Exploring Social Metacognition

Matt Jaquiery

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

For Mary Ann Furze, who would have got one too if it weren't for the meddling patriarchy.

Acknowledgements

It almost feels dishonest to check the little box on the submission form that claims this work as wholly my own when so much thanks and credit is due to others who have helped along the way. The greatest contributions are due to my supervisor, Nick, without whose gentle wisdom, perspicacious comments, and apparent ability to read limitless text of ludicrous density without losing concentration this thesis would be very much the worse. Many of the best ideas in this thesis arose from our discussions, if they didn't come from him directly, and every part of this thesis has been improved by his suggestions at one point or another.

I have enjoyed similarly helpful discussions with members of our Attention and Cognitive Control (ACC) Lab, especially with Joshua, Linda, Sarah, Maja, and Raj. Between journal clubs, maths clubs, lab socials, and side projects they've always been there to help me learn, think, and reflect. Similarly, members of the Oxford Experimental Psychology ReproducibiliTea journal club have inspired me – I've enjoyed our discussions immensely when we agreed and still more when we found points of difference to investigate. Special mentions must go to Dorothy, Laura, and Malika. Mentions also to Sam and Amy, for those journal clubs and also for welcoming me to the ReproducibiliTea Core Team. Likewise, Marcus, Jackie, Emma, and the UK Reproducibility Network for helping me wield my experience with open research practices for a practical purpose.

I thank my collaborators on non-thesis projects for all the discussion and support. Alex, whose academic companionship has been invaluable since Sussex. Olly and Lisa, whose willingness to listen to me ramble about things has saved me many hours of draft-writing and whose friendship has brought me joy. Jools, Marwa, and Danielle, whose work on our various projects has always constituted the lion's share – I'm sorry for all the times I've kept them waiting for something or had to rely on them catching some mistake I've made! Balazs, Marton, and Mate, whose response when we found a critical analysis bug while the proofs for our paper were with the editor was nothing short of exemplary.

My family, Gran, Caroline, and Edd, who kept a place for me to go to when breaks were needed, and to mum for meeting me there when she could. To Jo, whose support during this thesis is appreciated, and whose support in the years before I began studying was invaluable. The chance to unwind playing basketball,

ultimate, volleyball, and anything else that came my way was appreciated, too – thanks to Maarten, Richard, the Wolfson Ultimate Frisbee squad, and the West Oxford Dads. Thanks to the Film and Discussion Night crew for opportunities to relax in a less active but equally fun way. Some of the best discussions were about the films I liked the least!

On a less personal but equally important note, I thank the Medical Research Council for funding my studentship, and the very many programmers on whose work this thesis relies. Notable amongst the latter group is Ulrik Lyngs, whose work on the Oxforddown thesis template has saved many hours of fiddling about with code and made writing in RMarkdown a practical option. And thank you to anyone who has tried to read this thesis, proofreaders and reader-readers alike. Even if you only get this far.

Finally, thanks to Frances and Savvy. They enrich and enliven my life. One of them painstakingly proof-read every word and found several errors while the other lay around on the sofa and occasionally demanded to be fed. Thankfully, neither of them brought me too many mice in the middle of the night. Love and gratitude to them both.

Matt Jaquiery
Wolfson College, Oxford
8 October 2021

Abstract

This thesis explores two questions: does the way individuals seek advice produce echo chamber-like networks; and is the well-established phenomenon of egocentric discounting explicable as a rational process? Both parts are presented within a framework of advice as information transfer; the implications for wider interpretations of advice are discussed in the conclusion. Both parts are investigated with a mixture of computational simulations and behavioural experiments.

For the first question, behavioural experiments implementing a Judge-Advisor System with a perceptual decision-making task and a date estimation task are used to characterise people's propensity to use agreement as a signal of advice quality in the absence of feedback. These experiments provide moderate evidence suggesting that people do do this, and that experience of agreement in the absence of feedback increases their trust in advisors. Agent-based computational simulations take the results of the behavioural experiments and simulate their effects on trust ratings between agents. The simulations indicate that including the kind of heterogeneity seen in the participants in the behavioural experiments slows down the formation of echo chambers and limits the extent of polarisation.

In the second part, I argue that egocentric discounting deviates from a normative model of advice-taking because it is a rational response to concerns that always accompany advice: that the advice might be deliberately misleading, lazily researched, or misunderstood. Evolutionary computational simulations of advice-taking illustrate that when any of these circumstances might be true, egocentric discounting emerges as an adaptive response. Behavioural experiments using a date estimation task within a Judge-Advisor System test whether people respond adaptively to alterations in the circumstances explored in the evolutionary simulations. These experiments show that people respond flexibly to changes in the probability that their advisor will attempt to mislead them. Experiments attempting to explore people's ability to flexibly respond to acquiring information about an advisor's confidence calibration were inconclusive.

A web-book version of this thesis is available at `https://mjaquiery.github.io/oxforddown/`. Its RMarkdown source code is available at `https://github.com/mjaquiery/oxforddown`.

Contents

Appendices

List of Figures

List of Tables

Glossary

Advice . . . . Here broadly defined as information which comes from a social source, advice as commonly understood includes an intentional aspect on the part of the advisor (the advisor intends it to be advice).

ANOVA . . . . Analysis of Variance; a statistical analysis technique based on the general linear model.

BF . . . . Bayes Factor; the extent to which the data support one model over another.

CSS . . . . Cascading Style Sheets; used for specifying the presentation of web pages.

Echo chamber . . . . A social or information environment characterised by high levels of homogeneity in values, experiences, and opinions.

Egocentric discounting . . . The phenomenon wherein people take less advice than they should according to a normative model of advice-taking.

Feedback . . . . Objective information concerning the correctness of a decision.

GBP . . . . Great British Pound; Â£ – currency of the United Kingdom of Great Britain and Northern Ireland.

Homophily . . . . 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.

HTML . . . . Hyper-Text Markup Language; used for specifying the layout of web pages.

Hyper-prior An expectation that is not changed as a function of experience within the scope of a given scenario, forming part of the context within which the scenario occurs.

Influence The amount a binary decision with confidence is adjusted in the direction of advice.

Judge-Advisor System An experimental paradigm in which a *judge* makes an initial estimate, receives advice from one or more *advisors*, and then makes a final decision.

PHP PHP: Hypertext Preprocessor; server-side programming language for rendering and processing dynamic web content.

Polarisation The phenomenon whereby the distribution of opinions in a population trend toward extremes over time, with moderate stances becoming increasingly less popular.

Preregistration Specifying information about a study (and especially its analysis of data) prior to the analysis of the data.

R Open source statistical programming language.

ROC Receiver Operator Characteristic; a formal description of discriminative ability.

Staircasing Tuning the difficulty of a task adaptively based on performance to achieve a target level of difficulty.

Weight on Advice A measure of the relative contribution of advice (as compared to an initial estimate) to a final decision.

WoA Weight on Advice; see entry.

1
Introduction

Animals (and possibly other organisms) make decisions continually. They need to decide how to act, often on the basis of judgements on the state of the world. The average person will make hundreds of conscious decisions a day, and thousands of unconscious ones: what do I want for breakfast?; do I take an umbrella?; what should I get my child for their birthday?; is that other pedestrian going to bump into me?; and so on. Crucially, for many of these decision-makers, useful information can be obtained from other people: seeing what other diners are having in a restaurant; listening to a meteorologist forecast the weather; paying attention to your child's insistent requests. This kind of social information suffuses the world of human decision-making, and so too do judgements about its usefulness.

To use social information appropriately, people need to know whether the information is useful for them: it may be a mistake for someone looking for a lavatory in a theatre to join the longest queue on the assumption that it indicates the most popular choice. Beyond misinterpretation, there is the ever-present threat of duplicity: the history of cooperation is inextricably entwined with the history of free-loading, cheating, and deception. Who we trust, and how far, is a matter of great practical importance. This thesis investigates two questions in this area: whether we rely too much on those similar to us, and whether we rely too little on others in general.

Two key ideas in social influence concern echo chambers and egocentric discounting. Echo chambers are social environments where only one opinion is able to exist: everyone within the echo chamber expresses the same opinion, an opinion that is then 'echoed' back to them by others, making it appear ubiquitous and certain. Egocentric discounting is the name for the general observation that people do not use information from others as efficiently as they could.

This thesis explores two questions related to those ideas: does the way individuals seek advice produce echo chamber-like networks?; and is the well-established phenomenon of people 'irrationally under-weighting' advice actually explicable as a rational process? The first of these questions concerns how people learn about the trustworthiness of others who give them advice, especially where they have no objective way of knowing how useful that advice actually is. I ask whether people are more likely to seek advice from others whom they consider more trustworthy, and whether this might lead to forming echo chambers full of like-minded individuals who assess one another as trustworthy because they parrot their own opinions back again. This question is addressed using on-line empirical behavioural experiments to characterise the individual psychological processes underlying advice seeking, followed by agent-based computational simulations that implement the variation in those processes observed in human participants.

The second question takes as its starting point the observation that people do not take as much advice as they *should* (as defined by a mathematical model of information). There has been much discussion concerning the reasons for this, and I develop a contribution to this debate; the suggestion that most of the time it really is better to take a cautious approach to advice-taking to avoid being too vulnerable to ignorance, malice, and miscommunication. The question is addressed using computational agent-based evolutionary simulations to explore the adaptiveness of discounting advice under a range of circumstances. I then explore people's flexibility in responding to changes in those circumstances using on-line behavioural experiments. The organisation is as follows: this introduction establishes the core concepts invoked in this thesis, and describes their treatment in the literature. Then, the two questions are addressed in separate sections of the thesis. Each section includes a short introduction to the specific question addressed, a detailed description of the work conducted, and a short discussion of the conclusions drawn from the work. The final section offers broader conclusions arising as a consequence of the presented work, alongside some suggestions for future research.

In the remainder of this introduction, I introduce the key concepts that arise in the thesis. First, I introduce advice and highlight relevant features, including the measurement paradigm used in this work and a normative model that quantifies how advice *should* be used. Second, I briefly introduce the phenomena of interest. Finally, I develop the normative model with ideas of use of advice and advisor evaluation to form the framework within which the experiments and simulations take place.

1.1 Advice

Advice can be broadly defined as information which comes from a social source. Defined this way, it includes things most people would intuitively understand as advice, such as guidance from a mentor or lawn care tips, and things most people may not, such as advertisements, traffic, and recommender algorithms in on-line shops. Advice is different from other sources of information, such as the forbiddingness of a storm cloud or the rattle of a rattlesnake, in that it is the result of human mental processing of other information.[1] 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. This is generally within the marrow of what is understood by "advice", and the aim of the investigations herein are

[1] There are of course cases that defy this categorisation: for example, it seems uncontentious to allow that honeyguides and guide dogs offer advice to their human counterparts.

to shed light on the way humans exchange information with one another and the consequences of those exchanges.

1.1.1 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 operationalisations of advice-taking that the informational content of the advice determines the extent to which it is taken or rejected. For example, *weight on own estimate* and its derivatives quantify advice-taking as the amount an opinion is updated in the direction of advice (Yaniv and Kleinberger 2000). Insofar as the identity of the advisor matters, it matters because it functions as a cue to the informational content of the advice – for example where advice is more likely to be correct because it comes from someone with expert knowledge rather than someone who is guessing. 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 (Rader, Larrick, and Soll 2017; Ronayne and Sgroi 2018). 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 time-scales.

1.1.2 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, Davis, and Schoorman (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; Yonah and Kessler 2021)).

1.1.2.1 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 co-mingle 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, Moore, and MacCoun 2013; Schultze, Mojzisch, and Schulz-Hardt 2017; Sniezek, Schrah, and Dalal 2004; Sniezek and van Swol 2001; Soll and Mannes 2011), or are shown to be more expert empirically (Pescetelli and Yeung 2021; Sah, Moore, and MacCoun 2013; Yaniv and Kleinberger 2000).

1.1.2.2 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 renowned 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 participants are more inclined to reject advice when uncertainty is attributed to malice rather than ignorance (Schul and Peni 2015).

1.1.2.3 Integrity

Advisors with integrity exhibit adherence to principles which the decision-maker endorses. “Integrity” in this sense is related to the common-use sense of integrity as following a code or set of principles unwaveringly, but with the caveat that the principles have to be ones shared with or respected by the decision-maker. To borrow an example from Mayer, Davis, and Schoorman (1995), a ruthless dedication to “profit seeking at all costs” would lead to an advisor being highly mistrusted, except where the person doing the trusting also endorsed that value. While benevolence and integrity are 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). In many cases, integrity is difficult to disentangle from ability, because subscription to values often indicates a knowledge of a domain, which can act as a cue for ability. Where professional societies are involved, for example, many require formally-recognised training for membership. Providing a concise and accurate definition of integrity, complete with appropriate qualifiers and explanations, is well beyond the scope of this thesis: integrity has long suffered from conceptual confusion and disagreement between researchers (Palanski and Yammarino 2007), with recent consensus perhaps emerging that it is best understood as a multi-dimensional construct (Moorman, Blakely, and Darnold 2018) that may or may not bear close resemblance to the construct imagined by Mayer, Davis, and Schoorman (1995).

Within the scope of this thesis, integrity can be thought of as a commitment to fulfilling one’s role in an experiment, in the case of advisors giving advice according to stated goals (usually doing one's best to help the decision-maker). In combination with benevolence, integrity acts to determine the extent to which the helpfulness of advice approaches the limit imposed by the advisor's ability.

1.1.3 Source selection and advice-taking

From the perspective of this thesis, the two key components of advice are its source and its informational content. The underlying framework§1.3[2] used here views the informational content of advice as providing information not only about the external world and the best way to act, but also as providing information about the source of advice itself. Both the advice itself and information about its source can contribute to the likelihood of the advice being incorporated into a final decision. The literatures on source selection (where we look for advice) and advice-taking (how we use advice when it is provided) are grounded in different academic traditions. Source selection is usually studied within the context of Social and Personality Psychology, with choices viewed as minimising cognitive dissonance (Festinger 1957) or protecting a positive self-concept (Knobloch-Westerwick 2015). Advice-taking, on the other hand, is formally investigated in the Organisational and Cognitive Psychology literatures on judgement and decision-making and forecasting, and frames behaviour with reference to normative models§1.1.5.

In this work, I take steps towards joining these literatures by extending a framework for advisor evaluation in the absence of feedback§1.3 to the domain of source selection, referred to here as advisor choice. This thesis thus uses an approach more similar to the advice-taking literature than to the source selection literature.

1.1.4 Judge-advisor system

The paradigm used for the behavioural experiments and as a framing for the computational simulations is a Judge-Advisor System (Yaniv and Kleinberger 2000; Sniezek, Schrah, and Dalal 2004). In this paradigm, two or more individuals contribute to making a decision, one as the judge and the other as the advisor. The

[2] See section 1.3. Throughout the thesis, internal links are abbreviated in this manner.

judge makes an initial estimate alone, then receives advice from the advisor, and then makes a final decision integrating the advice with their initial estimate as they see fit. This paradigm allows advice to be quantified on the assumption that any systematic differences between initial estimates and final decisions are properties of the advice.

In theory, any decision-making task can be used for a Judge-Advisor System provided it allows for the judge to make two decisions and advisors to provide plausible advice. Estimation tasks are commonly-used in the literature, such as estimating life expectancies in difference countries (Trouche, Johansson, et al. 2018), the dates of historical events (Yaniv and Kleinberger 2000), the price of an item from its features (Sniezek, Schrah, and Dalal 2004), people's weights (Gino and Moore 2007), sports teams' performances (Soll and Mannes 2011), opinion prevalence rates (Liberman et al. 2012), and calorie values of foods (Yaniv and Choshen-Hillel 2012). Other studies have used perceptual estimation tasks requiring visual perceptual decisions about displays of dots (Pescetelli and Yeung 2021; Rouault, Dayan, and Fleming 2019), or the presence or absence of a visual target (Mahmoodi et al. 2015); these tasks have less obvious ecological validity but afford precise experimental control of participants' performance and advice accuracy. The behavioural experiments in this thesis use historical date estimation and perceptual dot-discrimination tasks.

1.1.5 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 (i) as the true weight (v) plus some error (ϵ):

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

$$\epsilon = \mathcal{N}(\mu = 0,\, \sigma^2) \tag{1.2}$$

The key insight is to observe that the error is drawn from a normal distribution (Equation 1.2)[3]. 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.

$$\frac{\sum_a^N (i_a)}{N} = \frac{\sum_a^N (v + \mathcal{N}(\mu = 0,\, \sigma_a^2))}{N} \tag{1.3}$$

$$\frac{\sum_a^N (i_a)}{N} = \frac{\sum_a^N (v)}{N} + \frac{\sum_a^N (\mathcal{N}(\mu = 0,\, \sigma_a^2))}{N} \tag{1.4}$$

$$\frac{\sum_a^N (i_a)}{N} = \frac{Nv}{N} + \hat{0} \tag{1.5}$$

$$\frac{\sum_a^N (i_a)}{N} \approx v \tag{1.6}$$

Observe that this formulation is true no matter the value of N. On average, it is always better to have more estimates than fewer, because as the number of estimates increases the sum of their errors approaches 0. Even in the 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.

Where variances of the error distributions are known (σ_i^2), estimates can be weighted by those variances:

$$f = \frac{\sum_a^N (\omega_a i_a)}{\sum_a^N \omega_a} \tag{1.7}$$

Where ω_a is $1/\sigma_a^2$. This will increase the accuracy of the estimates in proportion to the difference between the variances.

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

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

$$\begin{aligned} f &= \frac{\sum_a^N (i_a)}{N} && (1.8) \\ &\approx v && (1.9) \end{aligned}$$

This framework provides a formal basis for understanding and assessing how information is integrated in social and group scenarios. As a normative model, the framework provides an optimal strategy for using advice by modelling how advice can improve a decision-makers accuracy. Additionally, the framework formalises the notion of the "wisdom of crowds" – the observation that groups can often come to decisions that are better than the best individual's decision. Below, I show how this framework extends to cover the use of advice where agents have systematically different capabilities§1.3.1.

1.2 Advice-taking phenomena

This work is primarily interested in two advice-taking phenomena: one in which people weigh advice too heavily; and one in which people weigh advice too lightly. In one of these, people form echo chambers in which they are over-reliant on the advice of those like themselves, leading to missed opportunities to make accurate assessments. In the other, egocentric discounting, people take less advice than they should, according to the normative model presented above§1.1.5 when integrating multiple opinions with their own.

