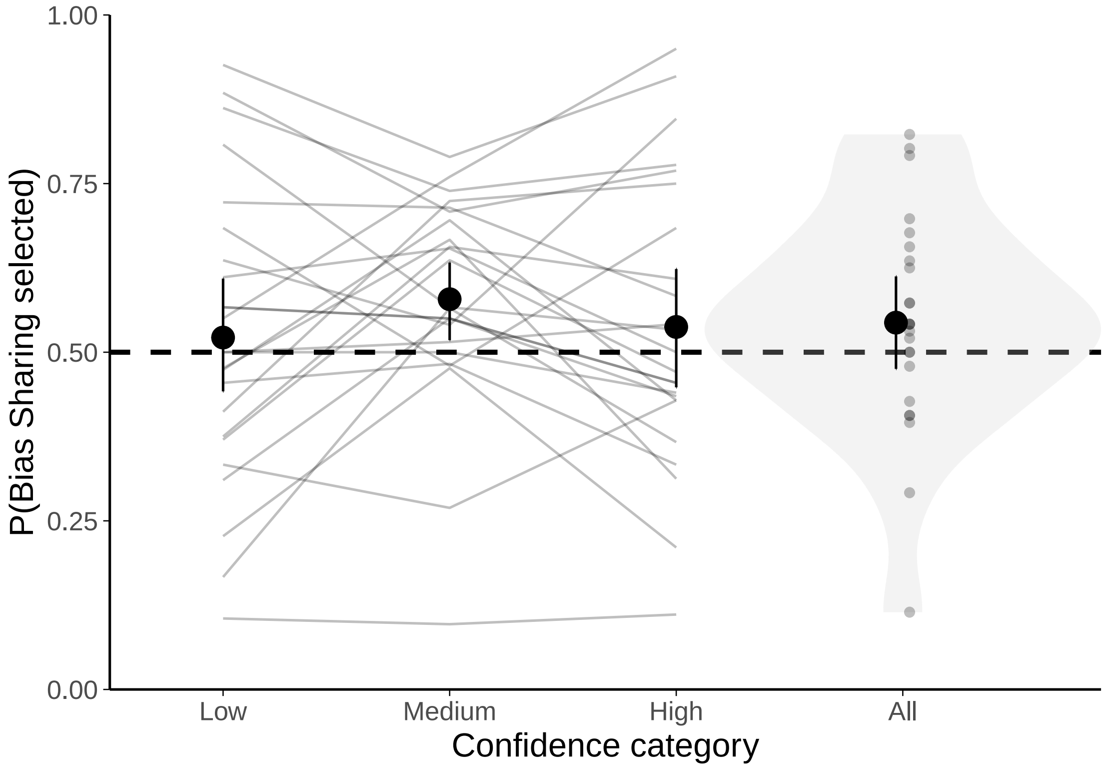

Figure 4.9: Advisor selection.
Proportion of the time each participant picked the Bias Sharing advisor. Faint lines and dots indicate data from individual participants, while the large dot indicates the mean proportion across all participants. The dashed reference line indicates picking both advisors equally, as would be expected by chance. Error bars give 95% confidence intervals.

There was considerable variability across participants in the overall pick rate for the Bias Sharing advisor (range = [.11, .82]).

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

Advisor selection on medium-confidence trials The advisers differed in their advice-giving as a function of the judge's initial confidence. In trials where the judge's initial decision was made with medium confidence, however, the advisers were equal on judge confidence and agreement rate. Comparing selection rates for these trials alone revealed a clear preference for the Bias Sharing advisor (; Figure

Table 4.5: Table : ANOVA of Advisor influence in Experiment 1

Effect	$F(1, 23)$	p		η^2
AiC	0.28	.602		.001
agree	13.88	.001	*	.175
hasChoice	4.23	.051		.001
AiC:agree	0.75	.395		.001
AiC:hasChoice	0.04	.842		$< .001$
agree:hasChoice	1.99	.172		$< .001$
AiC:agree:hasChoice	0.01	.935		$< .001$

Degrees of freedom: 1, 23

4.9 "Medium" confidence category), although the Bayesian analysis again indicated an insensitive result, albeit in the hypothesised direction.

Advisor influence Previous work in our lab demonstrated that the agree-in-confidence advisor exerted greater influence on the judges' final decisions than the agree-in-uncertainty advisor (Pescetelli 2017). Influence was examined with a 2x2x2 (Bias Sharing versus Anti Bias advisor; choice versus forced trials; agreement versus disagreement trials) ANOVA (Figure 4.10). No main effect was found for advisor !TODO[stats], meaning that the previous finding was not replicated. As shown in Table 4.5, the only statistically significant effect was the main effect of agreement, with disagreement producing higher influence than agreement !TODO[marginal means].

```
## Warning: 'fun.y' is deprecated. Use 'fun' instead.

## Warning: 'fun.y' is deprecated. Use 'fun' instead.
```

Advisor influence on medium confidence trials The agree-in-confidence and agree-in-uncertainty advisers differed by design in the frequency with which they agree with the participant as a function of the participant's confidence in their initial estimate. To control for the effects of initial confidence on influence, the above analysis was repeated using only those trials on which the initial estimate was correct and given with medium confidence. Two participants were missing

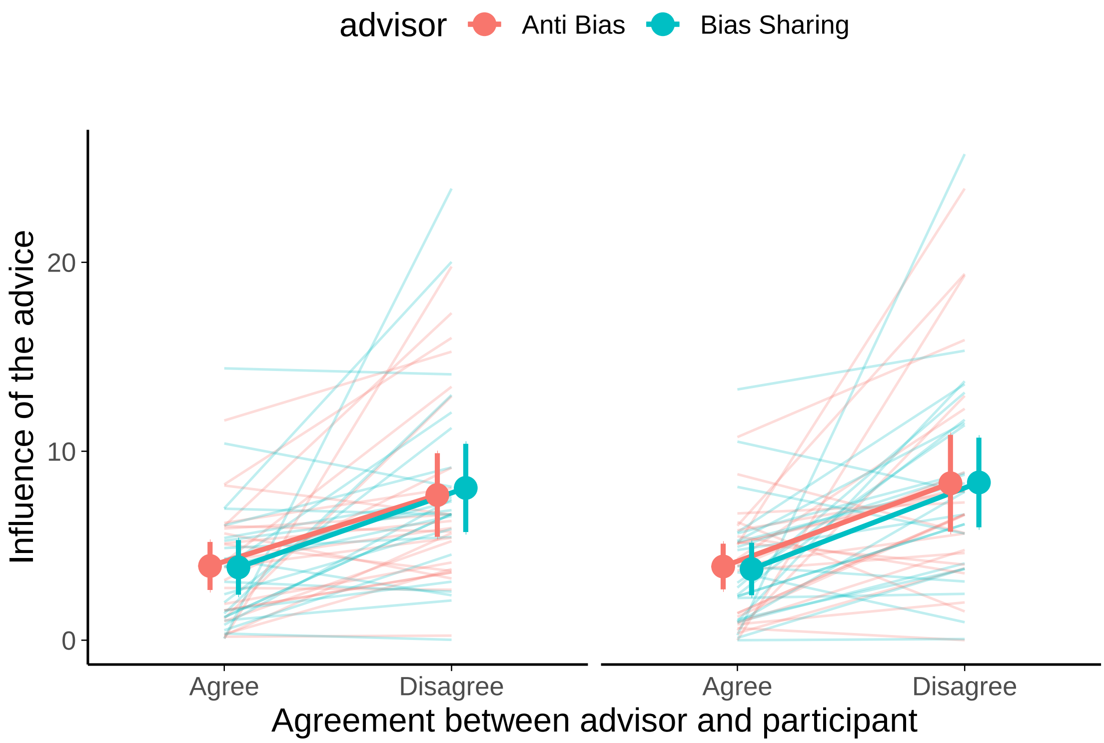

Figure 4.10: Advisor influence.
Influence of advice from each advisor by advisor, agreement, and trial type. Faint lines and indicate data from individual participants, while the dots indicate the mean proportion across all participants. Error bars give 95% confidence intervals.
Note: vertical axis is truncated to show group differences more clearly, the theoretical maximum influence given the scale is 110. The minimum is 0 as shown.

data in this analysis: one participant had zero mid-confidence choice trials in which the agree-in-confidence advisor disagreed with them; and the other had zero mid-confidence choice trials in which the agree-in-uncertainty advisor disagreed with them. These participants were removed from this analysis. As before, the only effect which was statistically significant was agreement. The low number of trials in some of the intersections meant the some participants had to be dropped. The analysis was rerun with forced trials only, !TODO[stats]

Subjective assessment of advisers Participants answered questionnaires about their trust in the advisers at four time points during the experiment. We hypothesised that this subjective trust measure would change over the course of the experiment, with the agree-in-confidence advisor becoming more preferred over time.

Table 4.6: Table : ANOVA of Advisor influence on medium confidence trials in Experiment 1

Effect	$F(1, 21)$	p		η^2
AiC	0.88	.358		.005
agree	9.32	.006	*	.119
hasChoice	0.50	.485		.001
AiC:agree	3.06	.095		.010
AiC:hasChoice	0.64	.433		.002
agree:hasChoice	0.21	.650		.001
AiC:agree:hasChoice	0.01	.919		$< .001$

Degrees of freedom: 1, 21

As indicated by Table @ref{tab:ex1-subjective-assessment}, however, no such effects were found: subjective assessments of the advisers did not differ post-experiment.

```
##
## Attaching package: 'broom'
## The following object is masked from 'package:modelr':
##
##     bootstrap
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(test, bf))`, with `mutate()` if needed
## Usually it is recommended to use column_spec before collapse_rows, especially ir
```

Sensitivity to the manipulation Finally, we planned to investigate the hypothesis that participants' choice of advisor would be sensitive to the differential agreement strategies of the advisers, e.g. participants might preferentially select the advisor with the greater likelihood of agreement given their initial confidence. This was investigated by testing the participants' mean bias sharing advisor pick rate in low- versus high-confidence trials. Pick rates did not differ (); Figure @ref{fig:ex1-selection}).

Exploratory analyses Below are reported analyses which were not part of the preregistration, but which were pursued to gain a greater insight into the behaviour

Table 4.7: Table : Questionnaire responses pre- and post-experiment

		Bias sharing			Anti bias					
			95% CI			95% CI				
	Question[a]	Mean	Low	High	Mean	Low	High	BF	$t(23)$	p
Pre-experiment	How **accurate** do you think this person was when performing the task?	61.0	56.9	65.1	59.9	55.5	64.4	0.30	1.1	.536
	How much are you **influenced** by the opinions of this person?	51.8	45.1	58.4	50.4	43.5	57.2	0.30	1.4	.642
	How much do you **like** this person?	60.5	56.0	64.9	61.3	56.1	66.6	0.30	-0.9	.613
	How **trustworthy** are the opinions of this person?	61.1	56.1	66.1	61.7	57.4	66.0	0.29	-0.6	.770
Post-experiment	How **accurate** do you think this person was when performing the task?	56.9	49.1	64.6	55.0	46.9	63.0	0.30	1.9	.607
	How much are you **influenced** by the opinions of this person?	54.0	44.7	63.4	53.3	43.1	63.6	0.29	0.7	.831
	How much do you **like** this person?	57.2	48.9	65.5	58.7	48.8	68.5	0.29	-1.5	.529
	How **trustworthy** are the opinions of this person?	57.2	50.5	64.0	56.0	47.8	64.2	0.29	1.3	.700

[a] Emphasis added.