1.2.1 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, Smith-Lovin, and Cook 2001). McPherson, Smith-Lovin, and Cook (2001) review a wealth of examples, including the tendency for married couples to be more similar to one another than chance would dictate in their ethnicity, age, and religion (and along many other dimensions). The use of homophily as a concept has increased dramatically in the last few decades, across a range of different research fields. Different fields (and researchers) provide different measurements, such as asking people to nominate their friends and comparing demographic information (McCormick et al. 2015) or comparing social media users' content similarity and interactivity (Cardoso et al. 2017), but the unifying idea is that more similar individuals are more likely to have closer contact with one another (Lawrence and Shah 2020).

Homophily is often discussed negatively, especially in the context of ethnic differences and intergroup tensions (McPherson, Smith-Lovin, and Cook 2001), but in many contexts it is not only expected but desired. Homophily on the basis of shared interests is unavoidable: provided people spend time doing things they are interested in, such as playing a sport, they will tend to spend time with other people doing those things. Similarly, many romantic matchmaking services deliberately use similarity on dimensions such as age to identify likely partnerships. Due to association between the many dimensions in which homophily operates, such as ethnicity and religion, it is often difficult in practice to separate out the dimensions along which homophily is operating.

Social influence processes combine with homophily to create a self-reinforcing spiral: individuals who are more similar to one another are more likely to associate with one another; and individuals who associate with one another are more likely to become more similar to one another. Again, this can be benign, as when newcomers to a social group are welcomed and absorb the social norms of interaction. Where this cycle generates concern is when reciprocal influence produces increasingly extreme attributes, for example political opinions. This process is known as "polarisation".

1. Introduction

Polarisation is the phenomenon whereby the distribution of opinions in a population trend toward extremes over time, with moderate stances becoming increasingly less popular.[4] Polarisation can be one-way, as when a consensus forms on a topic and that consensus shifts the entire ground of the debate (as with the core feminist principle that women and men deserve equal respect), or two-way as with divisive issues like the suitability of former President Donald Trump for office. While in both cases the whole population has acquired an extreme opinion, in the former the population remains united whereas in the latter the population is deeply divided, with little middle ground.

Theoretically, homophily may lead to polarisation because it isolates people from the checks and balances of the real world. Within an echo chamber inhabited by those who have similar values, experiences, and opinions to ourselves, it becomes easy to miss the drawbacks to new ideas. We receive positive feedback for opinions of a certain stripe, and join in with applauding them when we hear them. More extreme ideas in keeping with the views appear pioneering and courageous, while more moderate ideas become old-fashioned.

Most people who argue for the nefarious consequences of homophily and polarisation for society as a whole highlight social media, and the fluency of association that it offers, as a catalyst that sweeps away many traditional checks to the vicious cycle of homophily and polarisation. Whereas in the physical world we can only interact with those nearby, in the on-line world we can interact with people anywhere in the world, meaning we can find people whose opinions are far more similar to our own along far more dimensions (Davies 2017). Added to this, the algorithms that curate content on dominant social media platforms prioritise emotive content, selecting extremes of opinion to be reviled or applauded, and replace the previously universally-shared culture with a culture shared only by our echo chamber (Sunstein 2002; Sunstein 2018). Modelling work demonstrates that where there is a bias in assimilation of information, homophily exacerbates

[4] There are a variety of precise and somewhat different definitions in the literature (Bramson et al. 2016).

polarisation (Dandekar, Goel, and Lee 2013). Where polarisation in turn increases homophily, for example through selective exposure (where individuals seek out reinforcing views) or avoidance (where individuals avoid challenges to their views), a self-reinforcing spiral emerges wherein social connections become increasingly homogeneous and attitudes increasingly extreme (Song and Boomgaarden 2017).

Despite this compelling picture, whether homophily in on-line social networks actually *is* responsible for increases in polarisation is debated. Proponents of the idea note that political polarisation has increased in recent years (Perrett 2021; Boxell, Gentzkow, and Shapiro 2017); opponents note that this is not a new phenomenon (Perrett 2021) and that the sections of society polarising most rapidly are the least likely to use social media (Boxell, Gentzkow, and Shapiro 2017). Experimental investigation has shown that corroboration may increase the extremity of opinions (Baron et al. 1996; Schkade, Sunstein, and Hastie 2010), and this may outweigh the mollifying effect of contradiction (Lord, Ross, and Lepper 1979); but Grönlund, Herne, and Setälä (2015) found decreased extremism of opinion among anti-immigration participants after group discussion. Empirical studies demonstrate homophily in on-line social networks (Cardoso et al. 2017; Colleoni, Rozza, and Arvidsson 2014); but Barberá (2015) argue that these social networks are *less* homophilic than their off-line equivalents, and hence should *reduce* polarisation. There are many egregious examples of pathologically polarised insular on-line communities, but these are either pre-existing groups that have moved on-line (such as creationist communities), or new groups with deep similarities to old ones (such as Q-Anon and previous conspiracy theorist communities). Selective exposure is a potential driving mechanism for polarisation (Kobayashi and Ikeda 2009); although criticism of the idea goes back a long way (Sears and Freedman 1967) and the emerging consensus among researchers seems to be that while individuals may preferentially seek out reinforcing information they are unlikely to selectively avoid information that would challenge their views (Garrett 2009a; Garrett 2009b; Nelson and Webster 2017).

The discrepancies between computational models that demonstrate polarisation and empirical evidence that paints a mixed picture may be due to people's tendencies for selective exposure and selective assimilation of information being lower than those assumed by the models. Inter-individual variation may also contribute to this puzzle. This thesis explores biased source selection, its heterogeneity, and their impact on the way information flows through social networks.

1.2.2 Egocentric discounting

From the perspective of the normative model above§1.1.5, decision-makers should weigh their own estimate equally with each other estimate they receive in the process of coming to their decision. This is because the errors in the judgements made by the decision-maker and the advisors are interchangeable, mathematically-speaking, and on average one will approximate the correct answer most nearly by allowing all of the errors to cancel one another out, which is achieved by averaging across all estimates as show in Equation 1.6. However, one of the most robust findings in the literature on advice-taking is that people routinely under-weight 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, Larrick, and Soll 2017; Ronayne and Sgroi 2018; See et al. 2011; Soll and Mannes 2011; Trouche, Johansson, et al. 2018; Yaniv and Kleinberger 2000; Yaniv and Choshen-Hillel 2012; Yaniv and Milyavsky 2007).

Yaniv and Kleinberger (2000) provided a classic example of egocentric discounting in a task that required participants to provide initial estimates of historical dates (e.g. "When were the Dead Sea Scrolls first discovered?"), and then provide final decisions of the same after seeing advice (taken from a random participant's answer in a previous study). Despite both the initial estimates and the advice estimates being similarly accurate, participants' final decisions were much closer to their initial estimates than to the advice. Instead of producing decisions that were midway between the two estimates, as would be achieved with mathematically optimal averaging, participants weighted their initial estimate about 70:30 with the advice they received.

Egocentric discounting occurs in both feedback and no-feedback contexts (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 (Yaniv and Kleinberger 2000); overrate their own competence (Sniezek, Schrah, and Dalal 2004); may have a desire to appear consistent (Yaniv and Milyavsky 2007); may see opinions as possessions (Soll and Mannes 2011); may be loss-averse to providing a worse final decision due to advice-taking (Soll and Mannes 2011); or have difficulty avoiding anchoring (Schultze, Mojzisch, and Schulz-Hardt 2017) or repetition bias effects (Trouche, Johansson, 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, Johansson, et al. (2018) arguing that the potential for misaligned incentives between decision-maker and advisor motivate discounting of advice. The latter part of this thesis§6 explores whether egocentric discounting may be a stable meta-strategy which protects against exploitation, carelessness, incompetence, and miscommunication. From this perspective, the normative model is only normative within the very particular scenario it describes. When we broaden the scope to consider how that particular scenario fits into a background of advice-taking in humans as the product of genetic and cultural evolution and individual experience, we find that the normative model must be altered to account for these features.

1.3 Conceptual framework

How seriously advice is taken is a consequence of how much a judge trusts it. The framework underlying this thesis regards advice as being evaluated according to two key properties. The first of these is the *content* of the advice itself, i.e. how plausible it is given other evidence. The second property is the *source* of the advice, i.e. how trustworthy the judge considers the advisor to be. Thus, while we might trust a friendly-looking stranger as much as a meteorologist when they opine that the sunshine will hold out for the rest of the afternoon, we might be much more likely to trust the meteorologist if the forecast is that the blazing sunshine will turn into a thunderstorm.

Our framework additionally proposes that the reputation of a source of advice is built up over time as a function of the advice that the source provides. There are numerous factors that contribute to an advisor's trustworthiness, such as sobriety and professional expertise, and many of these can change over time, but this framework focuses on reputation as a function of advice quality. Reputation roughly captures the Mayer, Davis, and Schoorman (1995) dimension of *ability*, although it can incorporate other dimensions, too. The mechanism of trust updating differs slightly according to the availability of objective feedback.

In the case where advice is given on a task that has an immediately-verifiable answer, the utility of the advice can be evaluated on the basis of feedback and the evaluation of the advisor updated accordingly. If I am trying to remember whether I left my phone in my coat or my bag, and my partner tells me that it is in my bag, a brief examination of one or both of the potential locations will not only find my phone but will also allow me to evaluate the accuracy of my partner's advice. Over multiple interactions on these kinds of decisions, my partner will acquire a reputation in my mind as a relatively reliable or unreliable source of this kind of information.

Where feedback is unavailable, people may use their own sense of certainty as a yardstick for evaluating advice (Pescetelli and Yeung 2021): 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. Perhaps instead of a lost phone, my partner gives me advice on the name of an old acquaintance of ours: although we cannot verify the information, if it 'rings a bell' I may be highly confident my partner is correct, and judge the utility of the advice accordingly. In this way, confidence serves as a proxy for objective feedback. Confidence 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).

1.3.1 Use of advice

Normative models of advice-taking§1.1.5 state that averaging estimates minimises errors. As discussed at length later§5, the assumptions underlying the normative model do not always hold in the real world, but this normative framework is a useful starting point for considering how advice can or should be used. The performance of the normative model can be characterised according to differences between the advisor and the judge on ability and bias (Soll and Larrick 2009), proving effective for the kinds of decisions on which the errors made by judge and advisor are independent of one another. Equation 1.7 for our normative model states that advice should contribute to the final decision in proportion to the ability of the advisor compared to the judge. In the two-estimate case (initial estimate and advice), this can be expressed as:

$$f = \frac{\omega_a i_a + \omega_{a'} i_{a'}}{\omega_a + \omega_{a'}} \tag{1.10}$$

Where agent a is the judge, i_a the judge's initial estimate, agent a' is the advisor, $i_{a'}$ is the advice received (i.e. the initial estimate of agent a'), and ω_a and $\omega_{a'}$ the judge's weighting of their own and their advisor's answers, respectively. This weighting can be simplified to be expressed only in terms of the judge's weighting of the advisor because the two are constrained to sum to 1 by virtue of being relative to one another:

$$\omega_a + \omega_{a'} = 1 \tag{1.11}$$

$$\omega_{a'} = 1 - \omega_a \tag{1.12}$$

$$\therefore f_a = (1 - \omega_{a'}) i_a + \omega_{a'} i_{a'} \tag{1.13}$$

In the normative model, the weighting is equivalent to the ratio of variance of the errors made by each agent:

$$\omega_{a'} = \frac{\sigma_{a'}^2}{\sigma_{a'}^2 + \sigma_a^2} \tag{1.14}$$

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 later§5, 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§1.1.2 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, benvolence, and integrity.

$$\text{trust} \propto \text{ability} \cdot \text{benvolence} \cdot \text{integrity} \tag{1.15}$$

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

$$\omega_{a'} = \frac{\text{trust}_{a'}}{\text{trust}_{a'} + \text{trust}_a} \tag{1.16}$$

At this point we may question whether the variable trust_a (i.e., trust in one's own opinion) 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_a and integrity_a will be anything less than maximal, perceptions of one's own ability (ability_a) 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:[5]

$$\omega_{a'} \propto \frac{\text{ability}_{a'} \cdot \text{benvolence}_{a'} \cdot \text{integrity}_{a'}}{\text{ability}_{a'} \cdot \text{benvolence}_{a'} \cdot \text{integrity}_{a'} + \text{ability}_{a} \cdot \text{benvolence}_{a} \cdot \text{integrity}_{a}} \tag{1.17}$$

$$\propto \frac{\text{ability}_{a'} \cdot \text{benvolence}_{a'} \cdot \text{integrity}_{a'}}{\text{ability}_{a'} \cdot \text{benvolence}_{a'} \cdot \text{integrity}_{a'} + \text{ability}_{a}} \tag{1.18}$$

This model fits well with the variables manipulated throughout this thesis: increases in task difficulty and decreases in subjective confidence will alter the perception of ability$_a$, leading to greater advice-taking. Likewise, increases in perceived benevolence, integrity, or ability of an advisor will lead to greater advice-taking.

1.3.1.1 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 so that the work presented in this thesis can be judged and interpreted within its limitations.

Firstly, the model is an idealised situation approximated by the experimental method§1.1.4: a judge makes an explicit initial estimate, then receives advice, then makes an explicit final decision. Yaniv and Choshen-Hillel (2012) showed that preventing judges from making initial estimates 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 empirically performs poorly. At best, it could be argued that pre-exposure to the advice either anchors the initial estimate (thus moving

[5] If the concept of self-trust still appears untenable, note that trust$_a$ can be replaced with a constant without compromising the equations.

i_a systematically closer to $i_{a'}$), or that having to trust advice because one cannot make one's own decision inflates the weighting of the advisor.

Secondly, the model does not perform well when multiple advisors are consulted. The normative model, and the psychological derivative, predicts that a judge'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. Hütter and Ache (2016) presented evidence that this does not happen: the weight of the initial estimate stays relatively constant while the weights of the advisor estimates are reduced. This implies that if a judge 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 estimate remained constant ($\omega_i = .5$; $\omega_{j \neq i} = .25$), rather than the more transparently optimal policy of weighting all estimates evenly ($\omega_i = \omega_{j \neq i} = 1/3$). Similarly, Yonah and Kessler (2021) showed that, while increasing the number of advisors from whom an estimate is drawn from 20 to 200 does increase the weight placed on advice a little, it is nowhere near the level that would be expected normatively.

Finally, the model is supported by experiments that present advice-taking in terms of averages over several trials. These averages can obscure very different behaviours on a trial-by-trial basis: the same average advice weight of 50% would appear for a set of trials where initial estimate and advice were consistently evenly weighted as well as for a set of trials where the judge alternated between keeping their initial estimate and wholly adopting the advice. 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, whereby one or other answer is wholly adopted (Soll and Mannes 2011; Soll and Larrick 2009). 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.

The model could be extended to incorporate this distinction between picking and averaging behaviour by building in an additional parameter governing the likelihood of picking as opposed to averaging on a given trial. In the present model these two potential parameters, whether to pick or average and how much to average, are collapsed into one parameter governing the extent of averaging. Such additional complexity is not warranted here, but I highlight the distinction to allow readers to consider for themselves.

1.3.1.2 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. This work seeks 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 inexact descriptions of reality, 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.

1.3.2 Updating advisor weights

The weights assigned to the advisors (relative to the judge 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 (Önkal, Gönül, et al. 2017; Tost, Gino, and Larrick 2012; Schultze, Mojzisch, and Schulz-Hardt 2017) or have some summary of their performance displayed (Gino, Brooks, and Schweitzer 2012; Yaniv and Kleinberger 2000). Endogenous experience refers to the information that advice on a given trial carries about the trustworthiness of an advisor, and forms the basis of the Pescetelli and Yeung (2021) model used here.

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 simplifying 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; in this thesis only cursory attempts are made to manipulate perceptions of dimensions other than ability (e.g. Experiments 5§7.1.1 and 6§7.1.2).

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.* People may alter their advice-giving behaviour in anticipation of discounting on the part of the judge (Renault, Solan, and Vieille 2013; Azaria et al. 2016), somewhat akin to starting negotiations with a higher demand than one is hoping to settle for, in which case this assumption would not be wholly true. There is no evidence as yet as to whether people do this, and whether judges 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.

1.3.2.1 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 (Yaniv 2004)), while the properties of the advisor will include the advisor's trustworthiness (see above§1.3.1). The updating of trust following experience of advice is likely to be largely in the domain of ability§1.1.2.1, although other domains may be affected where the advice is particularly egregious.

1.3.3 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 and Yeung 2021; Sah, Moore, and MacCoun 2013; 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 ($\widehat{\text{ability}}_{a,a'}$) rather than one's own beliefs about the world, according to some learning rate (λ).

$$\widehat{\text{ability}}^{a,a'}_{t+1} = (1-\lambda)\cdot\widehat{\text{ability}}^{a,a'}_{t} + |i^{a'}_{t} - v_t|\cdot\lambda \tag{1.19}$$

1.3.3.1 Criticism of the advisor evalutation model

While many experiments have established the existence of 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. FeldmanHall and Dunsmoor (2019) argued in a review that a wide variety of social phenomena could be explained via reinforcement learning processes, and Behrens et al. (2008) demonstrated that a Bayesian reinforcement learning model provided a good fit to behavioural data from a social influence task. Additionally, Heyes et al. (2020) have argued that social learning is wholly explicable in terms of general reinforcement learning processes paired with attentional biases to social stimuli. 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 judge's confidence for objective feedback. While not foolproof, the method allows better-than-average approximation of the quality of advisors provided several plausible assumptions are met (Pescetelli and Yeung 2021).

1.3.4 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, Hauperich, and Yeung 2021). 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 (2021) suggest the mechanism for this ability to discriminate between advisors in the absence of feedback is performing updates based on confidence-weighted agreement.

1.3.4.1 Agreement

Consider first the non-weighted agreement case, where the advisor's estimate ($i_t^{a'}$) and the judge's estimate (i_t^a) at time t are binary ($\in 0, 1$).[6] The estimate of the advisor's ability ($\widehat{\text{ability}}^{a,a'}$) is updated positively if the advisor and judge agree, and negatively otherwise, according to the learning rate λ.

$$\widehat{\text{ability}}_{t+1}^{a,a'} = \begin{cases} (1-\lambda)\cdot\widehat{\text{ability}}_t^{a,a'} + \lambda, & i_t^a - i_t^{a'} = 0 \\ (1-\lambda)\cdot\widehat{\text{ability}}_t^{a,a'} - \lambda, & i_t^a - i_t^{a'} \neq 0 \end{cases} \tag{1.20}$$

[6] Where decisions are not binary, for example selecting between multiple options, the consequences of disagreement are less clear, but the qualitative insights concerning agreement hold.