Table 4.8: Table : Linear regression of pick rate in later blocks by initial agreement difference

Effect	Estimate	SE	*t*	*p*	
(Intercept)	0.54	0.03	16.92	< .001	*
agreeRateDifference.block3	0.37	0.17	2.27	.034	*

Model fit: F(5.1, 1) = 22; p = .034; R^2_{adj} = .152

Table 4.9: Table : Linear regression of pick rate in later blocks by initial agreement difference and preference

Effect	Estimate	SE	*t*	*p*	
(Intercept)	0.41	0.06	6.41	< .001	*
agreeRateDifference.block3	1.19	0.28	4.29	< .001	*
aicPickRate.block3	0.27	0.12	2.27	.034	*
agreeRateDifference.block3:aicPickRate.block3	-2.43	0.72	-3.37	.003	*

Model fit: F(7.6, 3) = 20; p = .001; R^2_{adj} = .461

of participants in the experiment.

Effect of initial agreement The hypothesised effect of the different advice-giving profiles of the advisers on their pick rates was not found, and we hypothesised that initial exposure to the advisers may have overshadowed information in subsequent blocks. To investigate this possibility, we examined the effect that agreement in the first experimental block had upon choices throughout the rest of the experiment (Figure @ref{fig:ex1-initial-agreement}). A simple regression (Table @ref{tab:ex1-initial-agreement-t}) indicated that the extent to which the Bias Sharing advisor agreed with the judge more frequently than the Anti Bias advisor in the first experimental block (Block 3) was a significant predictor of the preference for picking the Bias Sharing advisor in subsequent blocks.

The relationship between early experience of advisor agreement and overall advisor preference may be modified by an initial preference for one or the other advisor. We investigated this by adding advisor pick rate into the regression giving:

$$\text{PickRate}_{block>3} = \beta_1 \cdot \text{AgreementDifference}_{block=3} + \beta_2 \cdot \text{PickRate}_{block=3} + \beta_3 \cdot \text{AgreementDifference}_{block=3} \cdot \tag{4.4}$$

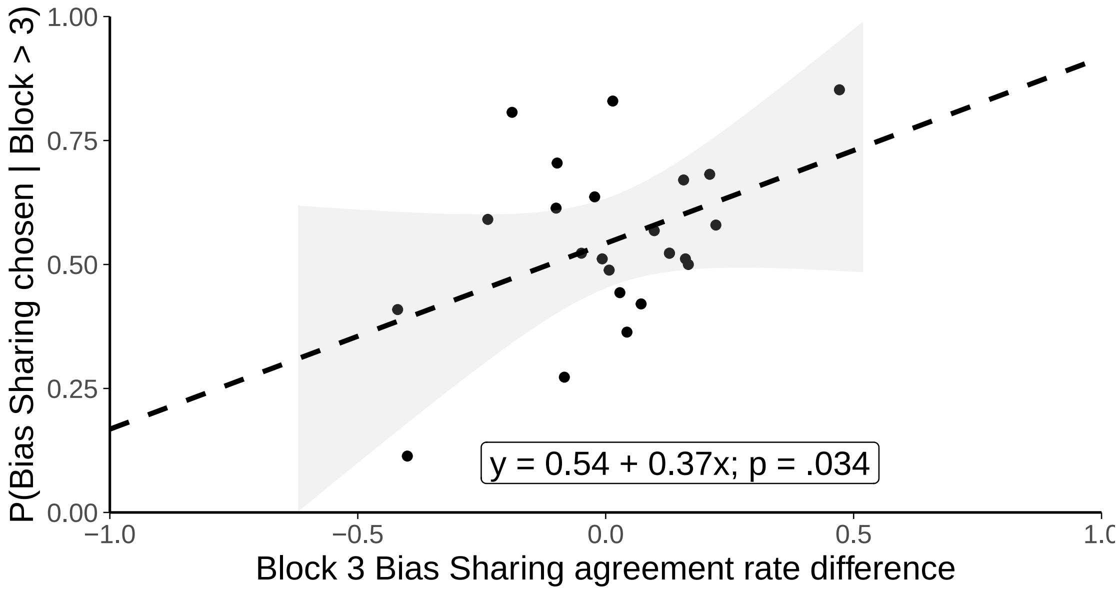

Figure 4.11: Initial agreement and subsequent preference.
figure shows the relationship between the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in later blocks. The dashed line shows the best-fit regression line, with shaded 99% confidence intervals.

This model (Table @ref{tab:ex1-initial-agreement-pick-rate-t}) also fit the data well enough for interpretation, and represented an improvement on the previous model ($F(20,\ 22) = 7.2990236$, $p = .004$). Early agreement difference remained predictive, and relative pick rate in block 3 was also predictive. Finally, the interaction between these predictors was also important to the model, although the negative sign of the interaction beta (β_3 = -2.43) was unexpected, and required further exploration. Figure @ref{fig:ex1-initial-agreement-pick-rate-mm} shows a marginal means plot for the interaction with the predictors collapsed to binary measurements using a mean split.

```
## Warning: ‘fun.y‘ is deprecated. Use ‘fun‘ instead.
```

```
## Warning: ‘fun.y‘ is deprecated. Use ‘fun‘ instead.
```

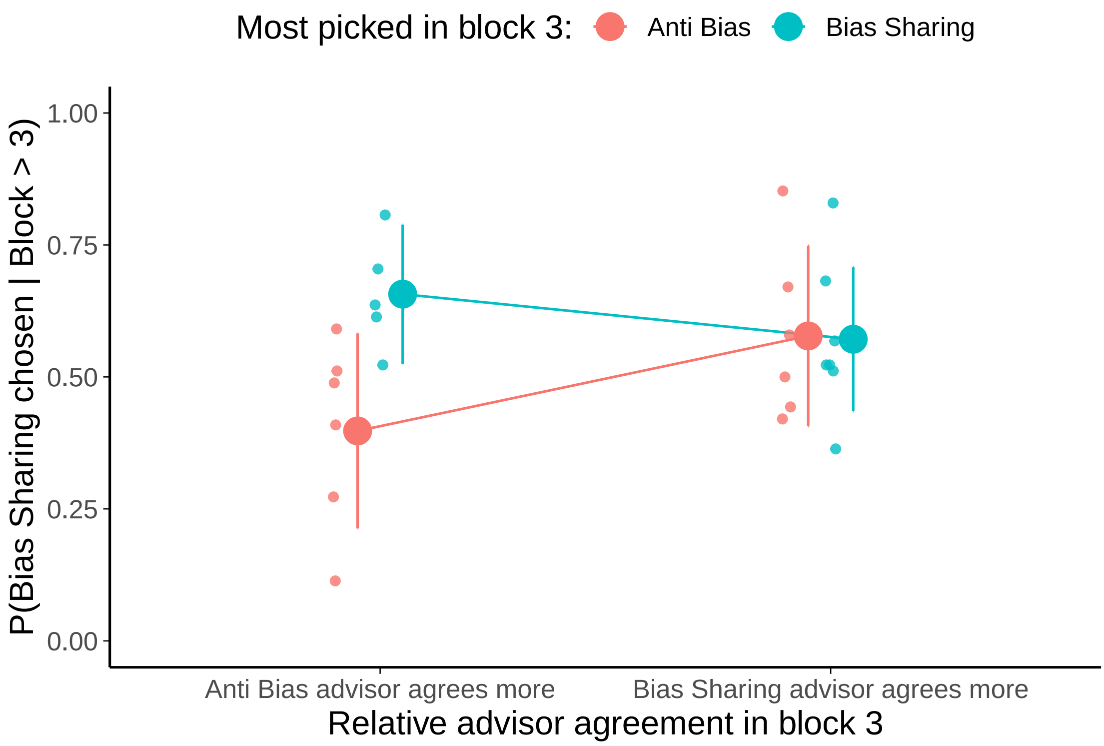

Figure 4.12: Initial agreement and preference predicting subsequent preference. Figure shows the relationship between whether the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in later blocks, split by whether the participant picked the Bias Sharing or Anti Bias advisor more often in the first block. Splits are based on the sample means, and error bars give 95% confidence intervals.

Table 4.10: Table : Linear regression of pick rate in later blocks by initial agreement difference by confidence

Effect	Estimate	SE	*t*	*p*	
(Intercept)	0.43	0.10	4.21	.002	*
agreeRateDifference.highConf.block3	0.02	0.08	0.23	.825	
agreeRateDifference.medConf.block3	0.12	0.13	0.92	.382	
agreeRateDifference.lowConf.block3	0.23	0.11	2.11	.064	
aicPickRate.block3	0.37	0.21	1.74	.116	

Model fit: \$F\$(2.6, 4) = 9; \$p\$ = .107; \$R^2_{adj}\$ = .331

The theoretical mechanism by which judges evaluate advisor accuracy in the absence of feedback is through their own metacognitive awareness. Thus, we would expect that, given preferences appear to be set by initial exposure to the advisors, those preferences would be more affected by high confidence agreement in block 3. This was tested using a regression model in which block 3 agree-in-confidence advisor agreement was used as a predictor split according to the three different levels of confidence. Several participants (5) had no trials in block 3 for one or both advisors at one or more confidence levels; these participants were removed from the analysis. Neither the overall model fit nor any of the predictors in this model (Table @ref{tab:ex1-initial-agreement-metacog}) were significant, but the low-confidence beta was substantially larger than the high- and medium-confidence betas.

A prediction of the high weighting of initial exposure is that this weighting should drop off over time. This does not appear to be the case. Figure @ref{fig:ex1-autoprediction-by-block} shows the regression coefficient for Bias Sharing advisor agreement in each block against the Bias Sharing overall pick proportion: no significant negative correlation was found between block number and the regression coefficient for relative agreement in that block predicting overall advisor preference.

```
## `geom_smooth()` using formula 'y ~ x'
```

Capped influence measure

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Preregistered analyses indicated that advisors are more influential when they disagree with the judge (Advisor influence). This result may simply be an artefact of the scale: although the scale extends equally far in both directions, there is necessarily potential for adjusting answers to follow agreeing advice by virtue of the fact that the decision is not placed in the middle of the scale. To redress this balance, a capped measure of influence (Method - Analysis - Capped influence) was used. Of the 6912 trials on which judges received advice, 193 trials (2.8%) had their influence scores capped by this process. These trials were predominantly disagreement

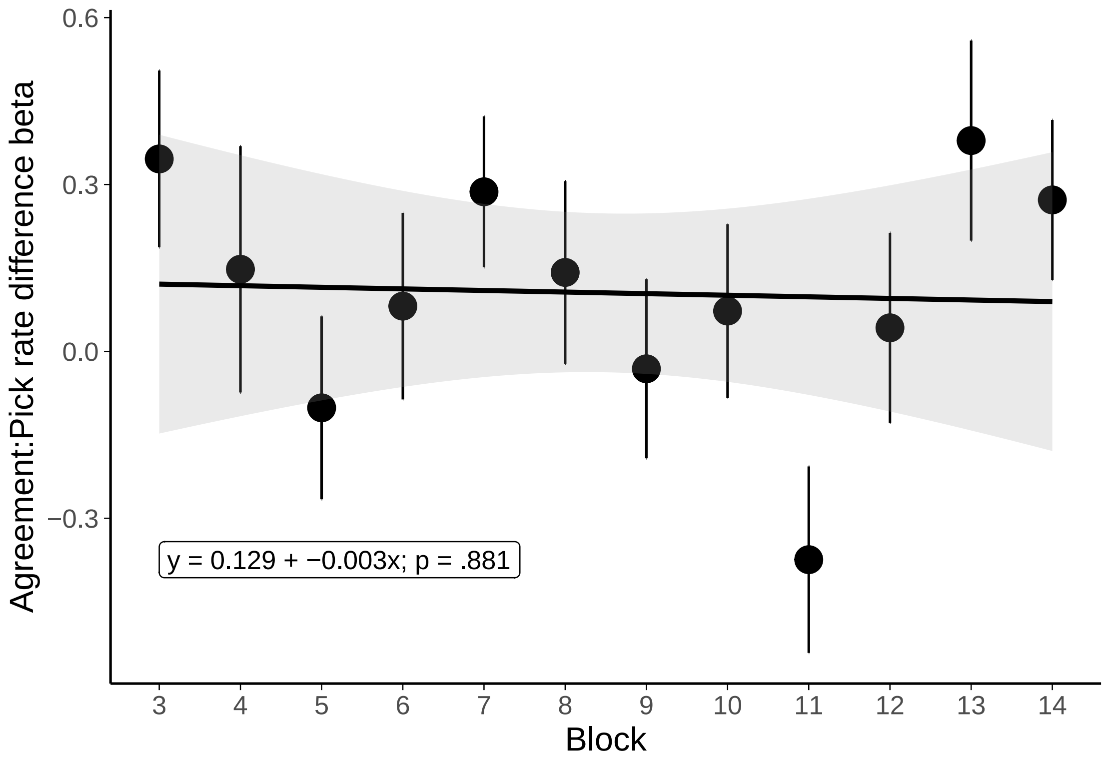

Figure 4.13: Block 3 agreement predicting pick rate in each block.
Figure shows the relationship between the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in each later block. Error bars give +/-1 standard error. Shaded area gives 95% confidence for the overall regression line.

trials[2] (178, 92.2%). The mean number of trials adjusted for each participant was $M_{TrialsAdjusted} = 8.04$ [2.55, 13.54], though the distribution was highly positively skewed and 10 participants had no trials adjusted this way (Figure @ref{fig:ex1-influence-cap-rates}). Analysis using the 2x2x2 (advice type/agreement/choice) ANOVA described above (Advisor influence) no longer showed a significant effect of agreement. As before, neither interactions nor other main effects were significant (Table @ref{tab:ex1-influence-cap-rates-t}).

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

Subjective assessment and choice of advisors We did not observe the hypothesized preference for the agree-in-confidence advisor, either as measured by

[2] Agreement trials can have capped values if the final decision goes in the opposite direction to the advice.

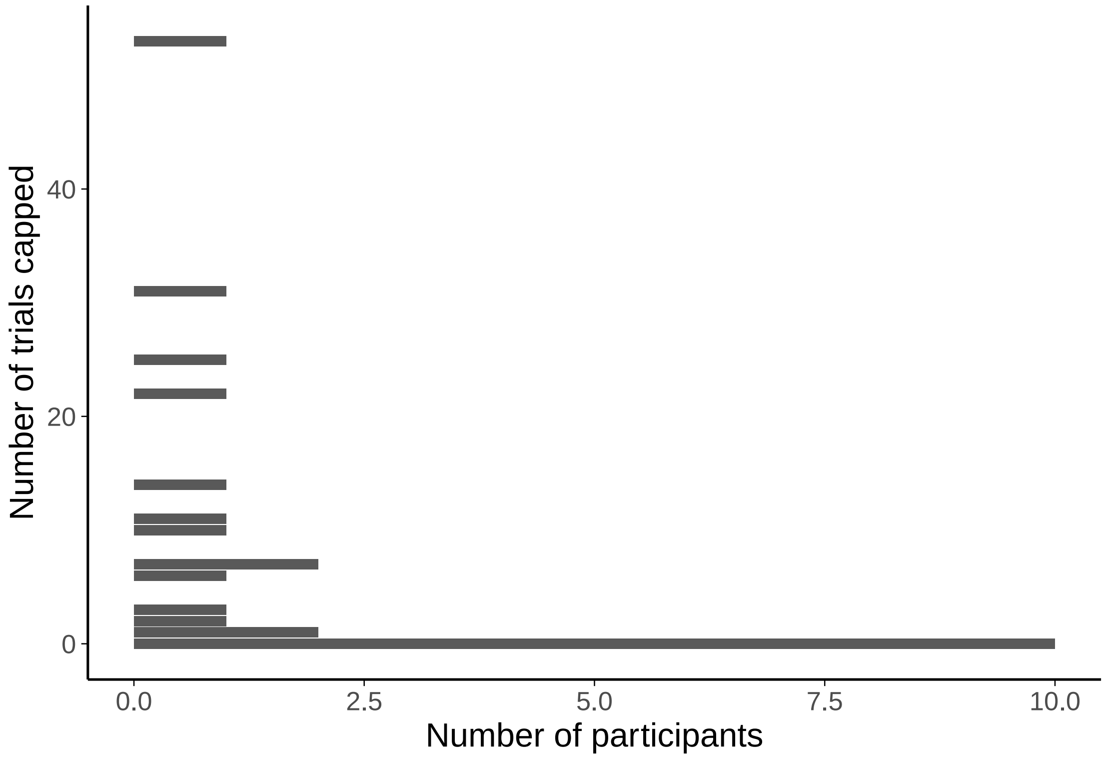

Figure 4.14: Advisor influence capping frequency.
Histogram showing the number of trials in which influence was capped for each participant.

Table 4.11: Table : ANOVA of Advisor influence with capped values in Experiment 1

Effect	$F(1, 23)$	p		η^2
adviceType	3.67	.068		.010
agree	0.30	.588		.008
hasChoice	3.19	.087		.003
adviceType:agree	2.44	.132		.013
adviceType:hasChoice	0.70	.411		.002
agree:hasChoice	0.71	.407		.001
adviceType:agree:hasChoice	0.21	.649		$< .001$

Degrees of freedom: 1, 23

pick rate or by questionnaire response. Nevertheless, it was possible to examine the relationship between the behavioural measure of pick rate and the self-reported questionnaire answers by examining the extent to which participants who picked the agree-in-confidence advisor more frequently rated that advisor more favourably. Questionnaire answers did not evolve systematically over the duration of the experiment, so the analysis was conducted on the final (post-experiment)

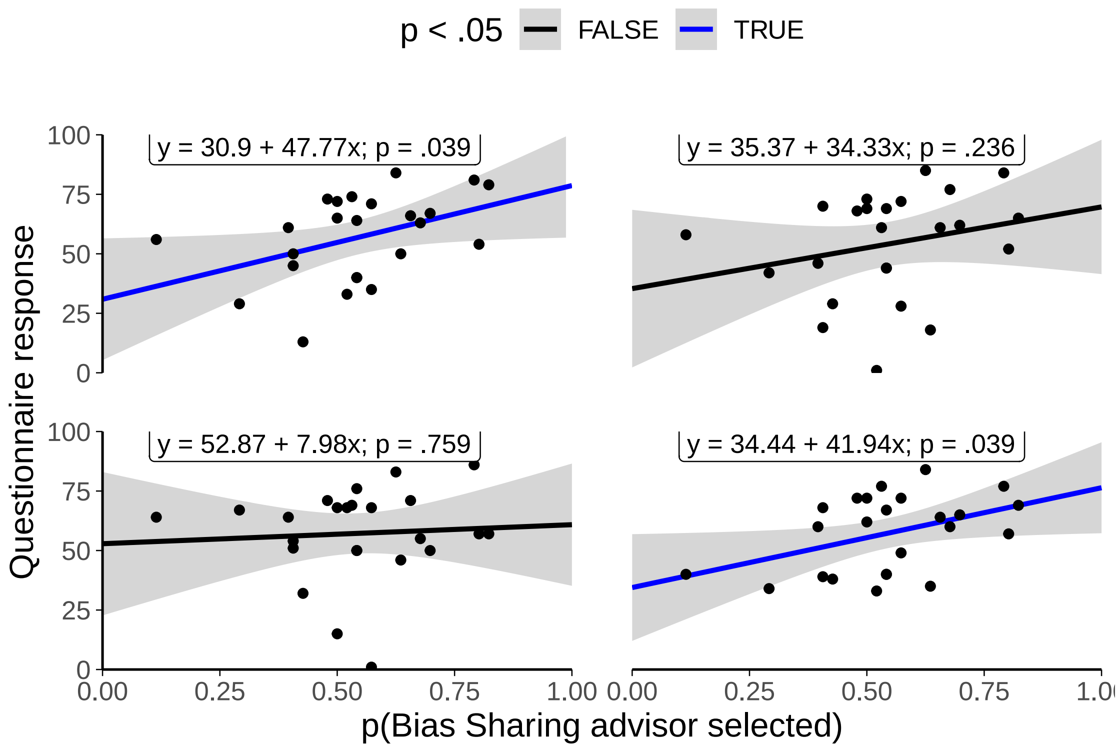

Figure 4.15: Behavioural and self-report consistency.
Relationship between questionnaire response scores on each item and overall pick rate for the Bias Sharing advisor. Lines show best-fit linear model and shaded areas give 95% confidence intervals for the parameters.

answers. Correlations (Figure @ref{fig:ex1-questionnaire-choice}) indicated that the hypothesized effect was present for the questionnaire dimensions of accuracy and trustworthiness, but not for influence or likeability.

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `geom_smooth()` using formula 'y ~ x'
```

We supported this raw score analysis by comparing the pick rate for the Bias Sharing advisor (which is by construction comparative with the Anti Bias advisor) to a comparative version of the questionnaire responses obtained by subtracting scores for the Bias Sharing advisor from the equivalent score for the Anti Bias advisor. The results were highly similar (Figure @ref{fig:ex1-questionnaire-choice-balanced}), with the explicit trust measures for the Bias Sharing advisor relative to the Anti Bias advisor correlating with the picking preference for the Bias Sharing

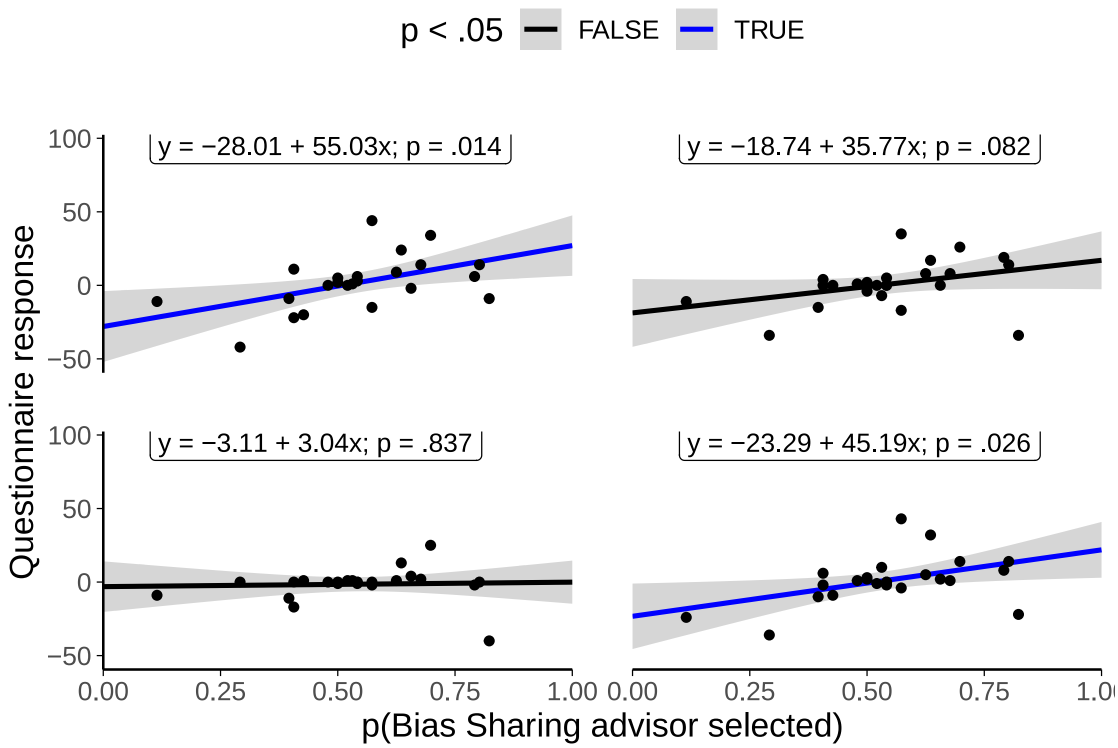

Figure 4.16: Behavioural and self-report consistency.
Relationship between questionnaire response scores on each item for the Bias Sharing advisor minus scores for the Anti Bias advisor and overall pick rate for the Bias Sharing advisor. Lines show best-fit linear model and shaded areas give 95% confidence intervals for the parameters.

advisor in the dimensions of accuracy and trustworthiness. Again, neither influence nor likeability showed the effect.