This agreement heuristic generally is quite useful: the likelihood of an independent other agreeing with you is monotonically related to their accuracy, so you can learn something about the accuracy of others simply by seeing how often they agree with you. This holds provided you are more accurate than chance: if you are less accurate than chance then the more accurate an advisor is the more likely they will *disagree* with you. This is because the probability that you and your advisor agree depends on both their accuracy and your accuracy.

Confidence-weighted agreement The insight that the usefulness of agreement as a proxy for accuracy scales with the judge's own probability of being correct means judges may be able to gain insights into that usefulness. The judge's accuracy varies from decision to decision, and this variation means that some decisions are accurate (and hence agreement is a useful proxy for advisor accuracy) while others are inaccurate (and hence agreement is a poor proxy). Insofar as a judge has insight into whether or not their decisions are more or less accurate, they have insight into whether or not agreement is a useful proxy for advisor accuracy. The judge's own confidence in their decisions is a metacognitive signal that, for well-calibrated judges, can serve this purpose.

Thus, the updating of advice contingent on agreement may be weighted by confidence in the initial estimate (c_t^a), 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.

$$\widehat{\text{ability}}_{t+1}^{a,a'} = \begin{cases} (1-\lambda)\cdot\widehat{\text{ability}}_t^{a,a'} + c_t^a\lambda, & i_t^a - i_t^{a'} = 0 \\ (1-\lambda)\cdot\widehat{\text{ability}}_t^{a,a'} - c_t^a\lambda, & i_t^a - i_t^{a'} \neq 0 \end{cases} \tag{1.21}$$

A well-calibrated judge who adopts this approach can thus exploit their insight into their own performance to improve their assessments of their advisors. On those decisions where the judge is most likely to be correct they will be more confident, and will therefore (rightly) take agreement and disagreement to be more diagnostic of their advisor's ability.

1.3.5 Use of the framework

This model of agreement-dependent advisor evaluation in the absence of feedback (whether confidence-weighted or non-confidence-weighted) forms the framework on which the experiments and computational models in this thesis are based. Framed in these terms, we can think of egocentric discounting as systematic under-weighting of advice from others relative to self, and ask why this might occur. We can think of echo chambers as systematically over-weighting advice from others who tend to agree, and explore why this occurs.

I present behavioural experiments that aim to explore the validity of this framework, and computational models that explore its implications. The behavioural experiments do not offer severe tests of the framework because they are primarily concerned with advisor choice: a lack of evidence for a preference between advisors on any given experiment may be explicable by a failure to translate a higher assessment of an advisor's advice into a preference for selecting that advisor rather than a failure to acquire a higher assessment of an advisor's advice.

1.3.6 Thesis structure

Chapters 2-4 consider the psychology of advisor choice, asking how people choose advisors and how these choices impact network-level dynamics. Chapter 2 introduces the general methodology used in the behavioural experiments and the analytical approach. Chapter 3 presents behavioural experiments that use both the previous perceptual decision-making task and the new general knowledge estimation task to investigate advisor choice. Chapter 4 presents computational simulations guided by the results presented in Chapter 4 that explore the impact of individual-level advice-seeking and advice-taking tendencies on overall network dynamics.

Chapters 5-7 explore advice-taking more closely, asking whether under-weighting advice could be rationally motivated. Chapter 5 reviews the literature on egocentric discounting, and introduces our perspective. Chapter 6 presents computational simulations that illustrate the adaptiveness of egocentric discounting as a response to several plausible features of the advice-taking context. Chapter 7 presents behavioural experiments that explore whether people can flexibly response to changes in the features modelled in Chapter 7.

Finally, Chapter 8 concludes the thesis with a short summary of results and interpretations. The broader implications of the work are considered, alongside its generalisability and limitations.

2
General 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 reported in subsequent chapters have truncated methods sections in which the specific deviations from the general method are noted.

2.1 Behavioural experiments

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.

2.1.1 Participants

2.1.1.1 Recruitment

Human participants were recruited from the on-line 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.

2.1.1.2 Payment

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

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 failed these attention checks were not paid. There is an ongoing ethical debate concerning non-payment of participants who fail attention checks in on-line studies. In on-line 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.

2.1.1.3 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 (Criado Perez 2019), initial concerns over General Data Protection Regulation resulted in a cautious approach to the collection of data concerning protected characteristics of participants. 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.

2.1.2 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).

2.1.3 Procedure

Participants visited the Uniform Resource Locator (URL) for the study by following a link from Prolific using their own device (Figure 2.1). 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. For studies with conditions, participants were assigned to a condition using an algorithm that produced shuffled sequences of conditions. The study introduced the software 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 participation 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 2.2.

2.1.3.1 Perceptual decision (Dots task)

The Dots task is a two-alternative forced choice task that has been used in various forms for many experiments in our lab (Steinhauser and Yeung 2010; Boldt

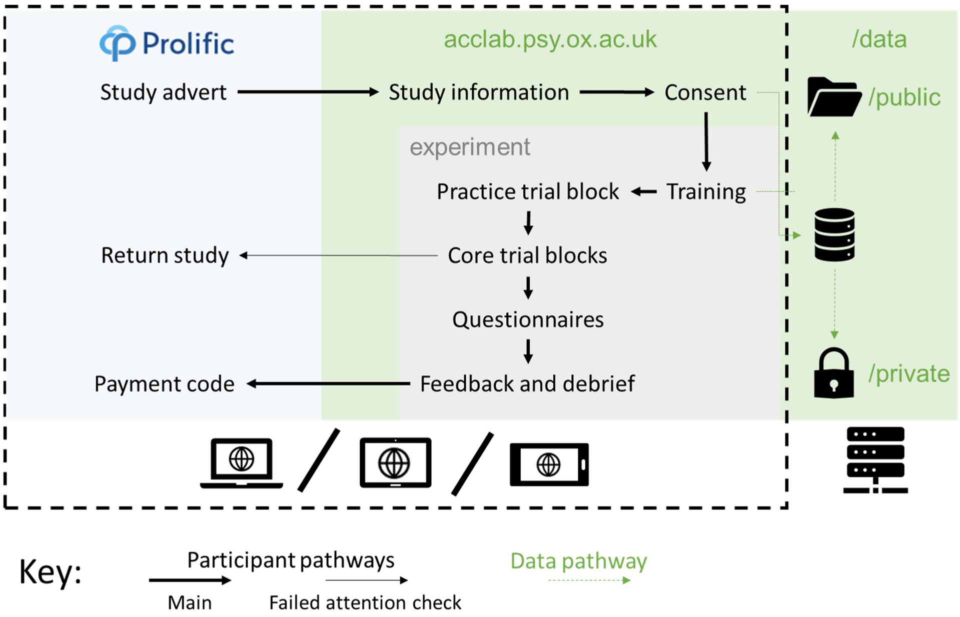


Figure 2.1: Participant pathway through the studies.
Participants used their own devices to complete the study, which was presented on a website written in HTML, CSS, and JavaScript. The data were saved on the server using PHP.

and Yeung 2015; Charles and Yeung 2019; Carlebach and Yeung 2020; Pescetelli, Hauperich, and Yeung 2021; Pescetelli and Yeung 2021) and beyond (Rouault, Dayan, and Fleming 2019; Rouault, Seow, et al. 2018; Fleming, Ryu, et al. 2014). This task has several key features that make it appealing for studying decision-making and advice. Firstly, stimuli are only presented very briefly, meaning that many trials can be performed in a relatively short experimental session. Secondly, the difficulty of the task can be titrated to bring all participants to a desired level of accuracy, and this process can be done continually and unobtrusively throughout the experiment. Thirdly, there is a good level of variation in the difficulty of individual trials due to the specifics of the random patterns generated, meaning that the task produces an array of confidence judgements while controlling overall performance. Fourthly, the variation in subjective experience of the difficulty of objectively similar trials means that advice is plausible: someone else really could have seen the presentation more clearly than the participant. Lastly, the task is neutrally-valenced and unlikely to provoke strong associations in participants that

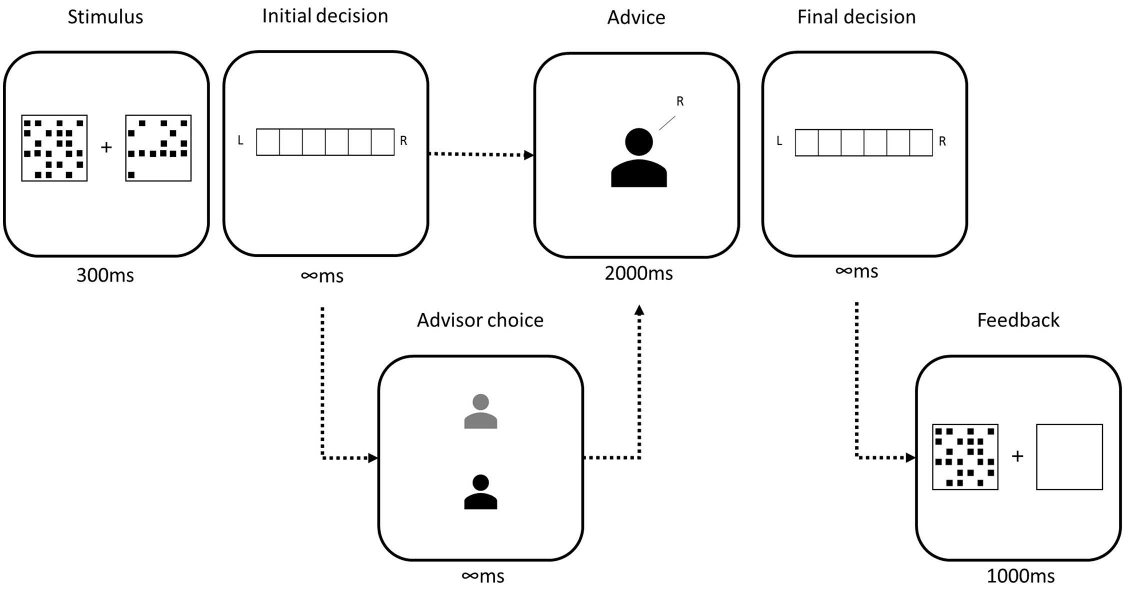
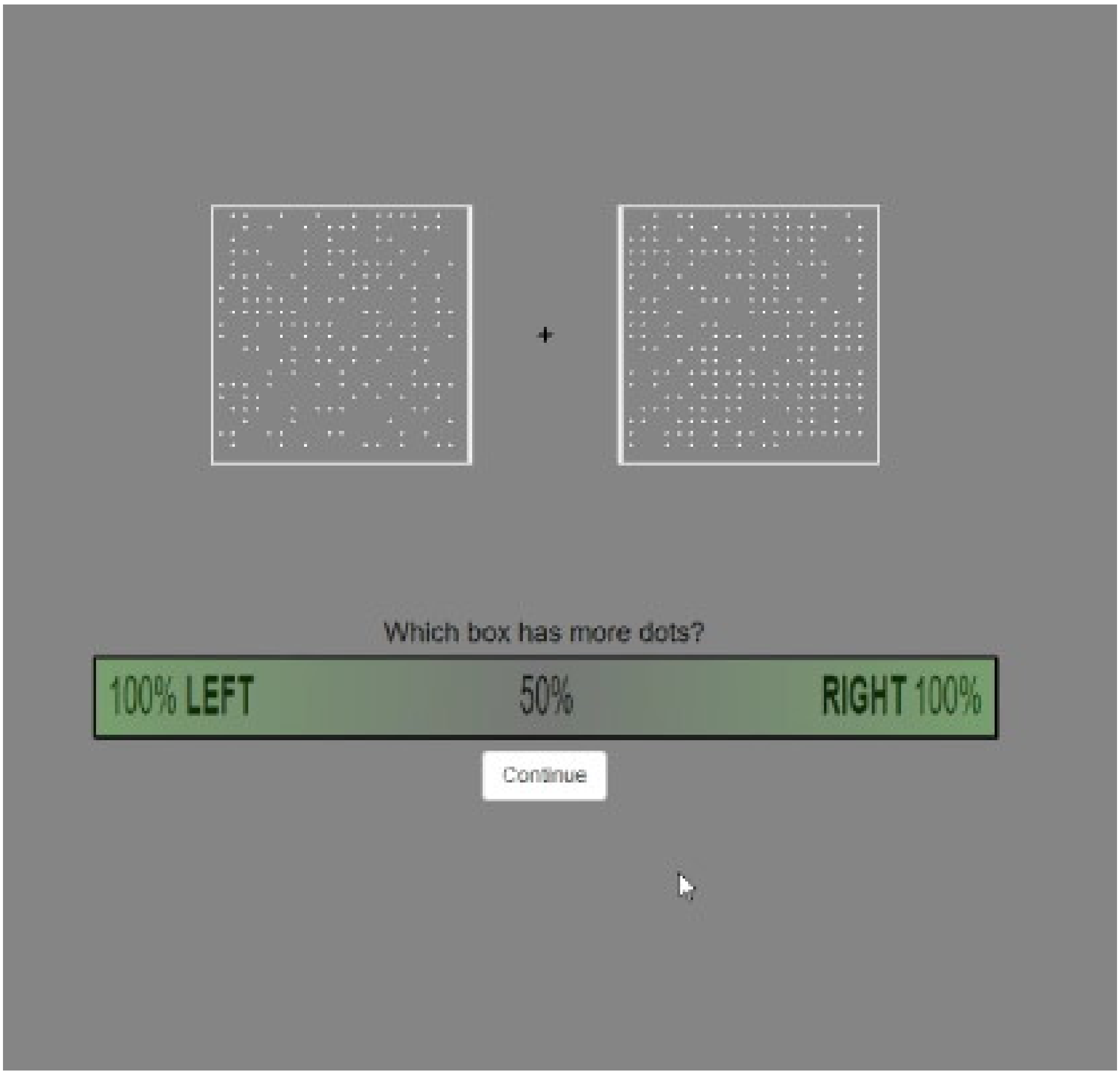


Figure 2.2: Trial structure of the Dots task.
In the initial estimate phase, participants saw two boxes of dots presented simultaneously for 300ms. Participants then reported whether there were more dots on in the left or the right box, and how confident they were in this decision. Participants then 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.

might alter their processing of information.

Stimuli in the Dots task consisted of two boxes arranged to the left and right of a fixation cross (Figure 2.3). These boxes were briefly (300ms) and simultaneously filled with an array of non-overlapping dots, and the participant was instructed to identify the box with the most dots. The participant submitted their response by selecting a point on a horizontal bar: points to the left of the midpoint indicated the participant thought the left-hand box had more dots, and points to the right that they thought the right-hand box had more dots. The further away from the midpoint the participant selected on the bar, the more confidence they indicated in their response.

The number of dots was exactly determined by the difficulty of the trial: the box with the 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.



Which box has more dots?

Figure 2.3: Dots task 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 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 71% using a 2-down-1-up staircase procedure. There were a substantial number 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. The specific number of practice trials differed between experiments as we sought to balance minimising the participant time requirement with the need to minimise practice effects in the core experimental blocks.

After the practice there were blocks where participants received advice on their initial estimates and made a final decision using the same response process as their initial estimate. Advice was presented with a representation of an advisor with a text bubble stretching out to the side the advisor endorsed, containing the text "I think it was on the RIGHT" or LEFT as appropriate. The advisors were represented in various manners in various versions of the task, but predominantly in a format with a central box containing a generic blank portrait icon and text at the bottom with "Advisor ##" where ## was a number between 10 and 60.

On some trials participants could choose between potential advisors. In these cases, the potential advisors were positioned vertically in the centre of the screen and the participant clicked on the advisor from whom they wished to receive advice. There was never a time limit for selecting an advisor.

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 decisions they had provided were correct and allowed to take a short, self-paced break.

At the end of the experiment participants were presented with a questionnaire asking them to rate their advisors' likeability, ability, and benevolence (Mayer, Davis, and Schoorman 1995), and offering them the opportunity to provide free-text comments on the advisors and the experiment in general. These questionnaire responses are not analysed because they were generally uninformative. Some free text responses may be redacted in the open data to prevent participants being identifiable.

Specific limitations of the Dots task The Dots task uses a perceptual decision with a high number of trials. There is a long-standing debate concerning whether social learning processes are unique or merely the operation of general learning processes on social cues (Lockwood and Klein-Flügge 2020), and this structure makes this task especially unsuitable for addressing this issue. It plausible that participants respond to the advice in this task primarily through simple reinforcement learning rather than through specific social processes even if those processes do exist. We can therefore draw only tentative conclusions that we are capturing fluctuations in trust in advisors during this task, as opposed to capturing fluctuations in the association between advice and stimuli.

Whether or not the Dots task taps into social processes, both the task itself and the experimental structure are very different from most advice-taking and advisor evaluation in the real world. Perceptual decisions are rarely the subject of advice, and thus the central task is an unusual one for joint decision-making. The amount of exposure to advisors also greatly exceeds that which would be obtained over a far longer period of time in most real world situations. This much greater level of exposure risks investing effects with artificial importance: while it is a strength of experimental designs to magnify the effects they aim to study we must not let such magnification blind us to the real relevance of these effects within the complex and dynamic context of real life.

To accommodate some of the limitations of this task, we aimed to replicate results from the Dots task in another task that afforded less precise control but that perhaps captures more everyday decision-making and advice.

2.1.3.2 Estimation (Dates task)

The Dates task is a real-world general knowledge trivia task that requires participants to estimate the dates of 20th century events. Participants were presented with historical events that occurred in a specific year in the 20th century. Participants were then asked to either: a) drag a marker onto a timeline to indicate a range of years within which they thought the event occurred (continuous version); or b) state whether the event occurred before or after a specific date (binary version). Once they had made this initial estimate, they were presented with advice from an advisor and then asked to make a final decision. The advice took the same form as the initial estimate, a range of years in the first case and an indication of whether 'before' or 'after' was the correct answer in the second.

During practice trials (either with or without advice), participants received feedback on their answers. Some participants were placed in a Feedback condition where they received feedback on all trials except those where they were asked to choose an advisor to provide advice.