```
## `summarise()` ungrouping output (override with `.groups` argument)

## `geom_smooth()` using formula 'y ~ x'

## `summarise()` regrouping output by 'participantId', 'advisorId' (override with `
```

There was a strong correlation between trustworthiness and accuracy ($r = .894$), and thus it was probable that the variance in the questionnaire responses explained by advisor preference was principally driven by this shared component. This was borne out by nested multiple regression Table @ref{tab:ex1-questionnaires-additive-variance-explained-t}, which demonstrated that a model using accuracy to predict advisor preference was not significantly improved by including trustworthiness.

Table 4.12: Table : Iterative model comparison predicting advisor choice from questionnaire scores

		Model			
		1	2	3	4
(Intercept)	β	0.54	0.54	0.53	0.53
	p	< .001	< .001	< .001	< .001
accurate	β	0.00	0.00	0.00	0.00
	p	.014	.284	.232	.246
trustworthy	β		0.00	0.00	0.00
	p		.808	.587	.603
influential	β			0.00	0.00
	p			.517	.782
likeable	β				0.00
	p				.370
F	df	1 / 22	2 / 21	3 / 20	4 / 19
	F	7.20	3.48	2.40	2.00
	p	.014	.050	.098	.136
R^2_{adj}	R^2_{adj}	.212	.177	.154	.148
	Δ		-.035	-.023	-.007
	F		0.06	0.43	0.84
	p		.811	.519	.370

Subjective and objective measures of influence Measures of influence were obtained in two distinct ways: through a self-report questionnaire and through observation of behaviour. To the extent that participants have an insight into their behaviour, these measures should correlate positively with one another. This proved to be the case (using the post-experiment responses for simplicity) for both the agree-in-confidence advisor and the agree-in-uncertainty advisor, as shown in Figure @ref{fig:ex1-subjective-objective-influence}.

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

4.4.3 Discussion

The data are pretty equivocal here: there’s no particularly solid reason for thinking that participants select Bias Sharing advisors more frequently than Anti Bias advisors.

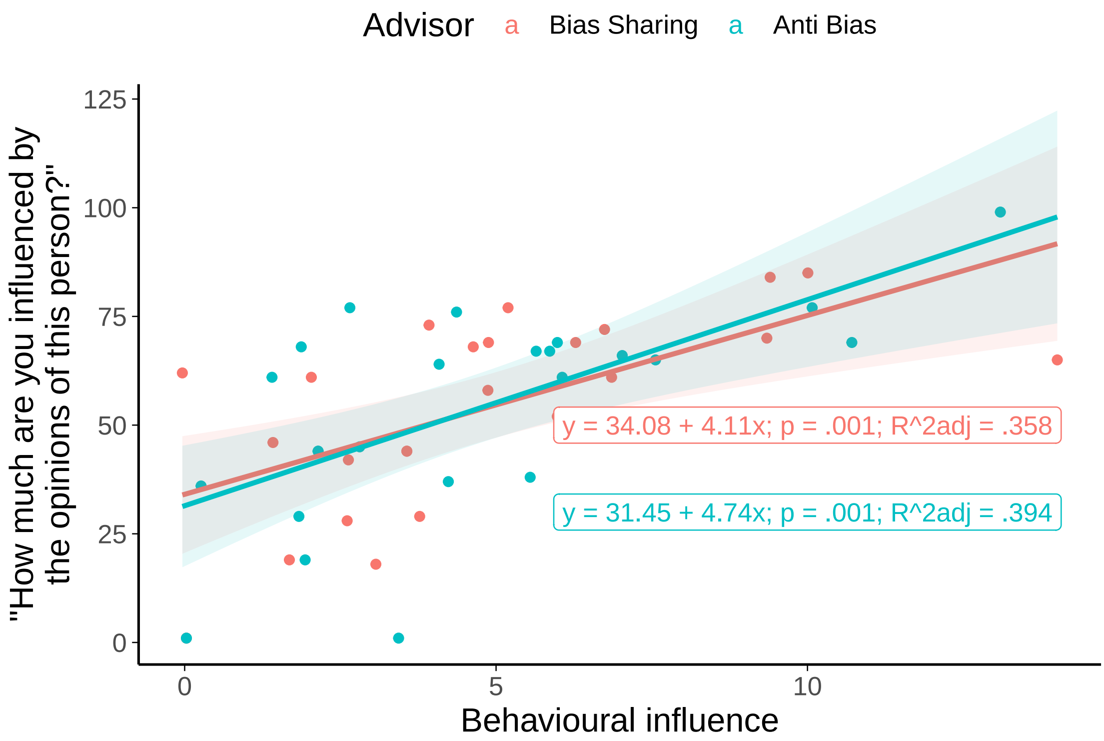

Figure 4.17: Behavioural and self-report measures of influence.
Scatterplot showing correlations between influence as measured by the change in confidence following advice vs self-report.

4.5 General discussion

The patterns observed for advice taking are also evident in advisor selection. Modelling work !TODO[lots of citations for this] indicates that biased source selection can dramatically reshape communication networks and create echo chamber effects where accurate but unpaletable information is ignored. Empirical research on source selection behaviour has found relatively little indication that people behave this way in the real world !TODO[source selection experiments citations], but here we are at least able to demonstrate in principle a psychological mechanism which could drive biased source selection. Furthermore, the mechanisms which produce biased source selection are rational and appropriate given the information available to participants.

5
Psychology of advice-taking

Contents

5.1 Agreement?

!TODO[check Niccolo covered this]

5.2 Accuracy vs Agreement

https://osf.io/fgmdw

!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACv2/index.html

ACv2/coreAnalysis_v0.0.21 ACv2/coreAnalysis_v1.0.1

The study reported here is a preregistered replication of a pilot study which produced the same results. Data, analysis script, and analysis for the pilot study are also available at !TODO[OSFify pilot data/analysis stuff].

5.2.1 Method

5.2.2 Results

5.2.3 Discussion

In consistent with previous studies, these results show that people are more influenced by accurate advice over advice which simply agrees with their own opinions where feedback is available. When feedback is not available, people have only their own opinions against which to validate advice, and are thus more influenced by agreeing advice. Intuitively, we expect the initial estimate to be the participant's "best guess" for the correct answer, and thus the closer the advice is to this guess (i.e. the stronger the agreement), the more likely that advice is to be correct. We deliberately violated an assumption which may be generally true in real life, that the advice is sufficiently independent as to convey at least some information regarding the correct answer. Had we told participants that the agreeing advisor would agree with them no matter what they the participants said, the participants may have disregarded the advice. We have shown that the agreement/accuracy distinction generalises to a continuous estimation decision rather than a categorical perceptual decision. We have illustrated that, even where they would have performed objectively better by preferring accurate over agreeing advice, participants were not able to detect the more accurate advice without objective feedback. These findings support the model of advisor evaluation put forward here.

5.3 Confidence-contingent advice

!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACv2/index.html

ACv2/withConfidence_coreAnalysis_v0.0.1

!TODO[This experiment is kinda crappy and failed its manipulation. Can we run a version which does its job properly and provide actual evidence for/against the confidence modulation?]

5.3.1 Method

5.3.2 Results

5.3.3 Discussion

Participants appeared to pay more attention to the advice than the advisor. In other words, participants distinguished between individual pieces of advice but did not translate these distinctions into distinctions between advisors. This study thus provided no evidence in favour of the confidence-weighting adjustement to the agreement model.

Part III

Context of advice

Contents

6

Context of Advice

6.1 Egocentric discounting

Egocentric discounting (also known as egocentric 'advice discounting') is a phenomenon wherein advice is under-weighted during integration with the advice-taker's existing opinion, relative to a normative expectation. Most experiments which explore egocentric discounting use the Judge-Advisor System. The Judge-Advisor System has roles for a judge, usually the participant, and one or more advisors, often another participant. The judge offers an initial estimate for some decision, e.g. the value of coins in a change jar, then receives advice from the advisor, and then makes a final decision. The difference between the initial estimate and the final decision is taken as measure of how influential the advice was, typically expressed in terms of the contributions of the initial estimate and the advice to the final decision.

In these experiments, the task performance for the advisor is usually as good or better than that of the judge. This performance structure can be well captured for most tasks with Gaussian answer + error distribution where the answer supplies the mean and the error supplies the variance. When combining individual estimates from multiple distributions, optimal results are obtained by weighting the estimates according to the relative precision of their parent distributions (**sollStrategiesRevisingJudgment2009**). Where the performance of the advisor is higher than the judge, the advisor's error will be lower than the judge's, and thus the variance of the advisor's distribution will be narrower and therefore the precision of the advisor's distribution will be higher. When a judge combines their own estimate with that of an advisor who is at least as good, an optimal judge will weight the advisor's opinion at least as highly as their own. !TODO[We'll have got a whole section on models of advice integration stuff, probably, so we can point to that here.] The classic presentation of egocentric discounting is when, in these scenarios, the weight applied to the advice is lower than the optimal weight.

Egocentric discounting is a robust phenomenon in advice-taking. It is not a generic inability to combine estimates: people can combine estimates which do not include their own opinion well (Ilan Yaniv and Choshen-Hillel 2012), even adjusting for differences in ability between advisors (Soll and Mannes 2011). Similarly, Trouche et al. (2018) showed that discounting occurred *towards advice* when that advice was labelled as the judge's initial estimate and vice-versa, suggesting that a person's own opinion has a privileged status.

In this chapter, I review the literature on egocentric discounting, with particular attention to manipulations that have been used and explanations which have been offered. I conclude the chapter by expounding an alternative perspective from which to view the phenomenon which, in my view, clarifies the phenomenon and opens the field for a wider range of effective explanations.

6.2 Manipulations affection egocentric discounting

A sizeable body of research has been conducted into egocentric discounting, including a wide variety of manipulations. These manipulations fit roughly into four categories: properties of the task, properties of the advice, properties of the advisor, and wider social factors.

6.2.1 Task properties

The most prominent feature of the task which affects egocentric discounting is difficulty. Gino and Moore (2007) asked participants to estimate a person's weight from a clear (easy condition) or blurry (hard condition) picture, and saw less discounting on the hard task. Likewise, Wang & Du (**wangWhyDoesAdvice2018**) used blurring to increase the difficulty of estimating the number of coins in a photograph of a jar and found that participants discounted less in the blurry compared to the clear condition.

Wang & Du saw full mediation of their difficulty manipulation by participants' confidence on the task, while Gino and Moore saw only partial mediation. Other studies have manipulated the judge's confidence through other mechanisms. See et al. (2011) used a power manipulation which was effective in part through raising judges' confidence; and Gino, Brooks & Schweitzer (2012) used anxiety manipulations to decrease judges' confidence. In both cases, partial or full mediation through confidence occurred such that higher confidence was associated with greater egocentric discounting. In many other experiments, including other experiments in Wang & Du's (**wangWhyDoesAdvice2018**) paper, confidence is not manipulated but is still associated with greater egocentric discounting. Using a similar methodology, but a different analytical approach, Moussaïd, Kämmer, Analytis, & Neth (**moussaidSocialInfluenceCollective2013**) observed that highly confident participants rarely updated their views following advice.

More complex task designs, in which reflection and discussion are encouraged, can reduce discounting. Liberman and colleagues (**minsonTwoTangoEffects2011**; Liberman et al. 2012) asked dyads to take simultaneous roles as judge and advisor, providing initial estimates, exchanging advice during a discussion, and then providing final decisions on estimation tasks. Discounting was still evident in this process, as it was in Van Swol's (2011) traditional Judge-Advisor System paradigm using face-to-face advice. Liberman and colleagues did manage to eliminate discounting where, between exchanging advice and providing a final decision, participants produced a single mutually satisfactory collaborative judgement, and showed that the value of this collaborative judgement was itself improved by open-minded discussion over justifying estimates or exchanging bids. Schultze, Mojzisch & Schulz-Hardt (2017) demonstrated that asking judges to selectively generate reasons why the advice might be correct or incorrect led to lower and higher levels of egocentric discounting respectively.

6.2.2 Advice properties

When judges are more confident, they tend to be less influenced by advice, and the expected corollary of this is that when advice is expressed more confidently the advice will be more influential. Soll & Larrick (**sollStrategiesRevisingJudgment2009**) measured the confidence of advice and saw that higher advice confidence was associated with higher influence of advice. **moussaidSocialInfluenceCollective2013** also found that differences in confidence between judges' and advisors' estimates were useful in producing a decision tree determining the extent to which advice was taken.

Similarity of advice to the initial estimate

The most frequently-manipulated property of advice, and the most interesting in the context of the first part of this thesis, is the similarity of the advice to the initial estimate. This is sometimes expressed as agreement or reasonableness of advice. The evidence on the effects of advice distance on advice influence is equivocal: some studies appear to demonstrate that advice is discounted more the further it is from the initial estimate, while other studies show a greater influence of more distant advice.

Yaniv (2004) manipulated advice to be nearer to or further away from the initial estimate and saw that the influence of advice decreased as the advice was further from the initial estimate (although this pattern did not hold for low-expertise judges in one experiment). Minson, Liberman, & Ross (**minsonTwoTangoEffects2011**) found that more distant advice was associated with less advice-taking behaviour once average distance between dyad members was controlled for, although their results were not expressed using standard advice-taking metrics. Yaniv and Milyavsky (2007) observed that advice closer to the initial estimate was also more influential when combining multiple pieces of advice simultaneously.

Moussaïd, Kämmer, Analytis, & Neth (**moussaidSocialInfluenceCollective2013**) identified a three-zone structure to the influence of advice according to distance. Similar advice fell into the 'confirmation zone', where opinion was unchanged but confidence increased; moderately distant advice fell into an 'influence zone' where averaging strategies come into play; and distant advice was generally ignored. Likewise, Schultze, Rakotoarisoa, & Schulz-Hardt (**schultzeEffectsDistanceInitial2015**) showed in an elegant series of Judge-Advisor System experiments that relationships between egocentric discounting and advice distance were non-linear. Advice was discounted heavily where it was close to the initial estimate, with discounting then dropping precipitously to a peak for middle-distance advice, after which it either remained stable or decreased slightly as the distance increased further. Schultze et al. also showed that confidence in final decisions was dramatically boosted by near advice, and that confidence gains decreased sharply with distance, consistent with Moussaïd et al.'s account. Hütter & Ache (2016) found consistently higher influence for advice which was further from the initial estimate both for single pieces of advice and for integrating multiple pieces of advice. Hütter & Ache used three discrete advice distances which may explain why they found monotonic increases in influence while Moussaïd et al. and Schultze et al. found U-shaped relationships: Hütter & Ache may not have sampled a broad enough swathe of the distance-space.

The argument that close advice updates confidence rather than judgement may offer an explanation of the more straightforward monotonic reduction in advice-taking with distance. This would be consistent with the results of Yaniv and Milyavsky (2007) if we assume that this confidence reduces advice influence (as described above) and this reduction of influence disproportionately affects distant advice. Rollwage et al. (**rollwageConfidenceDrivesNeural2020**) argue that high-confidence decisions reduce the processing of disconfirmatory evidence, and such a process may explain the above phenomenon: near advice inflates confidence, which then disproportionately hampers the processing of far advice, leading to an apparent monotonic decrease in advice influence with distance. These mechanisms do not suffice to explain the results in Yaniv's (2004) study, however, because there is only a single piece of advice which cannot be both near and far simultaneously. Yaniv's (2004) results stand as something of a mystery, especially because Schultze et al. (**schultzeEffectsDistanceInitial2015**) obtained results consistent with their other results in an experiment designed to reproduce Yaniv's methodology as closely as possible.

Related to the distance of advice is the reasonableness of advice, because where the judge has a somewhat reasonable estimate the distance serves as a reliable proxy for reasonableness. Gino, Brooks, & Schweitzer (2012) included an experiment in which non-anxious participants heavily discounted unreasonably high and unreasonably low advice, while discounting reasonable advice at a rate typically seen in Judge-Advisor System experiments. Similarly, Schultze, Mojzisch, & Schulz-Hardt (2017) saw judges discount wildly implausible advice more heavily, although it was still assigned some weight, even when labelled as coming from a random number generator.

Solicitation of advice

- Un/solicited nature of advice
  - Gino & Moore 2007; Gino 2008; Hütter & Ache 2016

- Paying for advice?
  - Gino & Moore 2007; Gino 2008

6.2.3 Advisor properties

- Advisor expertise
  - Experienced

  - Stated
    * Yaniv & Kleinberger 2000; Sniezek et al 2004; Soll & Larrick 2009 (exp 1, 2); Van Swol 2011; Tost, Gino & Larrick 2012; Gino et al 2012; Rakoczy et al 2015; Schultze et al 2017; Wang & Du 2018

  - *Note* with accuracy, the normative point for weighting shifts, so changes in advice weighting might not be changes in discounting

- Pre-existing relationship with advisor
  - Minson, Liberman, & Ross 2011
- Humanness of the advisor (vs e.g. algorithm)
  - Onkal 2009

  - vs RNG Schultze, Mojzisch, & Schulz-Hardt (2017)
- Advisor Conflict of Interest
  - Bonner & Cadman 2014; Schul & Peni 2015; Gino et al 2012 (exp 6)

6.2.4 Wider social factors

- Fairness (counter-effect to some manipulations!)
  - Mahmoodi et al 2015

- Ostracism
  - Byrne et al 2016

- Power
  - See et al 2011 (perhaps via confidence); Tost, Gino & Larrick 2012

- Anxiety?
    - Gino, Brooks, & Schweitzer 2012

6.3 Purported explanations for egocentric discounting

6.3.1 Access to reasons

Yaniv; Stern Schultze & Schulz-Hardt 2017

6.3.2 Anchoring

Bonner & Cadman 2014; Shultze et al 2017

6.3.3 Risk aversion

Also anticipated regret.

6.3.4 Sunk costs

Gino 2008; Ronayne & Sgroi 2018

6.3.5 Expertise/self-inflation bias / Confidence

Gino, Brooks & Schweitzer; See et al 2011; Tost, Gino & Larrick 2012

6.3.6 Stimulus-response model

Schultze, Rakotoarisoa, & Schulz-Hardt 2015

6.3.7 Naive realism

Minson, Liberman, & Ross 2011

6.3.8 Wariness

Trouche et al 2018

6.3.9 Effort heuristic (for free advice)

No effort in = low expected value out (see Gino 2008 on paying for advice)

6.3.10 Responsibility / feeling of deserving outcomes

Ronayne & Sgroi 2018; Mahmoodi et al 2015

6.3.11 (Heirarchiacal?) Bayesian account

Priors and updating. Possibly hierarchical with a probability distribution over your own estimate being incredibly wrong/having overlooked something.

6.4 A wider view of egocentric discounting

The explanations offered are all explanations pitched at the level of are all offered as explanations for a deviation from normative optimality. I have chosen to take a different approach, asking instead under which circumstances the observed behaviour would be an optimal policy, and exploring the plausibility of those circumstances continuing to have an influence in experimental settings where the normative behaviour is averaging one's own opinion with advice.

As a starting point, I note that if one tells anyone who is not an advice-taking researcher that people do not take others' opinions as seriously as they take their own when making decisions, the response is likely to be a flat "of course", perhaps accompanied by a perplexity as to why such an obvious statement is being presented as a valuable insight.

6.4.1 Accounting for the evidence

6.4.2 Compatability with existing explanations

6.5 Evidence

In the chapters which follow, I present evidence from computational agent-based evolutionary simulations and online human behavioural experiments to illustrate the plausibility of the claims made above. The evolutionary simulations demonstrate that an array of plausible factors affecting the relative utility of advice can create an environment in which egocentric discounting is adaptive, and the behavioural experiments demonstrate that some of these factors can be responded to by individual humans by adjustments in behaviour.

7

Sensitivity of advice-taking to context

Advice-taking is often overly-conservative as compared to the normative level of advice-taking for a given experimental design. I argue that participants' performances in advice-taking experiments reflect both the specifics of the experimental design and prior expectations about advice-taking situations. These prior expectations may be both learned, as where individuals who grow up in less stable environments show lower propensity to trust `reference?`!TODO, and inherited. Most useful for the current argument would be a demonstration that conservatism can emerge within a population even where detrimental advice is rarely experienced, and that this can thus produce individuals who exhibit conservatism without ever experiencing detrimental advice. This demonstration is presented in the form of evolutionary modelling.

As discussed in the [previous chapter](#chapter-context), conservatism is optimal under some circumstances, and thus we expect that simulated agents allowed to evolve an advice-taking policy in those circumstances will evolve a conservative policy. I explored this tendency as a function of three plausible scenarios. The first scenario is one in which agents occasionally give deliberately poor advice to their advisee, which represents situations where advisors' interests may sometimes be contrary to judges' interests, unbeknownst to the judges. In the second scenario, advice is simply noisier than the judge's own initial estimate, either because the judge is less competent at the task, less willing to exert the required effort for the task, or because the advice is communicated imperfectly. In the third scenario, agents belong to either a 'cautious' or a 'confident' group in how they express and interpret advice, which is a simple analogue of the observation that people's expressions of confidence are idiosyncratic (Ais et al. 2016; Navajas et al. 2017). In each of these three scenarios, it is hypothesised that some level of egocentric discounting will emerge as the dominant strategy, i.e., the mean population weighting for initial estimates versus advice will be greater than .50. ## General method

Agent-based computational models of an evolutionary process were programmed in R (R Core Team 2018) and run variously on a home computer and the Oxford Advanced Research Computing cluster (Richards 2015). The code is available at `https://github.com/oxacclab/EvoEgoBias`, and the specific data presented below are archived at !TODO.

The models reported here use 1000 generations of 1000 agents which each make 30 decisions/generation on which they receive the advice of another agent. Decisions are either point estimation (Scenarios [1](#models-scenario-1) and 2) or categorical decision with confidence (Scenario 3). Each agent combines their own initial estimate with the advice of another agent, with the relative weights of the initial estimate and advice set by the agent's egocentric bias parameter, to produce a final decision. Final decisions are evaluated by comparison with the objective answer, and an agent's fitness is the sum of its performance over the 30 decisions of its lifetime.

7.0.1 Initial estimates

The agents perform a value estimation (category estimation in Scenario 3) task. Agent i's initial estimate t is the true value (v_t), plus some noise drawn from a normal distribution with mean 0 and standard deviation equal to the agent's insensitivity parameter (s^i, which is itself drawn from a positive-clamped normal distribution with mean and standard deviation 10 when the agent is created).

An agent's initial estimate (e^i_t) is thus:

$$e_t^i = v_t + N(0, s^i) \tag{7.1}$$

7.0.2 Advice

Each agent receives advice from another agent which it combines with its initial estimate to reach a final decision. The advice has a probability of being mutated in some fashion. The mutation depends upon the scenario and is described separately for each.

7.0.3 Final decisions

In the basic model from which other models inherit their decision procedure, agent i produces a final decision t as the average of the agent's initial estimate (e_t^i) and another agent's advice (a_t^i), weighted by the agent's egocentric bias (b^i). The models typically change the value of a_t^i, which is typically a function of some other agent j's initial estimate e_t^j.

An agent's final decision (d_t^i) is thus:

$$d_t^i = \frac{e_t^i \cdot b^i + a_t^i \cdot (1 - b^i)}{2} \tag{7.2}$$

The final decisions in Scenario 3 are more complex, but follow a similar structure.

7.0.4 Reproduction

Roulette wheel selection is used to bias reproduction in favour of agents performing best on the decisions. Performance is determined by a fitness function which differs slightly between categorical and continuous decisions. For scenarios 1 and 2, which use continuous decisions, this fitness is obtained by subtracting the absolute difference between the final decision and the true value for each decision:

$$u^i = -\sum_{t=1}^{30} |v_t - d_t^i| \tag{7.3}$$

The selection algorithm proceeds as follows: The worst performance is subtracted from each agent's fitness and 1 added to put fitness scores in a positive range. These scores are then continually multiplied by 10 until the lowest score is at least 10 to improve resolution. Each agent is then given a probability to reproduce equal to their share of the sum of all fitness scores:

$$r^i = \frac{u^i}{\sum_{j=1}^{n} u^j} \tag{7.4}$$

where n is the number of agents and u has undergone the transformations described above.

Reproducing agents pass on their egocentric bias to their offspring. Other agent features, e.g. decision-making accuracy, are randomised when they are created. In the present simulations, agents receive no feedback on decisions, and cannot learn about or discriminate between their advisors. The key outcome of interest in each simulation is whether the population evolves towards egocentric discounting as the dominant adaptive strategy.

7.1 Scenario 1: misleading advice

In scenario 1, agents sometimes choose to offer misleading advice to their advisee.

7.1.1 Method

The true value (v_t) is fixed at 50 in this scenario. The agents do not learn about the true value over time, so a fixed and arbitrary value does not alter the results of the simulation. Advice in this scenario is either the advising agent's initial estimate (e^j_t), or an extreme answer in the opposite direction to the advising agent's initial estimate (i.e. lower than 50 if e^j_t was above 50, and vice-versa). !TODO[describe what we actually do with this work; it has moved on a bit since transfer report]

7.1.2 Results

Where advice is always genuine, egocentric discounting does not emerge. Where advice is sometimes misleading, however, egocentric discounting emerges rapidly and remains stable throughout the rest of the stimulation.

7.1.3 Discussion

As one might expect, where there is the potential for exploiting trust it is safer to trust less, even if this reduces some of the benefits which would be gained from trusting. Egocentric discounting may be a viable strategy, even if there is no difference in ability between advice-seekers and advice-givers, given social contexts where the interests of advice-seekers and advice-givers are not perfectly aligned.

7.2 Scenario 2: noisy advice

In this scenario, agents offer advice which is of slightly lower quality on average than their own intial decisions. This could reflect a difference in effort or in expertise.

7.2.1 Method

As with Scenario 1, the true value (v_t) is fixed at 50.

Advice in this scenario has additional noise added:

$$a_t^i = e_t^j + N(0, x) \tag{7.5}$$

Where $x = 0$ in the no-noise condition and 5 in the noise condition.

Models were also run where a new decision was used as the basis for advice, rather than an initial estimate that some other agent would be using as a component in its own final decision:

$$a_t^i = v_t + N(0, s^j + x) \tag{7.6}$$

but the results were the same.

7.2.2 Results

As before, egocentric discounting emerges rapidly in the active condition and remains stable throughout the rest of the simulation.

7.2.3 Discussion

Where the advice is of worse quality on average, encoded here as additional variation, egocentric discounting tailors the relative weights of the estimates according to their average quality. As with situations in which an advice-seeker is more competent at making the relevant decision than their advisor, some measure of egocentric discounting is warranted in this scenario. It is worth noting that different competencies are not the only reason why advice may be systematically less valuable than initial decisions: difficulties in communicating the advice or different levels of conscientiousness in decision-making may also produce this effect.

7.3 Scenario 3: confidence confusion

While lackadaisical, incompetent, deliberately bad, or poorly communicated advice produces an obvious adaptive advantage for egocentric discounting, it is plausible that scenarios may exist where equally competent, wholly well-intentioned advice may still favour egocentric discounting. A common feature of advice is the communication of confidence, and this improves outcomes (Bahrami et al. 2010). Notably, there are large and consistent individual differences in people's expressions of confidence (Ais et al. 2016; C. Song et al. 2011). In this scenario agents are equally competent at the task, and do their best to assist one another, but may be hampered by expressing their estimates and advice with different confidences.

7.3.1 Method

The true value was drawn from a normal distribution around 50:

$$v_t = N(50, 1) \tag{7.7}$$

This allowed categorical answers which identified whether or not v_t was greater than 50. Agents were equiprobably assigned a personal confidence factor (c) of 1 or 10. This was used to scale the difference between the advising agent's intial estimate and the category boundary to produce the advice:

$$a_t^i = (e_t^j - 50 \cdot c^j) + 50 \tag{7.8}$$

Each agent then used the reciprocal of this process to translate advice back into its own confidence scale before integrating it with the initial estimate and arriving at a final decision:

$$d_t^i = \frac{e_t^i \cdot b^i + (\frac{1}{c^i}(a_t^i - 50) + 50) \cdot (1 - b^i)}{2} \tag{7.9}$$

This process amounts to the outgoing advice being translated into the advising agent's confidence language, and incoming advice being translated into the advised agent's confidence language. Where these languages are the same, the resulting advice is understood equivalently by both agents, but where there are differences the advice will be of greater or lesser magnitude compared to the initial estimate.

7.3.2 Results

Egocentric discounting again emerges in the condition where agents can have different confidence factors. The adaptiveness of egocentric discounting in this scenario arises because advice from a mismatched partner (e.g. one using a confidence scaling of 10 when the judge uses 1, or vice-versa) requires a different aggregation approach than advice from a matched partner. The application of an inappropriate aggregation approach results in misleading advice, so there is a pressure for a middle-of-the-road strategy whereby advice is counted, but not too much.

7.3.3 Discussion

Even where there is no intent to mislead, and no difference in basic ability, it is possible for egocentric discounting to emerge as an optimal strategy, purely because of differences in how agents communicate and understand estimates.

7.4 General discussion

The computational models show that egocentric discounting is an adaptive strategy in an array of plausible advice contexts: misleading advice, noisy advice, and different interpretations of confidence. While these models are necessarily limited in applicability to real life, they do demonstrate that egocentric discounting, while irrational for simple estimation problems with an objective answer and with advice that is not systematically better or worse than an individual's own judgement, may be beneficial for many of the kinds of decision for which we have sought and used advice in everyday life throughout our evolution. This argument invites attention to the advice-taking task as much as to the properties of the advisor: it predicts that egocentric discounting will be attenuated where the outcome of decisions affects judges as well as advisors (Gino 2008, observed this effect but attributed it to judges falling prey to the sunk costs fallacy); where decisions rely more on objective than on subjective criteria (Van Swol 2011); where advisors and judges have opportunities to calibrate their confidence judgements with one another by completing training trials where they have to produce a shared decision with a shared confidence; and where incentives for judges and advisors are more closely aligned (Gino, Brooks, et al. 2012; Sniezek, Schrah, et al. 2004).

Notably, the utility of these heuristics does not depend on malice, mistake, or miscommunication: inconsistency in the usage of confidence terminology can produce adaptive pressure for egocentric discounting. More generally, the results indicate that properties of the advice-giving milieu can influence advice-taking strategies.

These models establish that it is plausible that people have deeply ingrained hyperpriors towards discounting advice. It is also possible, however, that people can flexibly respond to contexts, modulating their advice-taking appropriately. The voluminous experiments in advice taking discussed in the introduction to this section !TODO[that discussion?] can be seen as eliciting exactly this behaviour. In the next chapter I present new behavioural experiments which suggest that people do indeed modulate their advice-taking to the specifics of the context in which they are in.

8

Behavioural responses to advice contexts

8.1 Benevolence of the advisor population

!TODO[Make sure to report in brief/as post mortem the results for the in/out group studies and the early direct benevolence experiments, and point readers to the full details.]

The evolutionary models discussed in the last chapter demonstrated that optimal advice-taking strategies depend in part upon the advice one receives being a genuine effort to help. Difference in benevolence, or the extent to which the interests of the advisor and the judge overlap, have been shown to affect levels of advice-taking. !TODO[CITE] showed that, where advisors were paid contingent upon the quality of the final decision, advice-taking was higher than where advisors were paid a flat fee for providing advice. !TODO[lit review stuff - inc. Behrens/Hunt? social decision making]

Advice-taking can be contingent on the properties of the advice, or on the properties of the advisor. In order to maximise the value of advice while minimising the potential exposure to exploitation, advice-taking should be contingent on a combination of these factors. Where advice is plausible it should be weighted relatively equally, whether it comes from an advisor who is sometimes misleading or not, but where advice is more implausible it should only be trusted when

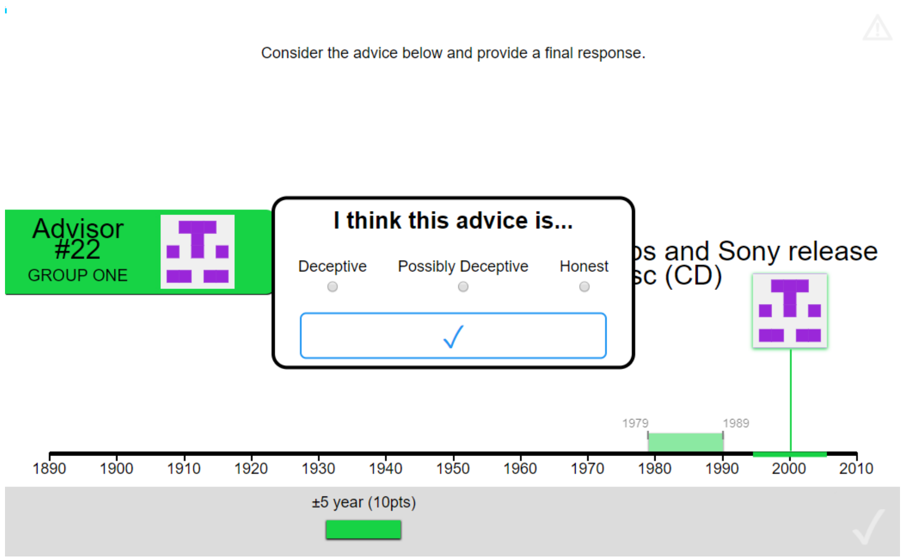

Figure 8.1: Advice honesty rating.
Participants rated advice on a three-point scale according to whether they thought the advice was deceptive or honest.

it comes from an advisor who is never misleading. To explore whether people's behaviour matches this pattern, participants were recruited for a series of behavioural experiments in which they were given advice on a date estimation task from advisors who were described as either always helpful or occasionally misleading. Results indicated that participants' advice-taking depended upon both the plausibility of the advice and the benevolence of the advisor, although the specific pattern differed from the one described above.

8.1.1 Method

Procedure

The procedure for these experiments follows the dates estimation task, described previously. Additionally, in this version, participants had to provide an assessment of the honesty of each advisory estimate they received before they could integrate it into their final decision.

Manipulation

Participants performed two blocks of experimental trials. In one block, participants received advice from an advisor who they were told would 'always try to help', while in the other they received advice from an advisor who 'may sometimes try to mislead'. The order of advisors was counterbalanced between participants.

The advice offered by the advisors was drawn from equivalent distributions for both advisors, meaning that there were no systematic differences in the quality of the advice. Both advisors offered advice sampled from a roughly normal distribution around the correct answer with a standard deviation of 11 years.

Experiments

There were a series of experiments in this topic during which the experimental design was tweaked in order to find an effective manipulation and operationalisation of the key concepts. The experimental code for previous versions can be found on the commit history on GitHub for the main experimental file, as well as its dependencies. To give a brief overview of the major changes:

- Version 2.0.0
    - introduced a much clearer manipulation
        - reminded participants of advisor description in a message they had to acknowledge
        - kept participant's group visible throughout
        - included a single trial where the misleading advisor did actually mislead the participant
- Version 2.1.0
    - removed actual differences in advice and adjusted advisor descriptions to match

- Version 3.0.0
  - added a question probing the perceived honesty of the advice between receiving advice and providing a final decision
- Version 3.0.1
  - pre-registered replication of version 3.0.0
  - Deviations from pre-registration:
    * Added a new exclusion rule for those people who use translation software.
- Added exclusions for participants with NA values in the in- vs out-group t-test (where e.g. no outgroup advice was rated as 'honest').