After completing all the trials, participants were presented with a feedback form for each advisor, requiring them to rate the advisor on their level of knowledge, helpfulness, and likeability, chosen to reflect the three aspects of trust identified by Mayer, Davis, and Schoorman (1995) (ability, integrity, and benevolence). Participants could also provide free-text responses containing further comments on the advisors. Participants then completed a general debrief form that told them: "There was a difference between the advisors. What do you think it was?" They also had the opportunity to add free text questions or comments about the experiment. The latter responses are not included in the shared data.

Finally, participants were given a screen that allowed them to inspect their performance and send links to the study to others (only participants directly invited on the Prolific platform were included for analysis and in the shared data). This final screen also contained the payment code participants should enter on Prolific.

Rationale There were several reasons behind the development of the Dates task. Firstly, we believed the theory that advisors are evaluated on the basis of agreement when objective information concerning accuracy is not available (Pescetelli and Yeung 2021) to be a general property of advice-taking and not constrained to the domain of perceptual decision-making. We therefore wanted to replicate the effects see by Pescetelli and Yeung (2021) in a different task domain.

Secondly, we wanted to replicate the effects in a task domain that was more similar to the kinds of decisions that people might seek advice about. While people do occasionally consult one another concerning perceptual decisions ("is that a bird or a plane?"), such consultation is rare, especially in comparison to the ubiquity of perceptual decision-making. We thus selected a more deliberative task.

Thirdly, much of the Judge-Advisor System and advice-taking literature has used tasks based on estimation (Sniezek, Schrah, and Dalal 2004; Gino and Moore 2007; See et al. 2011; Soll and Mannes 2011; Gino, Brooks, and Schweitzer 2012; Liberman et al. 2012; Minson and Mueller 2012; Tost, Gino, and Larrick 2012; Yaniv and Choshen-Hillel 2012; Bonner and Cadman 2014; Schultze, Rakotoarisoa, and Schulz-Hardt 2015; Hütter and Ache 2016; Schultze, Mojzisch, and Schulz-Hardt 2017; Trouche, Johansson, et al. 2018; Wang and Du 2018), and often specifically estimation of dates (Yaniv and Kleinberger 2000; Yaniv and Milyavsky 2007; Gino 2008). We felt that being able to replicate the effects seen by Pescetelli and Yeung (2021) in a task domain commonly used for evaluating advice-taking behaviour would be especially useful.

Lastly, we wanted to design a task in a way that would be engaging for participants. This decision did not point us towards date estimation specifically in the way the previous decisions did, but we did feel that it would be possible to make an engaging task along those lines. Subsequent to designing the task we discovered two tabletop games with similar designs, both of which are simple to learn and engaging to play (TimeLine, 2012; CONFIDENT?, 2018). In keeping with this approach, we decided to make the task as compact as possible, for which estimation questions are useful, as demonstrated by the typical number of questions in the estimation tasks in the literature.

Continuous In the continuous version of the Dates task, participants were shown a timeline below the event description (Figure 2.4). Below the timeline were the markers that they used to indicate their responses. In earlier versions of this task, multiple markers were available, each with a point value that decreased with the width of the marker. Markers widths were chosen to span odd numbers of years, meaning that they could mark an even number of years in an inclusive manner (e.g. in the example shown, a 3-year marker marks the years 1916-1918 inclusively).

Once participants had placed their marker they clicked the tick button in the bottom-right to confirm their choice.

Decisions in the continuous version of the Dates task did not have an explicit confidence rating. The participant's choice of a narrower or wider marker was taken as an implicit indication of their confidence. In order to score points, the participant's marker had to contain the correct year on the timeline, and thus the more confident a participant was about the answer the more likely they would be able to score points using a smaller marker, and consequently the higher their average return from using a smaller marker.

Advice in the continuous version of the task was presented using a one-second animation wherein the advisor's avatar slid along the timeline to the middle of the advisory estimate. The advisor's avatar had a marker attached. This marker covered a range of years, indicating that the advisor suggested the year of the event lay within that range. Once the advisor's advice had reached its target location, the advisor's advice remained visible throughout the remainder of the trial, and the participant could enter a final decision.

To enter a final decision the participant could either simply click their existing marker, confirming their initial estimate as their final decision, move the marker along the timeline, or, where multiple markers were available, drag a different marker onto the timeline. Once again, to confirm their response they clicked the tick button in the bottom-right.

Advice in the continuous version of the task was still conceptualised along dimensions of accuracy and agreement. While it is possible to operationalise these as binary properties, with accurate advice including all advice where the advisor's marker covered the correct year and agreeing advice including all advice where the advisor's maker intersected with the participant's initial estimate marker, we decided that this approach was prohibitively difficult to implement, especially where initial estimate markers were very wide. Instead, accuracy and agreement were taken as continuous properties, measured from the centre of the advisor's marker to the correct year (for accuracy) or the centre of the participant's initial estimate marker (agreement).

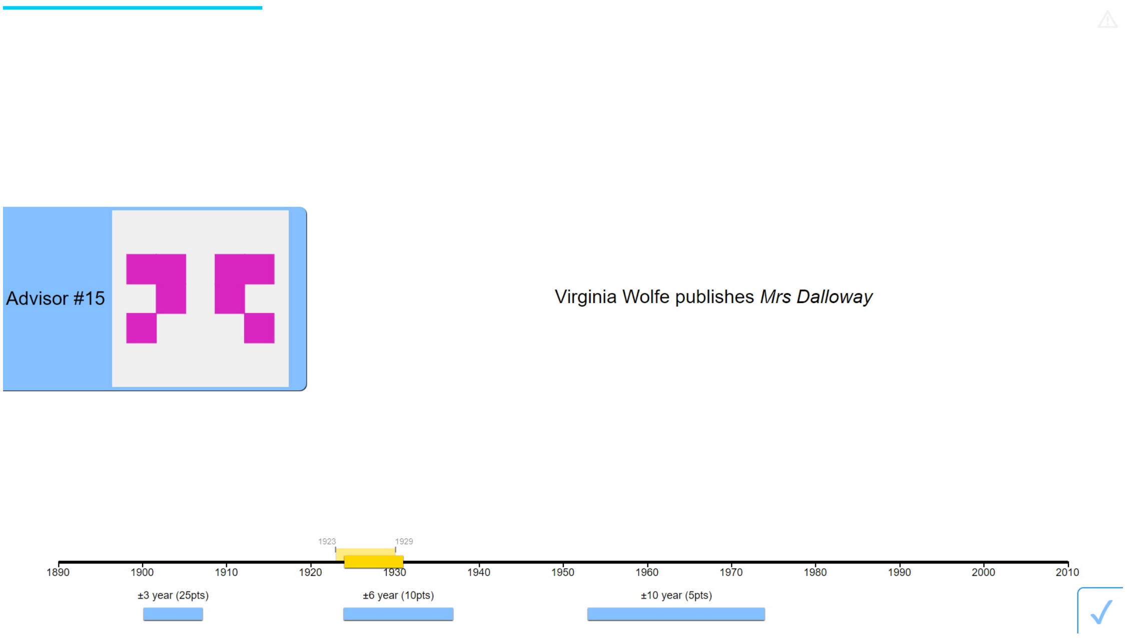
Advisors in these experiments typically placed their advice markers by identifying a target year, either the correct year (where the advisor was defined in terms of its objective accuracy) or the centre of the participant's initial estimate marker (where the advisor was defined in terms of its propensity to agree versus disagree with a participant's answer), and then placing the centre of their advice marker on a year sampled from a normal distribution around that target year. On occasion, advisors would offer 'Off-brand' advice, designed to neither be close to the participant's initial estimate nor the correct year, to allow for comparisons of influence across advisors that were not confounded with differences in the advice provided.

When a choice of advisors was offered, the two advisors were shown next to one another vertically on the left-hand side of the screen, and the participant clicked on the advisor from whom they wished to receive advice. There was never a time limit for making this choice.

On trials that contained feedback, the correct year was identified by the placement of a gold star placed over the timeline with a line pointing to the correct point on the timeline. The correct year was also displayed in text. This allowed participants to see both their own marker placement and the advisor's advice marker placement (if it was a trial with advice), and to compare both easily to the correct answer. The feedback screen lasted 2s before the next trial began.

Catch trials in the continuous Dates task consisted of an instruction to use the smallest marker to include a specific year, written out in words (e.g. 1942 would be "nineteen forty-two"). Participants who failed in this attention check task immediately failed the experiment, and were not allowed to continue.

Binary In the binary version of the Dates task, participants were shown an event and given an 'anchor' year. Their task was to correctly identify whether the event occurred before or after the date shown (Figure 2.5).



Virginia Wolfe publishes *Mrs Dalloway*

Figure 2.4: Dates task with continuous responses.

Below the event and anchor display were the answer bars used to make responses. Participants selected a point on the bar of their choice (the left-hand bar indicating they believed the event was before the anchor year, and the right-hand bar indicating they believed the event occurred after the anchor year). The higher the point they selected on the bar, the more confident their response. A blue bar appeared while they were choosing the height of their response, and the bar remained through the trial as a reference.

On trials with advice, an advisor's avatar would appear in the middle between the two bars with an arrow pointing to the bar that the advisor endorsed. On some versions of the paradigm, the advisor would provide an estimate with a measure of confidence. In these cases, the advisor's avatar appeared close to the bar they endorsed, and then slid up or down the bar to indicate how confident they were in the advice.

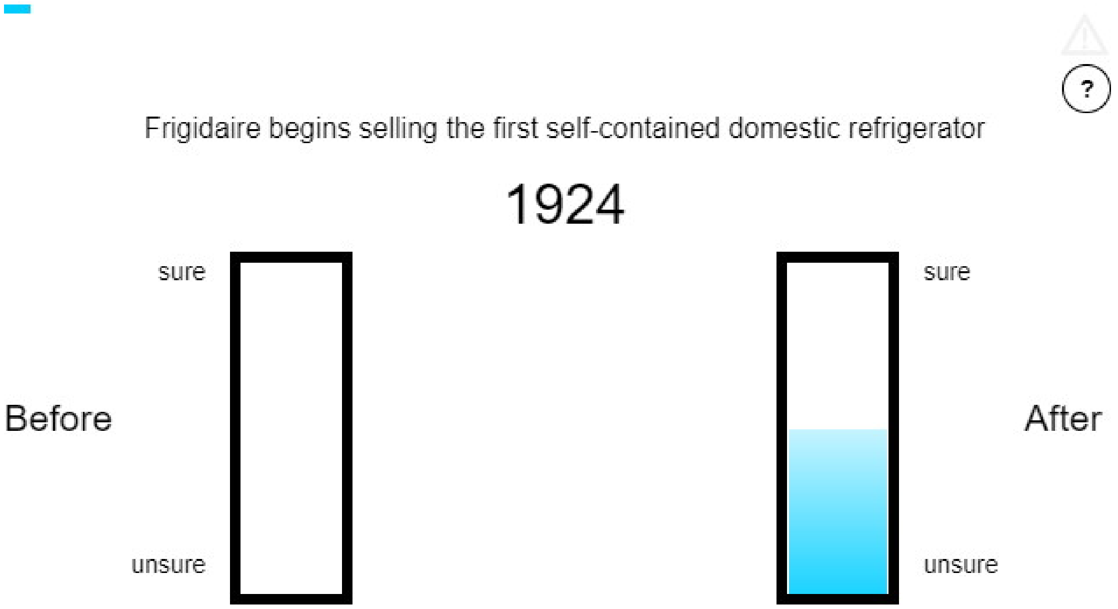
Once the advisor's avatar had indicated the advised response, participants again selected a height within a bar to enter their final decision. On trials with feedback, a gold star would appear next to the correct bar, along with the actual year of the event. The feedback lasted 2s.

As in the Dots task§2.1.3.1, advice was conceptualised along the binary dimensions of accuracy (whether the indicated bar was the correct one) and agreement (whether the indicated bar was the one selected in the participant's initial estimate). In some experiments advisors gave estimates with an indication of confidence. The confidence for these ratings was calibrated with regard to objective accuracy and varied with the difficulty of the question (i.e., with the discrepancy between the anchor year and correct year of the event). Specifically, the advice was generated by drawing a random year from a normal distribution centred around the correct year. The advice was given according to whether this year fell before or after the anchor year, with confidence scaled as a function of the distance between those years. This method is a somewhat rationalist version of how participants themselves may generate confidence judgements in this task.

When a choice of advisors was offered, advisor avatars were placed vertically in the centre of the screen, between the two bars, and participants clicked on the avatar of the advisor from whom they wished to get advice. There was never a time limit for this choice.

Attention check trials in the binary Dates task prompted the participant to enter a specific response with a specific confidence. For example, the prompt that usually contained an event might instruct the participant to "enter 'Before' with high confidence". As with the continuous version, entering an incorrect response would result in immediate termination of the experiment.

Selection of events for the Dates task The events selected for inclusion were determined through an iterative process of trial and error. Initially, a selection of events between 1850 and 1950 were compiled primarily from Wikipedia's timelines of the 1800s (*Timeline of the 19th Century* 2021) and 1900s (*Timeline of the 20th Century* 2021). Events were chosen that were felt by me to be somewhat guessable, and to have occurred on a specific year. A small website was created to allow people to enter a range of years within which they believed an event occurred, and these events were piloted by recruiting participants from Prolific.



Frigidaire begins selling the first self-contained domestic refrigerator

1924

Figure 2.5: Dates task with binary responses.

During piloting, each respondent saw each event in turn, and entered the year they believed the event to have occurred, along with years they were 90% sure the event occurred *after* and 90% sure the event occurred *before*. The data from the piloting indicated that people generally performed extremely poorly on the task, with a few exceptions for particularly famous dates. Performance was especially dire for events from the nineteenth century.

A new date range of 1900-2000 was chosen, and further events were added by revisiting the Wikipedia timeline for the 1900s (*Timeline of the 20th Century* 2021), as well as the Oxford Reference timeline of the 1900s (*20th Century* 2012). The new list of questions performed better during piloting. Questions where respondents' answers were particularly accurate or inaccurate were removed, leaving a final list of around 80 events.

Verification of the Dates task We ran a study with the Dates task that aimed to replicate effects previously observed in the Dots task, as a check on likely data quality and for the comparability of results across methods (Appendix B). The results indicated that the results were similar to those found by Pescetelli and Yeung (2021) for the feedback condition, in which participants were more influenced by advice from accurate compared to agreeing advisors, but not for the no feedback condition, in which participants were equally influenced by advice from both advisors.

Specific limitations of the Dates task There are several specific limitations of the Dates task. Firstly, and most importantly, we were unable to control participants' accuracy in the way we were able to control their accuracy in the Dots task. The advisor advice profiles' advice was determined through varying the probability of agreement contingent on the accuracy of the participant's initial estimate. Being unable to fix the overall accuracy of participants' initial estimates meant that the overall accuracy and agreement rates of the advisor advice profiles could differ quite substantially across participants and also between advisors for the same participant.

The inability to control overall accuracy rates is in part a consequence of the second limitation: the difficulty of any given question was idiosyncratic. The questions concerned dates of real historical events, and some of these may have had particular relevance to individual participants, or were perhaps encountered by those participants recently or memorably, allowing precise, accurate, and high-confidence responses to what would be for the average participant very difficult questions. The average performance of all participants on any given question is not necessarily a good guide to the individual experience of any given participant.

Lastly, participants found the task difficult overall. People tend to be more willing to take advice concerning difficult tasks (Gino and Moore 2007), perhaps as a way of diluting responsibility for the decision (Harvey and Fischer 1997). This meant that there was potential for some ceiling effects whereby advisors were not differentiated because all advisors were seen as useful, even if this was because of their potential for sharing blame rather than the informational content of their advice.

2.1.3.3 General limitations

There are several limitations common to both task designs. The most obvious limitation is that the advisors are not organically-interacting humans. There are other ecological validity limitations in the presentation of advice, the structure of the experiment, the absence of other cues, and the types of tasks used.

The use of artificial advisors means that advice can be carefully specified, and the experiments can be run easily, cheaply, and quickly. Transparently artificial advisors may, however, limit the generalisability of the experimental results in two ways. Firstly, if different integration processes exist for social and non-social information, it is plausible that, for at least some participants, the advice information is perceived as non-social information. While social and non-social information processing would not invalidate any findings (because such a factor would be unlikely to be systematically related to manipulations of interest), they may harm the ability of experimental results to inform us about the processes by which social information is integrated. Secondly, artificial advisors may not trigger a number of human-centred processes such as equality bias (Mahmoodi et al. 2015), meaning that effects revealed in these experiments may be much more difficult to observe in real human advice exchanges.

The advice presented to participants in the experiments is specific and impersonal. During real-life advice-taking, advice is often provided within a discussion, with estimates accompanied by reasons and points responded to interactively. Although studies have indicated that advice-taking behaviour remains similar where discussion is allowed (Liberman et al. 2012; Minson, Liberman, and Ross 2011), these experiments placed discussion within the context of advice exchanges over a decision made by both dyad members individually, and not with distinct roles for the advisor and judge as used in these experiments. The relationship between the judge and the advisor is also less rich than real-life relationships where numerous other factors may alter or overwhelm any advice-taking and advisor evaluation processes revealed by these experiments.

Further ecological validity limitations arise from the structure of the experiment. The task presents a series of trials sequentially, with a rapid procession through each. This structure is intended to condense a real-world relationship with an advisor, built up over repeated interactions over time, into as narrow a time window as possible. It is possible that this temporal compression does not fundamentally alter the processes of advisor evaluation and trust formation, but we have little positive evidence to support this supposition.

Lastly, both the Dots task and the Dates task had an objectively correct answer on every question. A substantial portion of everyday real-world advice-seeking behaviour concerns questions on which there is no readily determinable objectively correct answer, such as where might be a good holiday destination, or whether one should pursue postgraduate education. These more subjective questions have received some attention by van Swol (2011), but are seldom studied in the advice-taking cognitive psychology literature.

2.2 Analysis

The results of statistical analyses are included within the text. Where a number is expressed in the form $x[y, z]$, y and z are the lower and upper 95% confidence limits for x, respectively.

2.2.1 Dependent variables

2.2.1.1 Pick rate

Pick rate provides a measure of advisor choice behaviour. In most experiments there are some trials that offer participants a choice of which advisor they would like to hear from. There are always two choices, and a choice must always be made. The two choices are consistent within the experiment. Pick rate is the proportion of choice trials in which a specified advisor was chosen.