```
* Included frequentist stats in the trustworthiness questionnaire item t-tes
```

```
* Fixed some labels on graphs
```

```
* Exploratory analyses expanded to include analysis of trials with WoA > .05
```

Below, only the results from **version 3.0.1** are reported, but the results of this version are compatible with findings of previous versions. Data for all versions are accessible !TODO[archive data].

8.1.2 Results

!TODO[Restructure this code to use the esmData package. Should be easy...]
!TODO[consistent advisor naming, including in figures]

Exclusions

Participants (total $n = 20$) could be excluded for a number of reasons: failing attention checks, having fewer than 11 trials which took less than 1 minute to complete (one participant), providing final decisions which were the same as the initial estimate on more than 90% of trials, or using non-English labels for the honesty questionnaire (one participant). The latter exclusion was added after data were collected because it was not anticipated that participants would use translation software in the task.

The final participant list consists of 18 participants who completed an average of 11.83 trials each.

Task performance

Participants performed as expected, decreasing the error between the midpoint of their answer and the true answer from the initial estimate to the final decision, which suggests that they incorporated the advice, which was indicative of the correct answer.

```
## ‘summarise()‘ regrouping output by ’pid’ (override with ‘.groups‘ argument)
```

```
## Warning: ‘fun.y‘ is deprecated. Use ‘fun‘ instead.
```

The advice participants received was subject to variation, but there did not appear to be overall systematic differences between the advisors either in terms of error (distance between the centre of the advice estimate and the correct answer) or distance (distance between the centre of the advice estimate and the centre of the initial estimate). !TODO[error and distance graphs]

Hypotheses

The hypotheses tested were that * weight on advice would be higher for advice rated as ‘honest’ versus advice rated as ‘deceptive’ * weight on advice would be higher for advisors who were described as ‘never misleading’, even for advice rated as ‘honest’

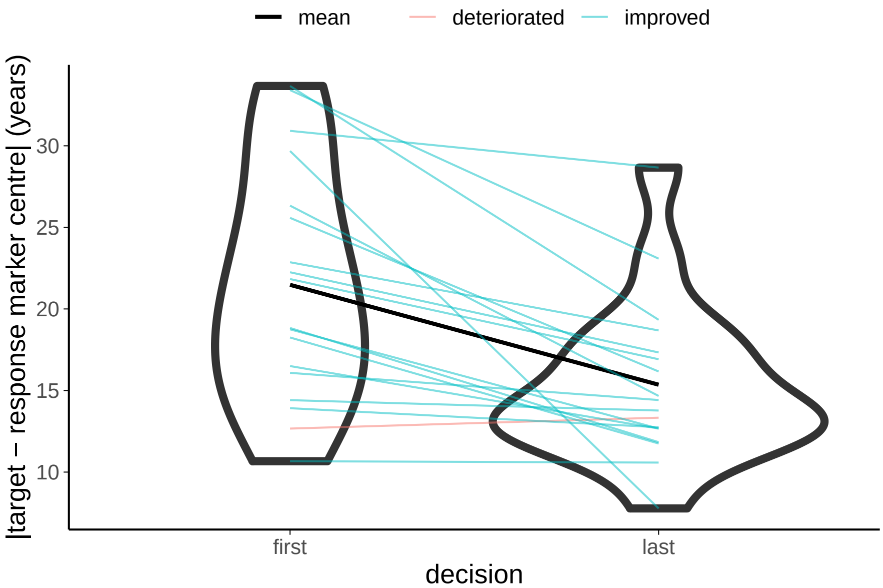

Figure 8.2: Mean answer error for initial estimates and final decisions.
Faint lines show means for individual participants collapsed across trials, while boxplots, violins, and the heavy dashed line show aggregate participant means.

Effect of advice

```
## `summarise()` regrouping output by 'pid' (override with `.groups` argument)
```

As expected, participants were substantially more influenced by advice they rated as 'honest' compared to advice they rated as 'deceptive' (!TODO[fix ttest]).

Effect of advisor

```
## `summarise()` regrouping output by 'pid' (override with `.groups` argument)
```

Participants were also more influenced by 'honest' advice from the advisor who was described as 'never misleading' (!TODO[fix ttest]).

!TODO[plot of woa by advice rating and advisor]

Results plot The details of the data are best presented in a combined figure.

The figure shows that there is a different pattern in the responses to advice across advice ratings depending upon the advisor supplying the advice. Advice perceived as not being honest is discounted relatively equally from the advisor described as 'always helpful', and is still fairly influential in the final decision. The advice perceived as 'honest' from this advisor is the most influential of all advice. Likewise, the 'honest' advice from the advisor described as 'sometimes-misleading' is more influential than the 'possibly deceptive' advice, which is in turn more influential than the 'deceptive' advice. The 'deceptive' advice from the 'sometimes-misleading' advisor is almost completely ignored. Within this broader pattern there is some variability between participants, including some participants whose responses violate the pattern quite dramatically - note the two participants with large decreases in influence for 'honest' advice from the 'sometimes-misleading' advisor as compared to the 'possibly deceptive' advice from the same. This may be a consequence of participants' values in each cell in the design being comprised of a low number of trials. #### Exploration

!TODO[below]

- were participants' criteria for judging advice as dishonest different between advisors?
- Type 1 SDT might be able to do this if a sensible definition of objectively-deceptive advice can be produced