A participant's pick rate is an aggregate over a number of trials, and expresses the observed probability of picking the specified advisor. The mentally represented preference for that advisor is not measured directly (if such a thing even exists), and cannot be determined from the observed pick rate without knowing the mapping function for each individual participant. Mapping functions (such as the logistic sigmoid function) produce stochastic choice behaviour from a preference marked on a continuous scale. The relationship between preference and pick rate is non-linear and idiosyncratic, but it is likely monotonic for all participants: the stronger the preference the higher the pick rate.

2.2.1.2 Weight on Advice

Weight on Advice, and its complement Advice-taking, are commonly used to quantify the relative contributions of advice and initial estimates in making final decisions. It is obtained by dividing the amount an initial estimate was updated by the amount the advisor recommended adjusting the initial estimate. It thus expresses the amount the estimate changed as a proportion of the advised change.

Formally, Weight on Advice is given by $(e^F - e^I)/(a - e^I)$ where e^I and e^F are the initial estimate and final decision, respectively, and a is the advice.

Where the final answer moves away from advice (i.e. the final decision is further from the advice than the initial estimate), the value of Weight on Advice is negative, and where the adjustment towards advice exceeds the advice itself (i.e. the advice falls between the final decision and the initial estimate) the value of Weight on Advice is greater than 1. These values are typically truncated to 0 and 1, respectively.

In cases where the advice is exactly equal to the initial estimate, the denominator is equal to zero and the value for Weight on Advice is thus undefined. Trials in which the advice is exactly equal to the initial estimate are consequently discarded when calculating Weight on Advice.

2.2.1.3 Influence

Participants on the binary tasks (Dots and Binary Dates task) make two responses with a direction and confidence. The amount that the confidence shifts between the initial estimate and final decision in the direction of the advice is termed the *influence* of the advice. In most cases, participants increase their confidence when advisors offer agreeing advice and decrease their confidence when advisors offer disagreeing advice. Both of these cases constitute positive influence of the advice.

Influence is calculated as the extent to which the judge's initial estimate is revised in the direction of the advisor's advice. The initial (C_1) and final (C_2)

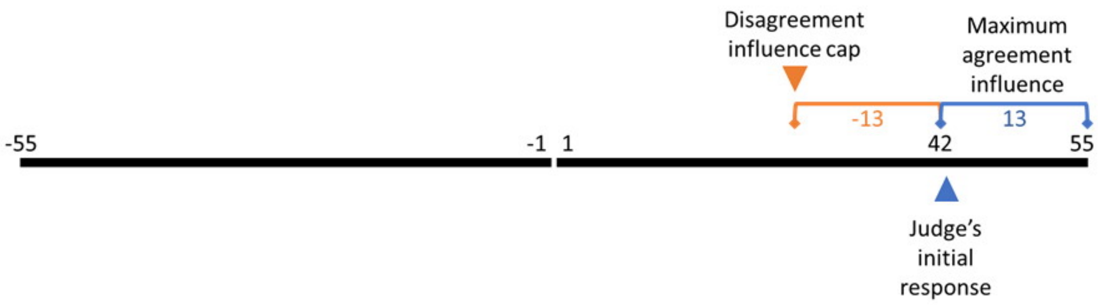


Figure 2.6: Capping influence to avoid scale bias.
In this example the judge's initial response is 42, meaning that their final decision could be up to 13 points more confident or up to 97 points less confident. Any final decision which is more than 13 points less confident is therefore capped at 13 points less confident.

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:

$$I|\text{agree} = f(C_1) \begin{cases} C_2 - C_1 & C_1 > 0 \\ -C_2 + C_1 & C_1 < 0 \end{cases} \tag{2.1}$$

And by the inverse of this for disagreement trials:

$$I|\text{disagree} = -I|\text{agree} \tag{2.2}$$

Capped influence 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 that were greater than the maximum influence that could have occurred had the final decision been a maximal response in the direction of the initial answer (Figure 2.6).

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

$$I_{\text{capped}} = \min(I, |S_{\max} - C_1|) \tag{2.3}$$

Where C_1 and C_2 are the initial estimate and final decision, respectively, I stands for influence, and $S_{\max}$ indicates the maximum value of the scale. Thus, to follow the example in Figure 2.6, the initial estimate was 42, and if disagreeing advice led to a final decision of 2 the raw influence measure would be 40. This would then be compared to the absolute value of the scale width minus the initial estimate (55 - 42 = 13) and the minimum of the two would be the capped influence value, in this case 13.

There are benefits to and limitations of this approach. A benefit is that it places agreement and disagreement on the same potential scale. A limitation is that the truncation happens most extremely where initial estimates are made with very high confidence (because there is almost no increase that could have happened following agreement, so the cap is very low on the influence of disagreeing advice). Conceptually, these changes where a participant adjusts a view previously held with near certainty may in fact be the most interesting of all.

An alternative approach to capping, not used here, is to scale the observed adjustment by the potential space for adjustment. This approach often has the effect of over-inflating influence of agreement, and can make very subtle differences in the use of the top end of the scale highly important. If, for example, a participant using a 50-point rating scale selected 48 for their initial estimate and received agreeing advice, a final decision of 49 would constitute an influence rating of 50% and a final decision of 50 an influence rating of 100%. It is not at all clear that participants differentiate the positions on the scales with this decree of precision, and so we considered this approach to capping influence inappropriate.

2.2.2 Inferential statistics

Statistical analyses are conducted using both frequentest and Bayesian statistics.

2.2.2.1 Frequentist statistics

In this thesis a range of frequentist statistical tests are used, most frequently t-tests and analyses of variance (ANOVA). The α is always set at .05 unless otherwise stated. Null hypotheses are always the expected distribution if the effect being tested is nil. Where the level of influence exerted by two different advisors is studied, for example, the null hypothesis would be that there were no systematic differences in influence exerted by those advisors.

2.2.2.2 Bayesian statistics

The 'Bayes factor' quantifies relative likelihood between statistical models given observed data and prior beliefs. This is usually presented as a comparison between a model representing the alternative hypothesis and a model representing the null hypothesis, written as $\text{BF}_{\text{H1:H0}}$. When using Bayes factors to explore the results of linear modelling, tests are conducted on whether the data are better fit by a model with an effect as compared to a model without that effect, written as $\text{BF}_{\text{+Effect:-Effect}}$.

In order to draw categorical inferences, thresholds are placed on this continuous outcome. Here these thresholds are $1/3 < \text{BF} < 3$, meaning that a BF of less than 1/3 constitutes evidence in favour of the simpler model while a BF greater than 3 constitutes evidence in favour of the more complex model. These values are those suggested by Schönbrodt and Wagenmakers (2018) as representing moderate evidence in a given direction. Where the BF lies between these thresholds it is labelled as 'uninformative'. An uninformative result supports neither the simpler nor the more complex model, and indicates that the data are insufficient to distinguish the hypotheses.

The Bayesian tests used here rely on the priors specified by the BayesFactor R package (Morey and Rouder 2015). These priors govern the expected distributions of observed differences between samples where there is or is not a genuine effect creating systematic differences. The use of the same, weakly-informative priors for all tests means the approach used here is an 'objective Bayesian' approach. This objective Bayesian approach can be contrasted with a 'subjective Bayesian' approach in which the goal is to specify the exact amount of belief one should have in one hypothesis over another. Neither the objective nor subjective approach is clearly superior. The objective approach is used here because it is simpler. There is some risk that results will be a poor fit to reality because the priors are inappropriate, but this risk is fairly low and somewhat mitigated by the additional inclusion of frequentist statistics.

2.2.2.3 Integrating statistical results

In most cases, Bayesian and frequentist statistics produce the same conclusion. Where this is not the case, results should be interpreted very cautiously: a significant frequentist test with an uninformative or null-favouring Bayesian test can indicate that the result may be a false-positive, while clear Bayesian support for the alternate hypothesis in the absence of a significant frequentist test can indicate that the priors in the Bayesian test are inappropriate.

Where null conclusions are to be drawn, i.e. the null hypothesis is to be retained, only Bayesian statistics can be considered informative. In these cases Bayesian statistics will be interpreted, with the caveat that the safeguard of using two independent approaches to draw statistical conclusions has lapsed.

2.2.2.4 Software

Data analysis was performed using R (R Core Team 2018), and relied extensively on the Tidyverse family of packages (Wickham 2021). For a full list of packages and software environment information, see Appendix C.

2.2.3 Unanalysed data

Early versions of several experiments had bugs in the experiment code that made the results unreliable. The data collected during these runs are available alongside the data collected in the final versions of experiments. 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 they submitted on the Prolific participant recruitment platform. For details on what went wrong with the experiments in which bugs were found, see the description column of the data files in the esmData:: R package. While not useful for the hypotheses of the experiments for which the data were collected, some excluded data can be used for other purposes such as analysing responses to advice. A breakdown of unanalysed data by experiment is provided as Appendix A.

2.3 Open science approach

2.3.1 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 fraudulent 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 (Simmons, Nelson, and Simonsohn 2011), Nosek and colleagues published a roadmap for improving the structure and function of academic research and publishing (Nosek and Bar-Anan 2012; Nosek, Spies, and Motyl 2012), and the Open Science Collaboration began (Collaboration 2015). 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.

2.3.2 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 Center and used according to the Center's rules. The preregistration badges are used throughout the thesis to indicate analyses that were specifically declared in preregistrations.

2.3.2.1 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 (Kerr 1998). 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.

2.3.2.2 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. In some cases links are provided without badges; this badge indicates that the shared materials are considered suitable for release.

2.3.2.3 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 re-purpose the data for other investigations. This increases the robustness of the results, and increases the efficiency of science as a whole. All data are available on-line, with links provided throughout the thesis. This badge is unused because there are no cases where non-open data need to be differentiated from open data.

2.3.3 Thesis workflow

This thesis is written in RMarkdown using the Oxforddown template (Lyngs 2019), with the data fetched and analysed at the time the document is produced using the publicly available pipeline - the entire document can be reproduced locally using the source code in an appropriate environment. Some parts of the thesis use computational modelling or model fitting. These parts use cached datasets, also available on-line, in order to reduce the time required to create the thesis from its source files from hours to minutes using a high-end desktop personal computer. The caching behaviour can be prevented for those wishing to evaluate the validity of the modelling components by setting the value of the R option `ESM.recalculate` to higher values (documented in the `index.Rmd` file). A Docker environment copying the environment used to produce this document can be produced by running the `Dockerfile` included in the repository.

3
Psychology of advisor choice

Advice is relied upon to a different extent depending upon a variety of markers for its trustworthiness, including its plausibility and the reputation of its source. The question at the core of this chapter is whether people preferentially seek out advice from advisors they believe to be more trustworthy. At first glance, it may seem a foregone conclusion that people will seek out advice from more trustworthy advisors: who after all wants to be advised by fools or liars? Nevertheless, empirical evidence for this kind of source selection behaviour has been somewhat mixed. The evidence that people do tend to seek out information from sources likely to agree with them is moderate ('selective exposure', Hart et al. 2009). The evidence that people avoid information likely to disagree is poor ('selective avoidance', Jang 2014; Weeks, Ksiazek, and Holbert 2016), with evidence becoming less persuasive as tasks become more ecologically valid (Sears and Freedman 1967; Nelson and Webster 2017).

There are also intuitive arguments for a range of potential findings. It would make sense that people seek out information they are more likely to use, because all information acquisition comes with some kind of cost, even if only attentional and opportunity costs, and rational actors should maximise their benefit-cost trade-off. It may make sense for people to seek out information they are likely to agree with, regardless of usefulness, because they may be exercising critical vigilance over their own side in a debate, ensuring that bad arguments are not used to support their ideological position. It would also make sense, however, for people to seek out information from sources they disagree with: perhaps those we disagree with have access to evidence or reasons we had not considered; or perhaps learning about others' views will allow us to better counter them and convert their adherents (Freedman 1965). People may even prefer a balanced or random diet of information because they feel unable to judge relative quality, or because all the reasons above are pulling them in different directions.

The vast majority of the source selection literature uses surveys or browsing tasks with stimuli being realistic politically-charged media items. Measurements are either active interest ratings or passive activity monitoring (usually on links clicked or reading time, but eye tracking is a recent innovation: Marquart 2016; Schmuck et al. 2020). The experiments here are more traditional cognitive psychology experiments: the complex contextual factors suspected of driving selectivity are removed (Festinger 1957; Knobloch-Westerwick 2015) and only the informational motive remains. While it is not impossible that a preference for agreement which makes sense from a self-image preservation perspective bleeds into a context where accuracy is key, the experiments here at least provide a context where a correct answer exists.[1]

3.0.1 Similar work

The source selection literature is largely from the domains of Social and Personality Psychology, in which the constructs that produce the phenomena are attitudes and self-concepts. The present work is grounded in the Cognitive Psychology domain, and consequently uses a model of advisor evaluation (Advisor Evaluation without Feedback§1.3.4) that posits measurable variables and mathematically describable processes. This model is based on a similar model from Pescetelli and Yeung (2021). It takes as a starting point the observation that, given objective feedback, people can use that feedback to learn about the trustworthiness of advisors (Yaniv and

[1] It is arguable that people may seek out politically-concordant information because they believe it to be more accurate rather than because they do not want their self-image to be challenged.

Kleinberger 2000; Pescetelli and Yeung 2021; Behrens et al. 2008). The extent to which advice is taken (§2.2.1.2) 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. As expected, people make greater use of advice they believe will be more accurate compared to less accurate (Gino, Brooks, and Schweitzer 2012; Rakoczy et al. 2015; Sniezek, Schrah, and Dalal 2004; Soll and Larrick 2009; Tost, Gino, and Larrick 2012; Schultze, Mojzisch, and Schulz-Hardt 2017; Wang and Du 2018; Önkal, Gönül, et al. 2017).

When objective feedback is unavailable, it is still possible for people to demonstrate a greater dependence upon advice from more as opposed to less accurate advisors. 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.

Pescetelli and Yeung (2021) demonstrated that advice is more influential from (equally accurate) advisors who tend to agree with a participant more frequently when objective feedback is not provided. This is despite the fact that advice is generally more influential when it disagrees with the participant's initial estimate.[2] Their data suggest that people may be using agreement as a proxy for accuracy, although they may also simply prefer agreement over disagreement when there is no accuracy cost to be paid. In this chapter, we partially replicate the findings of Pescetelli and Yeung (2021), and explore the consequences of pitting accurate advisors against agreeing advisors.

[2] 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.

3.0.2 Overview of experiments

We conducted a series of experiments to explore whether the advice-taking behaviour observed previously (Pescetelli and Yeung 2021) would translate into preferential advisor choice behaviour. In these experiments participants were familiarised with different pairs of advisors, and then given the opportunity to select which advisor they would like to get advice from (General Method§2.1).

Each experiment was repeated using two different tasks: a perceptual decision-making "Dots task", extended from Pescetelli and Yeung (2021); and a historical date estimation "Dates task" newly built for this project. To reduce expenditure, all participants in the Dots task experiments received no feedback on their answers while learning about advisors, because contrasting the presence and absence of feedback was done by Pescetelli and Yeung (2021). In the Dates task, due to its novelty, participants were split into conditions based on whether or not they received feedback while learning about the advisors.

The first of the advisor pairs was a high accuracy advisor and a low accuracy advisor (Experiment 1A§3.1.1 and Experiment 1B§3.1.2). We predicted that the high accuracy advisor would be selected more often, even where feedback was not available. This experiment would demonstrate the minimum phenomenon of interest – sensitivity to advisor accuracy in the absence of feedback translating into preferential advisor choice.

The second advisor pair was a high agreement and a low agreement advisor (Experiment 2A§3.2.1 and Experiment 2B§3.2.2). We predicted that the high agreement advisor would be selected more frequently, because we expect agreement to be the method by which the accurate advisors were detected in the previous task. This experiment would constitute a test of the purported mechanism.

The third advisor pair was a high agreement advisor and a high accuracy advisor (Experiment 3A§3.3.1 and Experiment 3B§3.3.2). We predicted that the high agreement advisor would be selected more frequently than the high accuracy advisor, but only where feedback was withheld. Where feedback was provided, we expected the high accuracy advisor to be picked more often. This experiment would test whether the absence of feedback invites using agreement as a substitute for accuracy.

The final advisor pair were confident contingent advisors like those used in Pescetelli and Yeung (2021) (Experiment 4A§3.4.2 and the Lab Study§3.4.1). These advisors agree at the same rate and are similarly accurate, but one agrees more when the participant expresses high initial confidence and less when the participant expresses low initial confidence, and the other vice-versa. We predicted that the former 'bias-sharing' advisor would be selected more often because participants would use their own sense of confidence to weight the value of agreement. This experiment would explore whether metacognitive processes are able to finesse the basic agreement-for-accuracy substitution.

3.1 Effects of advice accuracy on advisor choice

Pescetelli and Yeung (2021) demonstrated that more accurate advisors are more trusted and 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 choice in two on-line tasks: a 'Dots task' requiring similar perceptual decision-making, and an estimation-based 'Dates task'. We predicted that participants would choose a more accurate advisor over a less accurate one, and would do so even in the absence of objective feedback (based on the hypothesis that they can infer accuracy from differing agreement rates).

The ability to distinguish between accurate advisors in these experiments is important because discrimination based on accuracy is crucial to the phenomenon we are attempting to explain: rational advice-seeking behaviour in the absence of feedback. Pescetelli and Yeung (2021) demonstrated that people could identify and exploit more accurate advice, and these experiments seek to determine whether people will use that ability to obtain advice from a more reliable source. Experiment 1A§3.1.1 addressed this issue using the same perceptual decision task as use by

Pescetelli and Yeung (2021); Experiment 1B§3.1.2 extended the approach in a task asking participants to estimate historical dates.

3.1.1 Experiment 1A: advice accuracy effects in the Dots task

3.1.1.1 Open scholarship practices

This experiment was preregistered at `https://osf.io/u5hgj`. The experiment data are available in the `esmData` package for R (Jaquiery 2021c), and also directly from `https://osf.io/kn23p/`. A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from `https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html`.

3.1.1.2 Method

59 participants each completed 368 trials over 7 blocks of a perceptual decision-making task. Each trial consisted of three phases: participants gave an initial estimate (with confidence) of which of two briefly presented boxes contained more dots; received advice on their decision from an advisor; and made a final decision (again, with confidence).

Participants started with 2 blocks of 60 trials that contained no advice to allow them to familiarise themselves with the task; to allow the staircasing process to titrate the difficulty to their ability in order to maintain approximately 71% initial estimate accuracy; and to allow estimating of participants' idiosyncratic confidence reporting style. The first 3 trials were introductory trials that explained the task. All trials in this section included feedback indicating whether or not the participant's response was correct.

Participants then did 5 trials with a practice advisor to get used to receiving advice. They were informed that they would "get **advice** from an advisor to help you make your decision [original emphasis]", and that "advice is not always correct, but it is there to help you: if you use the advice you will perform better on the task."

Table 3.1: Advisor advice profiles for Dots task with in/accurate advisors

	Probability of agreement			
Advisor	Participant correct	Participant incorrect	Overall	Overall accuracy
High accuracy	.800	.200	.626	.800
Low accuracy	.600	.400	.542	.600

Table 3.2: Participant exclusions for Dots task with in/accurate advisors

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	3
Skewed confidence categories	6
Too many participants	0
Total excluded	**9**
Total remaining	**50**

Participants then performed 2 sets of 2 blocks each. These sets consisted of 1 Familiarisation block of 60 trials in which participants were assigned one of two advisors. The Familiarisation block was followed with a Test block of 60 trials in which participants could choose between the advisors they encountered in the Familiarisation block. The participants saw different pairs of advisors in each set, with each pair consisting of one advisor with each of the advice profiles.

Advice profiles The two advisor profiles (Table 3.1) used in the experiment were High accuracy and Low accuracy. The advisors' advice was stochastically generated according to the participant's response. The High accuracy advisor predominantly agreed with correct participant responses and disagreed with incorrect ones. The Low accuracy advisor did likewise, but was less likely to agree with correct responses and more likely to agree with incorrect ones. Overall, given an expected participant accuracy of 71% obtained by the staircasing procedure, the High accuracy advisor was correct 80% of the time while the Low accuracy advisor was correct 60% of the time. The advisor profiles were not balanced for overall agreement rates.

3.1.1.3 Results

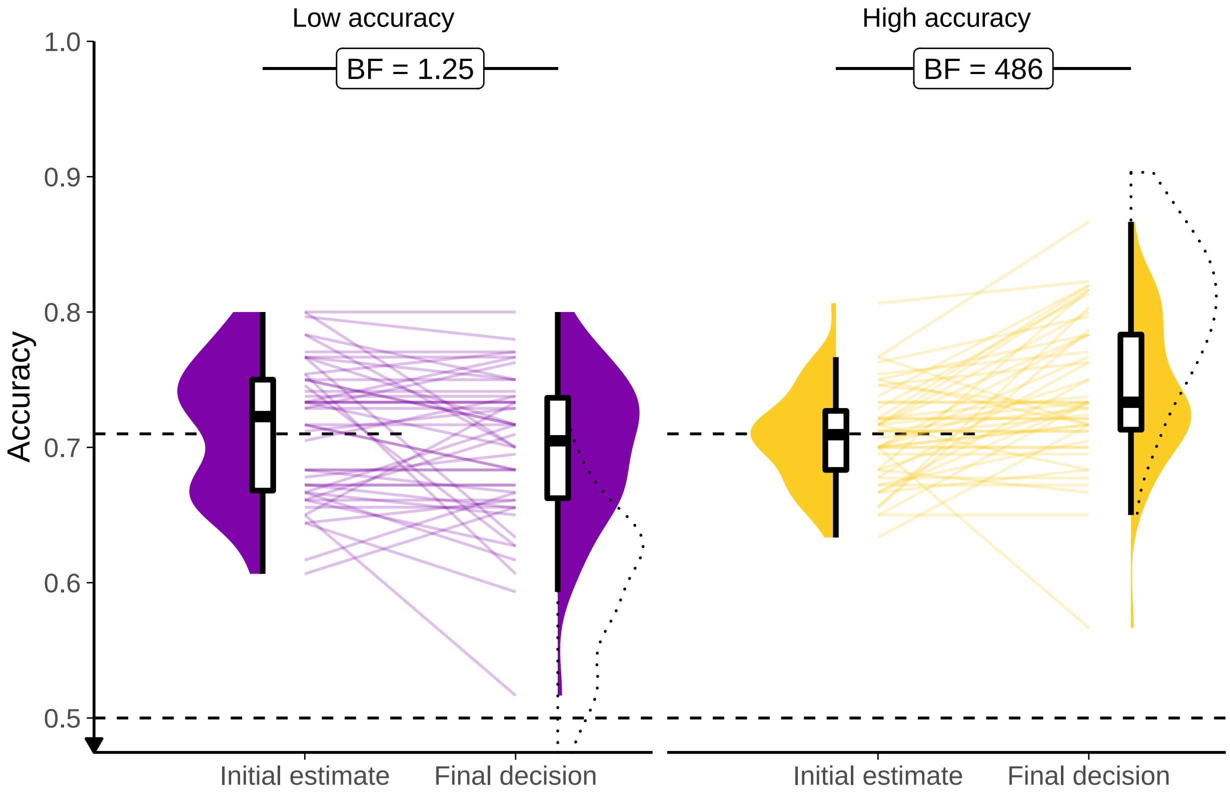


Figure 3.1: Response accuracy for the Dots task with in/accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

Exclusions In line with the preregistration, participants' data were excluded from analysis where they had an average accuracy below 0.6 or above 0.85, did not have choice trials in all confidence categories (bottom 30%, middle 40%, and top 30% of prior confidence responses), had fewer than 12 trials in each confidence category, or had completed the experiment after the preregistered amount of data had already been collected. Overall, 9 participants were excluded, with the details shown in Table 3.2.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor choice behaviour, it is useful to verify that participants interacted with the task in a sensible way, and

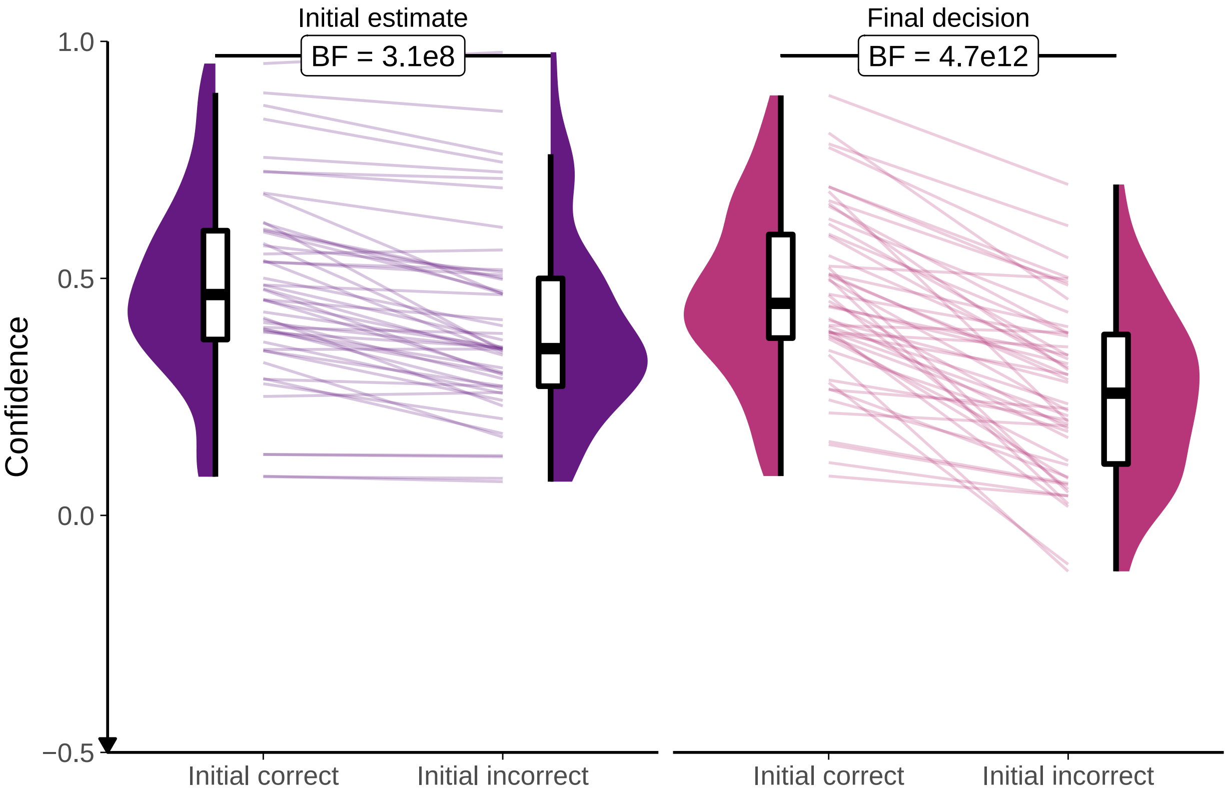


Figure 3.2: Confidence for the Dots task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

that the task manipulations worked as expected. In this section, task performance is explored during the Familiarisation phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Accuracy Accuracy of initial estimates was controlled by a staircasing procedure which aimed to pin accuracy to 71%. The accuracy of final decisions was free to vary according to the ability of the participant to take advantage of the advice on offer. As Figure 3.1 shows, participants' accuracy scores for initial estimates were close to the target values (partly because participants whose accuracy scores diverged considerably were excluded). Participants tended to improve the accuracy of their responses following advice from High accuracy advisors, while the evidence was unclear as to whether there was any difference in response accuracy with Low accuracy advice. This is supported statistically by an ANOVA of response accuracy by Advisor and Time: there was no discernible effect of Advisor (F(1,49) = 3.86, p = .055; $\mathrm{M}_{\mathrm{LowAccuracy}}$ = 0.71 [0.69, 0.72], $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 0.72 [0.71, 0.73]),[3] but there was an interaction between Advisor and Time (F(1,49) = 17.32, p < .001; $\mathrm{M}_{\mathrm{Improvement|LowAccuracy}}$ = -0.01 [-0.03, 0.00], $\mathrm{M}_{\mathrm{Improvement|HighAccuracy}}$ = 0.03 [0.02, 0.05]); and there was an effect of Time F(1,49) = 4.80, p = .033; $\mathrm{M}_{\mathrm{Final}}$ = 0.72 [0.71, 0.73], $\mathrm{M}_{\mathrm{Initial}}$ = 0.71 [0.70, 0.72].

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions (Figure 3.2; Figure 3.3). There is considerable variation between participants both on their baseline confidence and on its variability (not shown), despite all participants being roughly matched for accuracy. The narrow accuracy range, and its random nature mean there was no evidence of a correlation between participant accuracy and confidence for initial ($\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/2.80) or final ($\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/2.08), although neither of these correlations indicated good evidence that no correlation existed. Variation between individuals' confidence reports is expected (Ais et al. 2016; Navajas et al. 2017).

We ran an ANOVA on confidence by Time (initial estimates versus final decisions) and Correctness of the initial estimate (Correct initial estimates versus Incorrect initial estimates). For this analysis confidence was directionally coded so that final decision confidence was negative if the answer side changed between the initial estimate and final decision. The analysis indicated that participants were more confident in their initial estimates than their final decisions (F(1,49) = 15.82, p < .001; $\mathrm{M}_{\mathrm{Final}}$ = 17.99 [15.55, 20.43], $\mathrm{M}_{\mathrm{Initial}}$ = 22.06 [19.27, 24.84]). This makes sense given that there is more scope for participants to reverse their confidence than to increase it on any given trial given the way the scale works. There was also a

[3] Throughout, where means or parameter estimates are reported with bracketed figures afterwards, those figures give 95% confidence intervals.

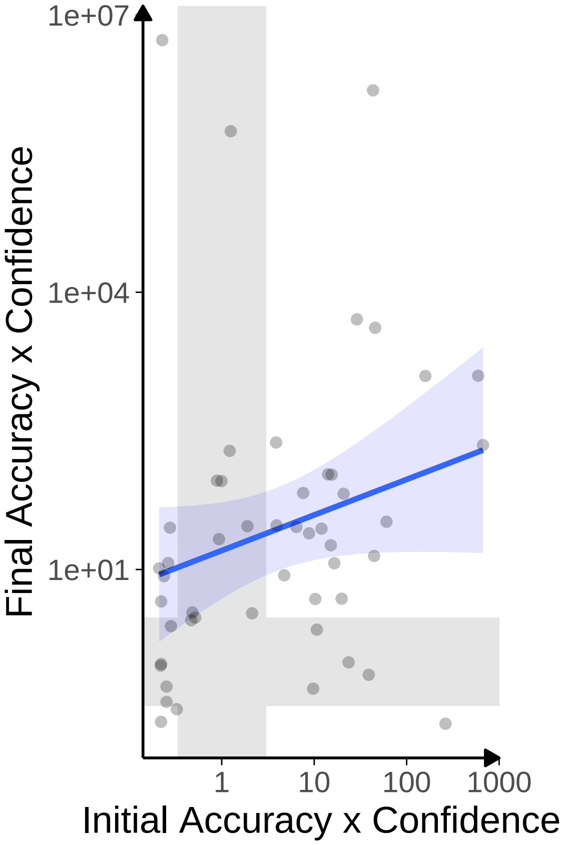


Figure 3.3: Accuracy x Confidence correlations for Dots task with in/accurate advisors. Each point marks the Bayes factor for a participant's correlation between their initial accuracy and confidence (horizontal axis) and final accuracy and confidence (vertical axis). Shaded bands show areas of no information ($1/3 < \text{BF} < 3$), with evidence for a correlation rightwards and upwards of the area and evidence against below and leftwards. Note that the Bayes factor is a measure of the likelihood the correlation is not 0, not a direct measure of the strength of the correlation. The blue line indicates the overall pattern, with shaded area giving the 95% confidence intervals. Axes use $\log_{10}$ scale.

main effect of Correctness, with participants being more confident overall where their initial estimate was correct as compared to when it was incorrect (F(1,49) = 168.89, $p < .001$; $\text{M}_{\text{Correct}}$ = 23.64 [21.05, 26.24], $\text{M}_{\text{Incorrect}}$ = 16.40 [14.07, 18.74]). There was an interaction, with participants becoming even less confident between initial estimates and final decisions for trials where the initial estimate was incorrect (F(1,49) = 46.01, $p < .001$; $\text{M}_{\text{Increase|Correct}}$ = -0.97 [-2.62, 0.69], $\text{M}_{\text{Increase|Incorrect}}$ = -7.17 [-9.89, -4.45]). This indicates that confidence is behaving in a sensible manner.

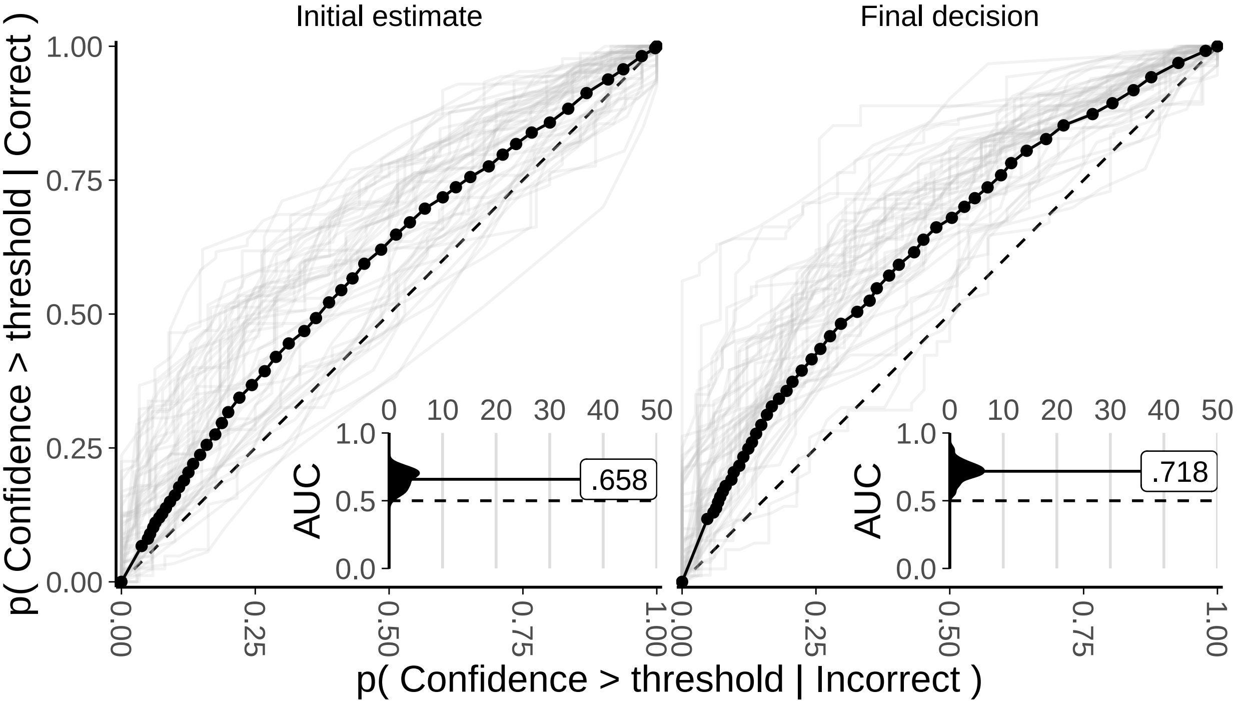


Figure 3.4: Metacognitive performance for the Dots task with in/accurate advisors. Faint lines show Receiver Operator Characteristic (ROC) curves for individual participants, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers. The inset plot shows the distribution of areas under the ROC, and the label gives the mean value.

Metacognitive ability Where performance on the underlying task is held constant, as here at least for participants' initial pre-advice decisions, metacognitive sensitivity can be measured in a bias-free way by plotting Receiver Operating Characteristic (ROC) curves for metacognitive responses (Fleming and Lau 2014).[4] ROC curves are obtained by calculating at each of a number of different points on a confidence scale, the probability that the confidence is at least that high for correct versus incorrect answers. The area under the ROC curve gives a measure of the ability of confidence ratings to distinguish correct and incorrect responses. An area under the ROC curve of .5 indicates chance performance, and a value of 1 indicates perfect discrimination.