```
* E.g.
```

whether averaging with the advice would reduce error

```
* There will be all sorts of problems with this, noteably it'll classify 'ne
```

  - May be able to investigate this with probabilities of classification for several distance bins - ROC type curve
    - what to do with 'possibly deceptive'?

8.1.3 Discussion

- Both the source and the plausibility of advice matter, and they seem to interact in a complex way.

- Flexible adaptation to advice context, both in terms of categorisation and influence.
- Support for the premise of the evolutionary simulations.
  - Does not prove people actually do have these hyperpriors

Limitations

- N
- No gender breakdown
  - simplicity/privacy vs data gap
- Difficult task (high levels of advice-taking by default)

8.2 Noise in the advice

The second scenario explored in the evolutionary models added noise to the advice agents received and demonstrated that this provided an evolutionary pressure towards egocentric discounting. This scenario is not explored in behavioural experiments because its conclusions are well supported by existing literature. Specifically, the literature demonstrates that the normativity of discounting where the judge outperforms the advisor, that advice-taking is sensitive to advisor expertise, and that people are likely to consider themselves superior to the average advisor.

The addition of noise in a point-value estimation task lowers the relative performance, and thus this scenario was essentially a manipulation of advisor expertise. Normative models support egocentric discounting where the judge is systematically more accurate than the advisor !TODO[PAR model paper, early advice-taking theory papers]. The decrease in advice-taking for novice as opposed to expert advisors is a robust result in the literature: !TODO[expertise manipulation experiment lit review]. The evidence above makes an empirical investigation on this point somewhat redundant, provided it can be shown that people have a naive assumption of superiority on a given task compared to the average other they are likely to encounter. This point is also well supported in the literature on self-serving biases !TODO[describe driving self-assessment, naive realism?, etc.]. It is expected that this is somewhat dependent upon the difficulty of the task presented; people faced with a difficult task will under- rather than overestimate their ability relative to others !TODO[hard-easy effect of advice]. Taken together, the above arguments demonstrate that, on average, people are likely to consider themselves more able on a given task than an arbitrary advisor, and consequently that they are likely to downweight advice relative to their own initial estimate. This behaviour is supported by normative mathematical models which show biasing towards the better estimator (in this case the judge) is the optimal strategy. As discussed in the scenario, the belief that one is better at a task than the average advisor may not be misguided: advisors may not dedicate the same amount of time, concentration, or thought to producing advice as judges do for initial estimates. Judges have to live with the consequences of their decisions, while advisors do not. ## Confidence mapping

The third scenario explored in the evolutionary models assigned each agent a confidence mapping, and demonstrated that discounting emerged as an appropriate response where the advisor's confidence mapping was unknown. The key difficulty in conducting behavioural experiments to test the effects of known vs unknown confidence mapping is finding a manipulation of confidence mapping knowledge which is not confounded by familiarity with an advisor or the amount of information provided by an advisor. We train people in two contexts, in both of which they receive advice with confidence from one of two advisors and see feedback on their final answers. In the first context, all the advice is well calibrated, and both advisors use either high or low confidence distributions. In the second context, one advisor uses a high and the other a low confidence distribution. Participants are then tested with a brief exposure to an advisor whose confidence distribution is drawn from the middle of the scale. Where participants are unprepared for different confidence mappings (the first context), they will treat the new advisor like the old advisors - considering the advice as high- or low-confidence (and weighting it accordingly) depending on whether it is high or low for the advisors' distributions. Where participants are prepared for different confidence mappings (the second context), they will treat the new advice as medium confidence, and weight it accordingly. There will be a difference in the weight on advice in the probe trials between contexts: WoA will be highest where both training advisors have a low-confidence distribution, middling where they have different distributions, and lowest where they both have a high-confidence distribution.

8.2.1 Individuality as a cue to confidence

Method

Results

8.2.2 Identifiability of advice

Method

Results

8.3 General discussion

The results of the direct benevolence experiments show that people are sensitive to the motivations of their advisors, and exercise appropriate caution where those motivations may mean the advice is misleading. They also demonstrate that, even where advice is considered trustworthy, it is discounted when coming from a less trusted advisor. Together with existing literature on the effects of advisor expertise, this supports the idea from the models that advice-taking can be fleixbly adjusted according to the context. The results of the confidence mapping experiments were less conclusive. We were unable to produce a manipulation which was sufficiently clear and strong to produce observable effects. It is plausible that people are sensitive to their knowledge of an advisor's confidence mapping, but it is also plausible that this degree of flexibility is beyond most people, and that simple heuristics (e.g. relying on cultural norms about the meanings of metacognitive terms !TODO[cite Bang, Heyes, et al. on cultural metacognition]) are used instead of complex calculations. The lack of flexibility on the timescale of a behavioural experiment does not, of course, negate the possibility of flexibility on the timescale of a human interpersonal relationship (perhaps there are a certain number of people whose confidence mappings we can track, a la Dunbar's number). Nor does it negate the possibility of an evolutionary or cultural-evolutionary mechanism baking in advice discounting as a protection against unknown confidence mapping, although it is questionable whether advice has been occuring in human societies for long enough to allow the former to take place. We have a plausbile explanation of egocentric which only relies on rational responses to environmental effects. These effects are ubiquitous, and thus a ubiquitous egocentric bias prior to engaging with the specifics of a situation makes sense. It is entirely plausible that experiments showing egocentric discounting behaviour fail to overcome the hyperpriors held by participants that taking advice is risky for various reasons.

We saw in the previous section that people are sensitive to the quality of advice and advisors, and will curate their information environments based on that sensitivity. We have seen in this section how the properties of the information environment can change people's sensitivity to advice. !TODO[If we have time, it'd be neat to see if we can get people's selection rates for advisors to recover after the advisors have been lying to them, showing that that how we curate information environment (source selection) also depends on the existing information environment properties.]

Part IV

Interaction

9

Interaction of psychological processes across minds

!TODO[extend the simulation/agreement-effects-analysis stuff a bit and it will fill this chapter out well.]

- Even when biases don't change (Niccolo covered that), source selection on the basis of agreement is inherently pathological.
  - This is true whether or not confidence weights updating.
- Why should we care if it's not changing biases (i.e. opinions?)
  - because it reinforces final decisions through agreement instead of attenuating them through disagreement.
- Our psychological mechanisms from Section II thus produce bad network effects when feedback is absent.
- Can we relate the networks back to the advice-giving environments from Section III?

Models say this will go horribly wrong.

10
Network effects of interaction

My models also suggest horrible things.

11 Real-world network effects

My models may not be good models.

Part V

Conclusion

12 Conclusion

I haven't wasted 3 years of my life and Nick's time.

- I've demonstrated that the structure of information networks may be partially dependent upon psychological mechanisms of advisor evaluation.

- I've provided some evidence for a model of advisor evaluation in the absence of feedback and demonstrated how that model produces adverse network effects.

- I've shown how the information environment can affect advice taking, and I've provided and reviewed empirical evidence that humans can alter advice-taking flexibly depending on these contextual factors.

- I've modelled how the psychological mechanisms create an information network structure, and how that information network structure might use the flexibility to context to change advisor evaluation processes.

- I've demonstrated how those changes might attenuate or exacerbate the formation of pathological network structures.

12.1 Open questions

- People's social networks constrain whom they can receive advice from. Within these constraints, source selection processes operate to evaluate the relative (and absolute?) trustworthiness of each (or the most salient?) members. Is trustworthiness (or at least the ability component) modelled discretely for different domains, or is it unitary?

  – Perhaps a matter of association with a particular concept?

  – Quite possible associative processes drive source selection rather than direct comparisons.

- Relative magnitude of advice from selected individuals in a social network vs. adverts/web searches/consulting experts.

- Frequency of advice-seeking as characterised in the thesis.

- Why do people lump group advice together before discounting - this seems idiotic even with the adapativeness of discounting...

Appendices

A
The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

B
The Second Appendix, for Fun

Works Cited

Ais, Joaquín et al. (Jan. 2016). “Individual Consistency in the Accuracy and Distribution of Confidence Judgments”. In: *Cognition* 146.Supplement C, pp. 377–386.

Bahrami, Bahador et al. (2010). “Optimally Interacting Minds”. In: *Science* 329.5995, pp. 1081–1085.

Barberá, Pablo (2015). “How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S.” en. In: p. 46.

Byrne, Kaileigh A. et al. (Oct. 2016). “Ostracism Reduces Reliance on Poor Advice from Others during Decision Making”. English. In: *Journal of Behavioral Decision Making* 29.4. WOS:000383623700006, pp. 409–418.

Cardoso, Felipe Maciel et al. (July 2017). “Topical Homophily in Online Social Systems”. en. In:

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (Apr. 2014). “Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter”. en. In: *Journal of Communication* 64.2, pp. 317–332.

Dana, Jason and Daylian M Cain (Dec. 2015). “Advice versus Choice”. en. In: *Current Opinion in Psychology* 6, pp. 173–176.

Dandekar, Pranav, Ashish Goel, and David T. Lee (Apr. 2013). “Biased Assimilation, Homophily, and the Dynamics of Polarization”. en. In: *Proceedings of the National Academy of Sciences* 110.15, pp. 5791–5796.

Galton, Francis (1907). “Vox Populi (the Wisdom of Crowds)”. In: *Nature* 75, pp. 450–451.

Garrett, R. Kelly (Jan. 2009a). “Echo Chambers Online?: Politically Motivated Selective Exposure among Internet News Users”. en. In: *Journal of Computer-Mediated Communication* 14.2, pp. 265–285.

— (Dec. 2009b). “Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate”. en. In: *Journal of Communication* 59.4, pp. 676–699.

Gino, Francesca (Nov. 2008). “Do We Listen to Advice Just Because We Paid for It? The Impact of Advice Cost on Its Use”. en. In: *Organizational Behavior and Human Decision Processes* 107.2, pp. 234–245.

Gino, Francesca, Alison Wood Brooks, and Maurice E Schweitzer (2012). “Anxiety, Advice, and the Ability to Discem: Feeling Anxious Motivates Individuals to Seek and Use Advice”. en. In: *Journal of Personality and Social Psychology* 102.3, pp. 497–512.

Gino, Francesca and Don A. Moore (Jan. 2007). “Effects of Task Difficulty on Use of Advice”. en. In: *Journal of Behavioral Decision Making* 20.1, pp. 21–35.

Hertz, Uri and Bahador Bahrami (Apr. 2018). “Intrinsic Value of Social Influence over Others”. In:

Hertz, Uri, Stefano Palminteri, et al. (Dec. 2017). “Neural Computations Underpinning the Strategic Management of Influence in Advice Giving”. en. In: *Nature Communications* 8.1, p. 2191.

Hütter, Mandy and Fabian Ache (2016). "Seeking Advice: A Sampling Approach to Advice Taking". en. In: *Judgment and Decision Making* 11.4, p. 16.

Kleiner, M, D Brainard, and D Pelli (2007). "What's New in Psychtoolbox-3? Perception 36 ECVP Abstract Supplement". In:

Kobayashi, Tetsuro and Ken'ichi Ikeda (Sept. 2009). "SELECTIVE EXPOSURE IN POLITICAL WEB BROWSING: Empirical Verification of 'Cyber-Balkanization' in Japan and the USA". en. In: *Information, Communication & Society* 12.6, pp. 929–953.

Liberman, Varda et al. (Mar. 2012). "Naïve Realism and Capturing the "Wisdom of Dyads"". en. In: *Journal of Experimental Social Psychology* 48.2, pp. 507–512.

MATLAB (2017). The Mathworks, Inc.

Mayer, Roger C., James H. Davis, and F. David Schoorman (1995). "An Integrative Model of Organizational Trust". In: *Academy of management review* 20.3, pp. 709–734.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001). "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* 27.1, pp. 415–444.

Minson, Julia A. and Jennifer S. Mueller (Mar. 2012). "The Cost of Collaboration: Why Joint Decision Making Exacerbates Rejection of Outside Information". en. In: *Psychological Science* 23.3, pp. 219–224.

Navajas, Joaquin et al. (Nov. 2017). "The Idiosyncratic Nature of Confidence". En. In: *Nature Human Behaviour* 1.11, p. 810.

Nelson, Jacob L. and James G. Webster (July 2017). "The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience". en. In: *Social Media + Society* 3.3, p. 205630511772931.

Pescetelli, Niccolo (2017). "On the Use of Metacognitive Signals to Navigate the Social World". DPhil. University of Oxford.

Pescetelli, Niccolo and Nick Yeung (Sept. 2018). "The Role of Decision Confidence in Advice-Taking and Trust Formation". In: *arXiv:1809.10453 [cs]*. arXiv: `1809.10453 [cs]`.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rader, Christina A., Richard P. Larrick, and Jack B. Soll (Aug. 2017). "Advice as a Form of Social Influence: Informational Motives and the Consequences for Accuracy". en. In: *Social and Personality Psychology Compass* 11.8, n/a–n/a.

Richards, Andrew (Aug. 2015). "University of Oxford Advanced Research Computing". In:

Ronayne, David and Daniel Sgroi (2018). "Ignoring Good Advice". In:

Sah, Sunita, Don A. Moore, and Robert J. MacCoun (July 2013). "Cheap Talk and Credibility: The Consequences of Confidence and Accuracy on Advisor Credibility and Persuasiveness". In: *Organizational Behavior and Human Decision Processes* 121.2, pp. 246–255.

Schul, Yaacov and Noam Peni (Oct. 2015). "Influences of Distrust (and Trust) on Decision Making". English. In: *Social Cognition* 33.5. WOS:000364258500004, pp. 414–435.

Schultze, Thomas, Andreas Mojzisch, and Stefan Schulz-Hardt (May 2017). "On the Inability to Ignore Useless Advice: A Case for Anchoring in the Judge-Advisor-System". en. In: *Experimental Psychology* 64.3, pp. 170–183.

Sears, David O. and Jonathan L. Freedman (1967). “Selective Exposure to Information: A Critical Review”. In: *Public Opinion Quarterly* 31.2, pp. 194–213.

See, Kelly E. et al. (Nov. 2011). “The Detrimental Effects of Power on Confidence, Advice Taking, and Accuracy”. In: *Organizational Behavior and Human Decision Processes* 116.2, pp. 272–285.

Sniezek, Janet A., Gunnar E. Schrah, and Reeshad S. Dalal (July 2004). “Improving Judgement with Prepaid Expert Advice”. en. In: *Journal of Behavioral Decision Making* 17.3, pp. 173–190.

Sniezek, Janet A. and Lyn M. Van Swol (Mar. 2001). “Trust, Confidence, and Expertise in a Judge-Advisor System”. In: *Organizational Behavior and Human Decision Processes* 84.2, pp. 288–307.

Soll, Jack B. and Albert E. Mannes (Jan. 2011). “Judgmental Aggregation Strategies Depend on Whether the Self Is Involved”. en. In: *International Journal of Forecasting* 27.1, pp. 81–102.

Song, Chen et al. (Dec. 2011). “Relating Inter-Individual Differences in Metacognitive Performance on Different Perceptual Tasks”. In: *Consciousness and Cognition.* From Dreams to Psychosis: A European Science Foundation Exploratory Workshop 20.4, pp. 1787–1792.

Song, Hyunjin and Hajo G. Boomgaarden (Apr. 2017). “Dynamic Spirals Put to Test: An Agent-Based Model of Reinforcing Spirals Between Selective Exposure, Interpersonal Networks, and Attitude Polarization”. en. In: *Journal of Communication* 67.2, pp. 256–281.

Sunstein, Cass R. (2002). *Republic.Com.* en. Princeton University Press.

— (Apr. 2018). *#Republic: Divided Democracy in the Age of Social Media.* en. Princeton University Press.

Trouche, Emmanuel et al. (Jan. 2018). “Vigilant Conservatism in Evaluating Communicated Information”. en. In: *PLOS ONE* 13.1. Ed. by Alexander N. Sokolov, e0188825.

Van Swol, Lyn M. (Jan. 2011). “Forecasting Another’s Enjoyment versus Giving the Right Answer: Trust, Shared Values, Task Effects, and Confidence in Improving the Acceptance of Advice”. en. In: *International Journal of Forecasting* 27.1, pp. 103–120.

Yaniv, I. (Jan. 2004). “Receiving Other People’s Advice: Influence and Benefit”. English. In: *Organizational Behavior and Human Decision Processes* 93.1. WOS:000188504000001, pp. 1–13.

Yaniv, Ilan and Shoham Choshen-Hillel (Dec. 2012). “Exploiting the Wisdom of Others to Make Better Decisions: Suspending Judgment Reduces Egocentrism and Increases Accuracy”. en. In: *Journal of Behavioral Decision Making* 25.5, pp. 427–434.

Yaniv, Ilan and Eli Kleinberger (Nov. 2000). “Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation”. en. In: *Organizational Behavior and Human Decision Processes* 83.2, pp. 260–281.

Yaniv, Ilan and Maxim Milyavsky (May 2007). “Using Advice from Multiple Sources to Revise and Improve Judgments”. English. In: *Organizational Behavior and Human Decision Processes* 103.1. WOS:000246807900007, pp. 104–120.