As shown by Figure 3.4, almost all participants showed above-chance metacognitive sensitivity for initial estimates and final decisions. Participants generally showed higher metacognitive sensitivity for final decisions, in line with their improved performance on these trials. Participants' metacognitive sensitivity was not particularly high, reflecting the difficulty of the task, and in line with previous datasets with this task (Pescetelli, Hauperich, and Yeung 2021). There was no evidence of participants' metacognitive sensitivity being correlated with their task performance (Initial estimates: r(48) = -.213 [-.306, .250], p = .833; Final decisions: r(48) = .266 [-.243, .313], p = .791). This is expected when task performance is tightly controlled, because under these conditions variation in task performance reflects variation in ability within a participant rather than between participants.

Advisor performance The advice is generated probabilistically from the rules described previously in Table 3.1. It is thus important to get a sense of the actual advice experienced by the participants.

The advisors' performance was stochastic, with the advisors agreeing or disagreeing with set probabilities depending upon whether the participant was correct or incorrect in their initial estimate. The performance of the advisors in practice

[4] The constant underlying task performance is only true for initial estimates in the paradigm used here, and thus the ROC curves for final decisions should be interpreted with caution because they cannot be proven to be unaffected by metacognitive bias.

Table 3.3: Advisor agreement for Dots task with in/accurate advisors

Advisor	Target|correct	Actual|correct	Target|incorrect	Actual|incorrect
High accuracy	.800	.800	.200	.203
Low accuracy	.600	.608	.400	.422

Table 3.4: Advisor accuracy for Dots task with in/accurate advisors

Advisor	Target accuracy	Mean accuracy
High accuracy	.800	.799
Low accuracy	.600	.601

was as specified (Table 3.3). The participants' accuracy rates were controlled with an adaptive staircase, meaning that the advisors' agreement strategies produced overall advice accuracy at target rates. The advisors' actual accuracies matched the target accuracies (Table 3.4), with 49/50 participants experiencing the planned relationship wherein the High accuracy advisor's advice was more accurate than the Low accuracy advisor's advice.

Hypothesis test With basic task performance as expected, our key analysis focused on participants' choice of advisors. As predicted, and as shown in Figure 3.5, participants selected the High accuracy advisor at a rate greater than would be expected if their choosing were random ($t(49) = 3.09$, $p = .003$, $d = 0.44$, $\text{BF}_{\text{H1:H0}} = 9.96$; $M = 0.57$ [0.52, 0.61], $\mu = 0.5$). The modal choice remained at chance level (.5), but almost all participants manifesting a preference preferred the High accuracy advisor.

While this effect is interesting, it is substantially smaller than participants' preference for picking the top advisor regardless of identity ($t(49) = 5.47$, $p < .001$, $d = 0.77$, $\text{BF}_{\text{H1:H0}} = 1.0\text{e}4$; $\text{M}_{\text{P(PickFirst)}} = 0.65$ [0.60, 0.71], $\mu = 0.5$), an effect that we would hope would be random and even out across participants. Note that because the advisor position is well balanced ($\text{BF}_{\text{H1:H0}} = 1/6.29$; $\text{M}_{\text{P(HighAccuracyFirst)}} = 0.50$ [0.48, 0.51], $\mu = 0.5$) across advisors, the presence of a preference for advisor by position would not cause a preference for an individual advisor.

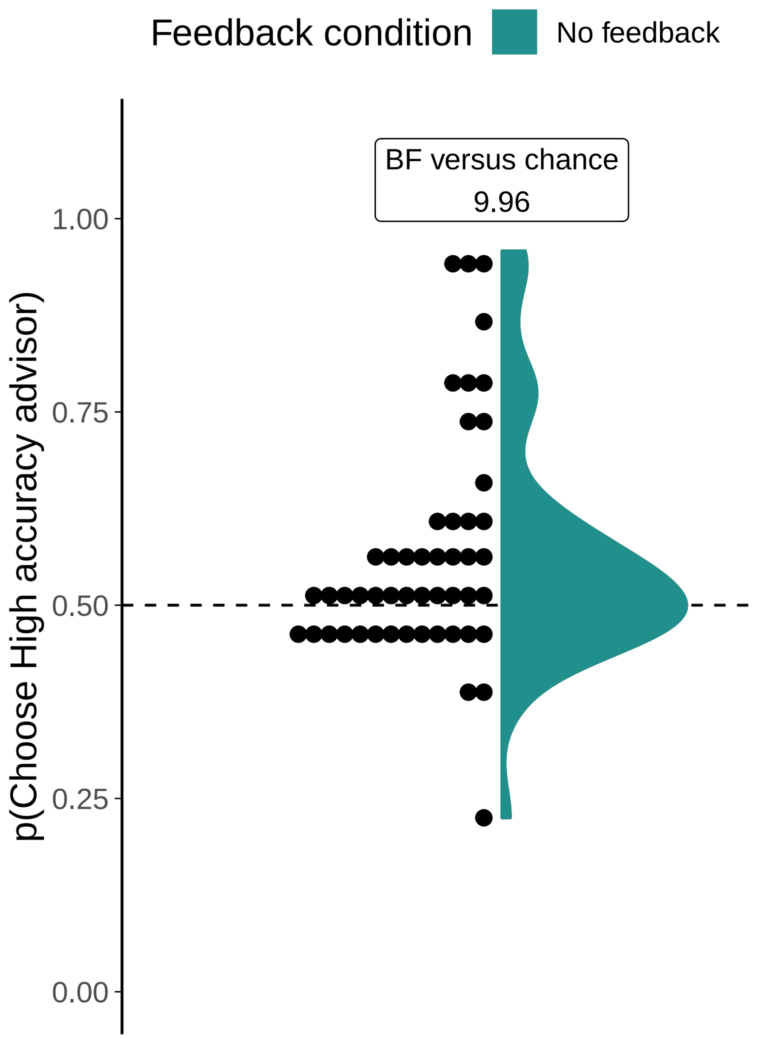


Figure 3.5: Dot task advisor choice for in/accurate advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

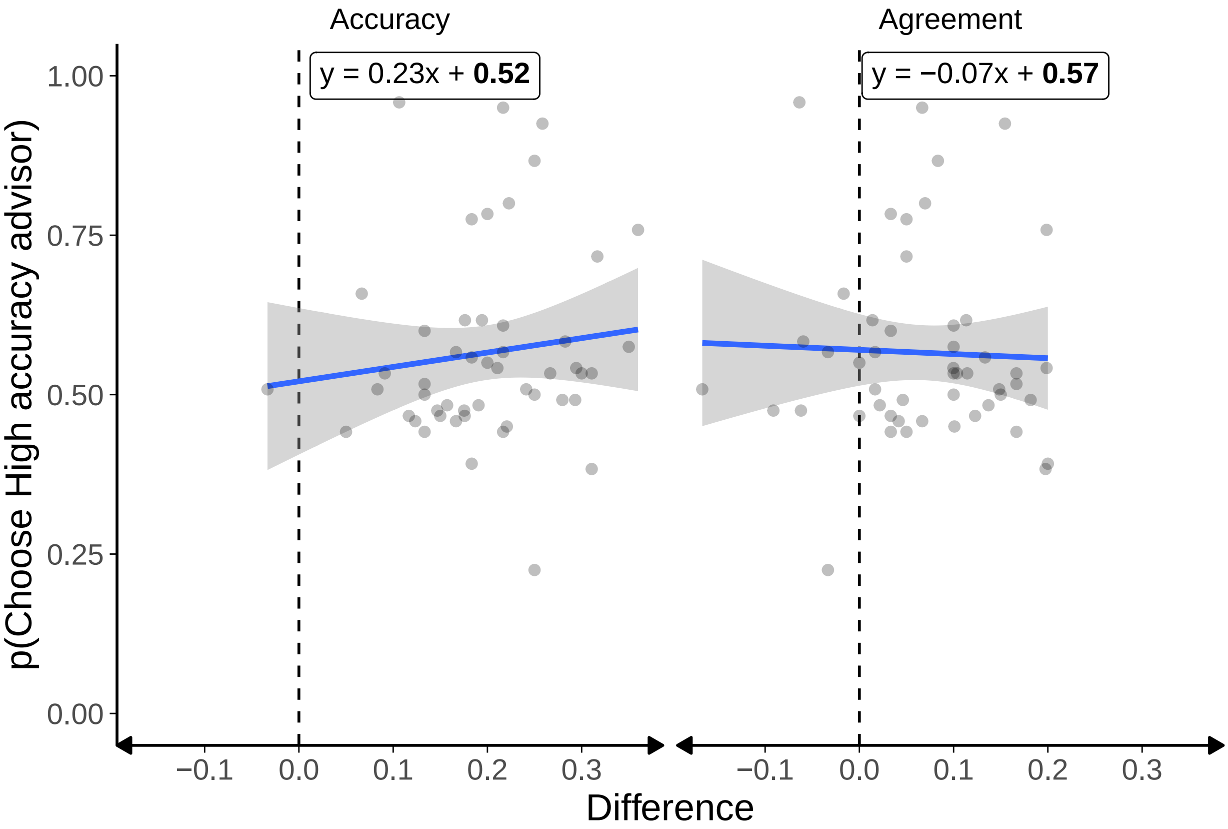


Figure 3.6: Advisor choice by experience in the Dots task with in/accurate advisors. Each dot is a participant's proportions. The difference in accuracy rates is calculated as the proportion of correct answers seen from the High accuracy advisor minus the proportion of correct answers seen from the Low accuracy advisor, and the difference in agreement rates similarly for agreement. The participants did not receive feedback on the correct answers.

Follow-up tests As noted above, the stochastic nature of the advisors' advice meant that there was some variation in the participants' experience of the advisors. Despite this difference, there was no evidence of a relationship between participants' advisor preference and their experience of either advisor accuracy ($\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.28$) or advisor agreement ($\mathrm{BF}_{\mathrm{H1:H0}} = 1/3.05$), with the latter indicating an absence of a relationship (Figure 3.6). This is not entirely surprising because, as with the accuracy correlations discussed above, there was relatively little variation in experience of advisors so effects might be expected to be small and difficult to detect.

We might also expect advisor preference to vary as a function of initial confidence. Perhaps, for example, participants may have a strong preference for the High accuracy advisor but only exercise that preference where they are unsure about the answer themselves (i.e. where advice is most valuable to them). This appeared not to be the case: we split participants' trials into high and low confidence based on their idiosyncratic median confidence, and conducted a paired t-test to compare pick rate of the High accuracy advisor for high versus low confidence trials. The Bayes factor for this t-test indicated good evidence of no difference in pick rates ($\text{BF}_{\text{H1:H0}} = 1/6.49$).

These uninformative results are typical of those across all Dots task experiments in this chapter. To save space, these analyses are not reported for subsequent Dots task experiments.

Discussion In the absence of feedback, it should be possible for a person to evaluate advice using the proxy of whether or not the advice accords with their initial opinion, at least given some reasonable assumptions about the independence of the initial opinion and the advice. In this experiment, we tested whether participants would be able to exploit this heuristic to detect that one advisor was more useful than another, and whether they would choose to hear advice from the more useful advisor. Participants showed a tendency, where they had a preference, to prefer the High accuracy advisor. In the next experiment, we aimed to replicate the results in a different task.

3.1.2 Experiment 1B: Advice accuracy effects in the Dates task

This experiment attempted to replicate the results of the previous experiment using a different task. The replication used a binary version of the Dates task reported in Experiment B.1. Unlike in the Dots task above, and because the study was newly designed for this work, participants in the Dates task were split into conditions so that half received feedback while learning about the advisors and half did not.

3.1.2.1 Open scholarship practices

This experiment was preregistered at https://osf.io/5xpvq. The experiment data are available in the esmData package for R (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html.

3.1.2.2 Method

62 participants each completed 52 trials over 4 blocks of the binary version of the Dates task§2.1.3.2. On each trial, participants were presented with an historical event that occurred on a specific year between 1900 and 2000. They were given a date and asked whether the event occurred before or after that date, indicating their confidence in their decision by selecting an appropriate point on the relevant answer bar. Participants then received advice indicating which of the two bars (before or after) was supposedly the correct answer. Participants could then mark a final response in the same manner as their original response.

Participants started with 1 block of 10 trials that contained no advice to allow them to familiarise themselves with the task. All trials in this section included feedback for all participants indicating whether or not the participant's response was correct.

Participants then did 2 trials with a practice advisor to get used to receiving advice. They also received feedback on these trials. They were informed that they would "receive advice from advisors" to "help you complete the task". They were told that the "advisors aren't always correct, but they are quite good at the task", and informed that they should "identify which advisors are best" and "weigh their advice accordingly".

Participants then performed 3 blocks of trials that constituted the main experiment. The first two of these were Familiarisation blocks where participants had a single advisor in each block for 14 trials, plus 1 attention check.

Table 3.5: Advisor advice profiles for Dates task with in/accurate advisors

	Probability of agreement (%)			
Advisor	Participant correct	Participant incorrect	Overall[a]	Overall accuracy[a]
High accuracy	.800	.200	.500	.800
Low accuracy	.590	.410	.500	.590

[a] Where participants' initial estimate accuracy is 50%

Participants were split into four conditions that produced differences in their experience of these Familiarisation blocks. These conditions were whether or not they received feedback, and which of the two advisors they were familiarised with first.

Finally, participants performed a Test block of 10 trials that offered them a choice on each trial of which of the two advisors they had encountered over the last two blocks would give them advice. No participants received feedback during the test phase.

Advice profiles The High accuracy and Low accuracy advisor profiles issued binary advice (endorsing either the 'before' or 'after' column) probabilistically based on whether or not the participant had selected the correct column in their initial estimate (Table 3.5). The High accuracy advisor agreed with the participant's initial estimate on 80% of the trials where the participant was correct, but on only 20% of the trials in which the participant was incorrect, meaning that the High accuracy advisor was correct 80% of the time. Using an analogous setup, the Low accuracy advisor was correct 59% of the time. To the extent that a participant was better than chance in answering the questions, the High accuracy advisor profile would agree more frequently. This mimics the hypothesised relationship wherein agreement between advisors and judges is driven by shared access to the truth.

3.1.2.3 Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. 4 participants had a total of 5 trials removed in this way, representing 0.21% of all trials. Participants were then excluded for having fewer than 11 trials remaining, fewer than 10 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. These criteria led to no participants being excluded from this experiment.

Task performance Before exploring participants' advisor choice behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarisation phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor, although as mentioned above a small number of trials were dropped from analysis where response times were overly long.

For the purposes of exploring participants' performance on the task, the conditions are pooled together. The participants were randomly assigned to conditions, and thus we know any differences in performance between conditions are random.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial estimates and final decisions can be seen in Figure 3.7. The distribution of these response times helps characterise some differences between the Dots task and the Dates task. In the former, decisions for both initial estimates and final decisions are tightly clustered, with a clear structure and pattern to the responses for all participants. In the Dates task however, response times are not only longer, but they are also much more varied within participants. Some increase in variance is expected with an increase in mean, especially with fewer trials for each participant, but the extent of the differences clearly shows that the tasks provide participants with different experiences: the Dots task is tightly rhythmic and repetitive, while the Dates task is more heterogeneous.

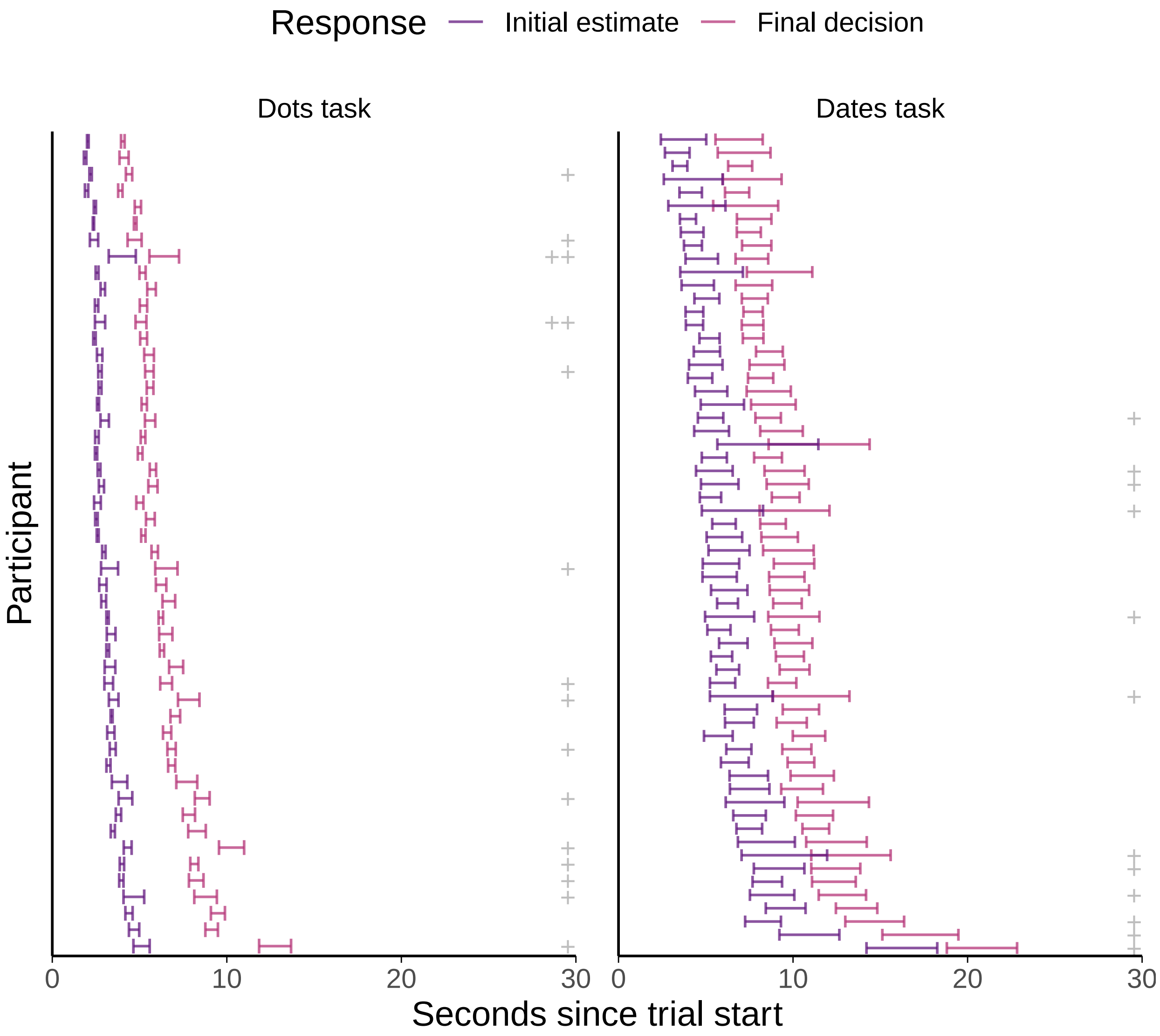


Figure 3.7: Response times for the Dots and Dates tasks with in/accurate advisors. Each row indicates a single participant's trials. The error bars show the 95% confidence intervals of the mean response time for each decision. The plusses on the right show the number of trials where response times were more than 3 standard deviations away from the mean of all Dates task final response times (rounded to the next 10s): + = 1-5 trials, ++ = 6-10 trials.

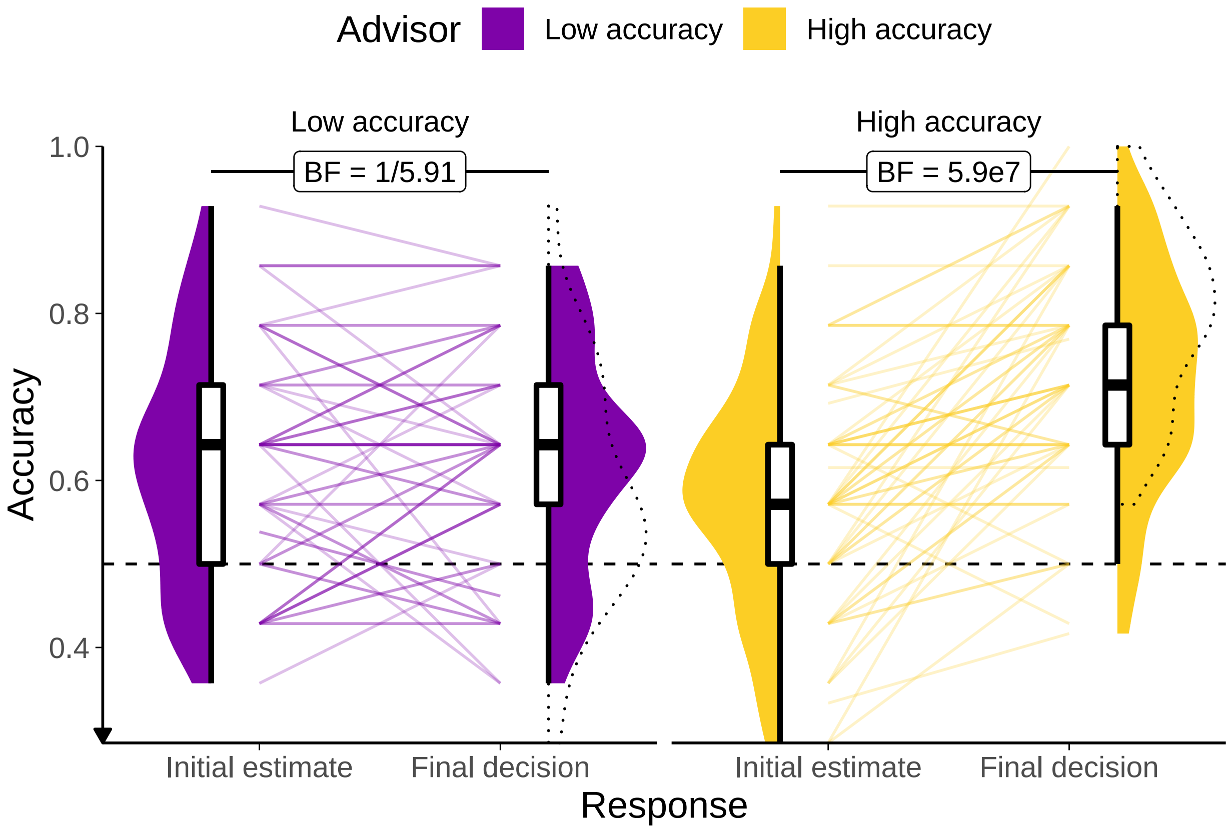


Figure 3.8: Response accuracy for the Dates task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.
Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

Accuracy Unlike in the Dots version of the task, participant accuracy is not controlled because it depends on participants' existing knowledge (and guesses) across a relatively small and varied set of questions. Correspondingly, accuracy varied substantially across participants (Figure 3.8). Figure 3.8 also shows that participants managed to improve their performance from their initial estimates to their final decisions with both advisors (F(1,61) = 36.40, p < .001; $\mathrm{M}_{\mathrm{FinalDecision}}$ = 0.68 [0.65, 0.70], $\mathrm{M}_{\mathrm{InitialEstimate}}$ = 0.60 [0.57, 0.63]). This is likely because the advisors themselves were more accurate than the participants, so following their

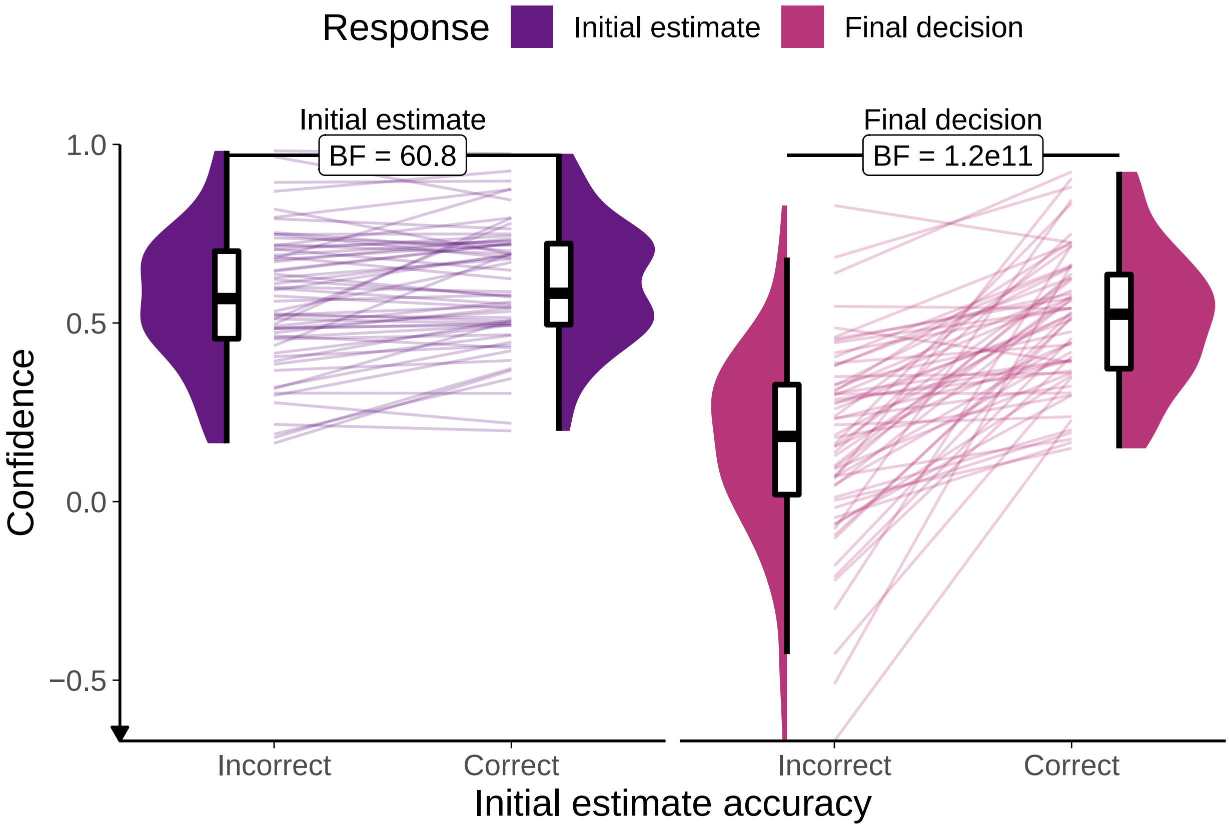


Figure 3.9: Confidence for the Dates task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

advice was generally a good strategy, and the difficulty of the task meant that participants were very willing to be influenced by advice.

As would be expected from participants following advice, the improvement in accuracy from initial estimates to final decisions was greater for the High accuracy advisor than the Low accuracy advisor (F(1,61) = 32.46, $p < .001$; $\text{M}_{\text{Improvement|HighAccuracy}}$ = 0.15 [0.11, 0.18], $\text{M}_{\text{Improvement|LowAccuracy}}$ = 0.01 [-0.02, 0.04]).

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect (Figure 3.9). Participants' initial estimates and final decisions were both systematically more confident when the initial estimate was correct as compared to incorrect (F(1,61) = 102.69, $p < .001$; $\text{M}_{\text{Correct}}$ = 0.56 [0.52, 0.60], $\text{M}_{\text{Incorrect}}$ = 0.36 [0.32,

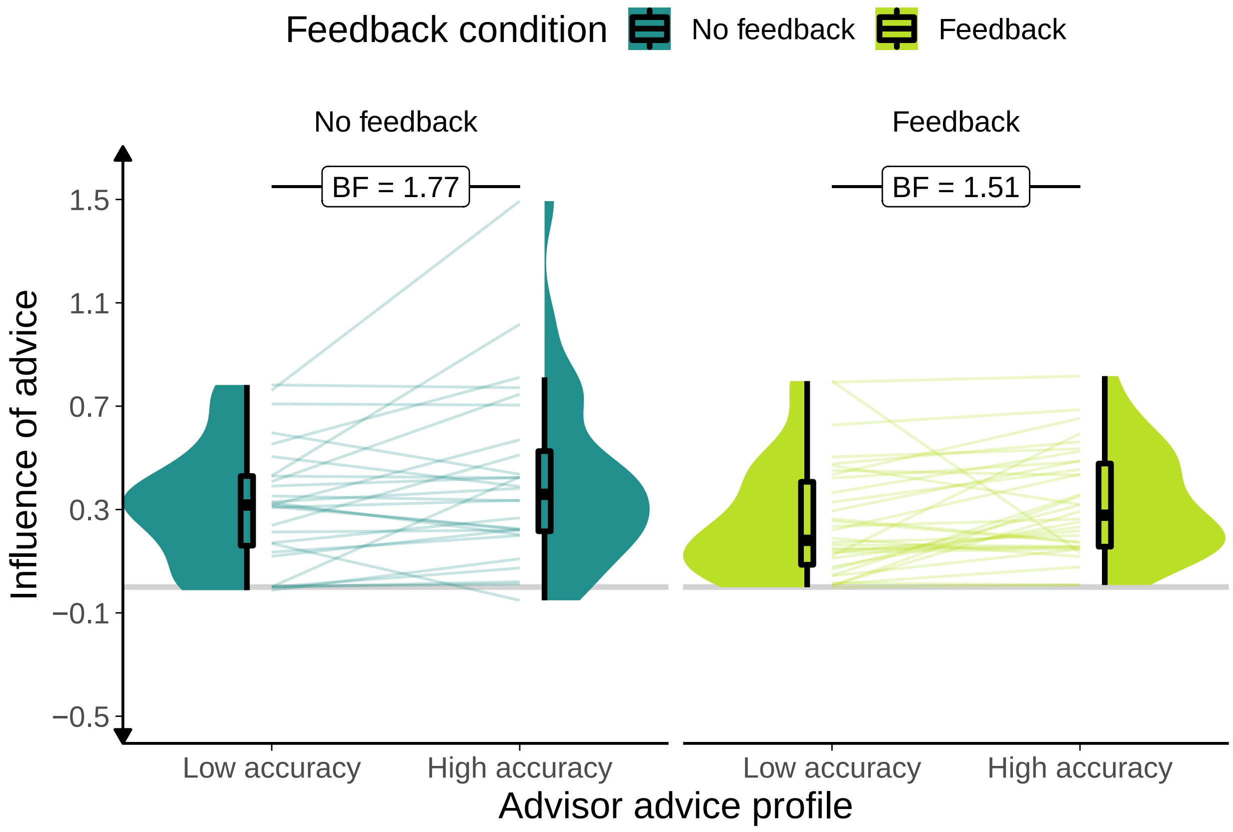


Figure 3.10: Influence for the Dates task with in/accurate advisors.
Participants' weight on the advice for advisors in the Familiarisation stage of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

0.41]). Participants were less confident on final decisions than on initial estimates (F(1,61) = 72.30, p < .001; $\mathrm{M}_{\mathrm{FinalDecision}}$ = 0.34 [0.29, 0.39], $\mathrm{M}_{\mathrm{InitialEstimate}}$ = 0.58 [0.54, 0.63]), and the decrease over time was greatest for the trials where the initial estimate was incorrect (F(1,61) = 68.17, p < .001; $\mathrm{M}_{\mathrm{Increase|Correct}}$ = -0.09 [-0.14, -0.05], $\mathrm{M}_{\mathrm{Increase|Incorrect}}$ = -0.40 [-0.48, -0.31]).

Metacognitive ability Estimates of the participants' metacognitive abilities were highly variable, with many participants displaying below-chance metacognitive ability (Figure 3.11). While this may appear concerning, recall that metacognitive sensitivity and bias vary substantially and cannot be reliably estimated using ROC curves where performance accuracy on the underlying task is highly variable, these

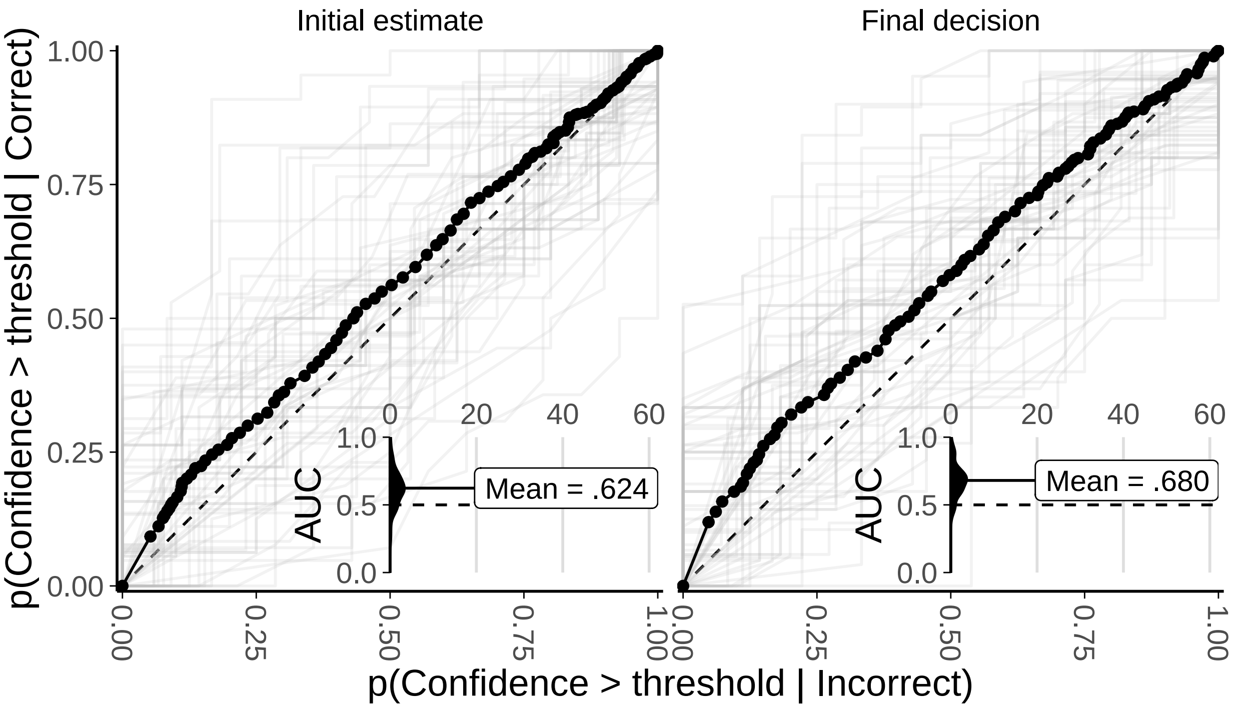


Figure 3.11: Metacognitive performance for the Dates task with in/accurate advisors. Faint lines show Reciever Operator Characteristic (ROC) curves for individual participants, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers. The inset plot shows the distribution of areas under the ROC, and the label gives the mean value.

Table 3.6: Advisor agreement for Dates task with in/accurate advisors

Advisor	Target|correct	Actual|correct	Target|incorrect	Actual|incorrect
High accuracy	.800	.787	.200	.216
Low accuracy	.590	.606	.410	.438

Table 3.7: Advisor accuracy for Dates task with in/accurate advisors

Advisor	Target accuracy	Mean accuracy
High accuracy	.800	.797
Low accuracy	.590	.585

values do not necessarily give cause for alarm.

Performance on the underlying task and metacognitive ability were correlated (Initial estimates: r(60) = 2.550 [.068, .522], p = .013; Final decisions: r(60) = 4.798 [.319, .686], $p <$ .001), showing that, as one might expect, participants with a greater ability to perform the Dates task have a greater insight into their performance on the Dates task. This in turn suggests that, despite the low number of trials on the task, we are able to obtain meaningful insights into participants' metacognitive abilities, albeit without being able to precisely estimate the metacognitive sensitivity or bias for an individual participant.

Advisor performance The advice is generated probabilistically from the rules described previously (Advice profiles§3.1.2.2). The advisors agreed with participants contingent on the accuracy of the participants' initial estimates at close to the target rates (Table 3.6). This meant that advisors were as accurate overall as they were intended to be in the Familiarisation phase (Table 3.7). Most (57/62, 91.94%) participants experienced the High accuracy advisor as providing more accurate advice than the Low accuracy advisor.

Advisor influence The High accuracy advisor was substantially more influential than the other Low accuracy advisor (F(1,60) = 9.98, p = .002; $\text{M}_{\text{HighAccuracy}}$ = 0.36 [0.29, 0.43], $\text{M}_{\text{LowAccuracy}}$ = 0.28 [0.22, 0.34]). This tendency did not differ significantly between the group who received trial-by-trial feedback and the group who did not receive feedback (F(1,60) = 0.20, p = .652; $\text{M}_{\text{High-LowAccuracy|NoFeedback}}$

= 0.09 [0.01, 0.18], $\mathrm{M}_{\text{High-LowAccuracy|Feedback}}$ = 0.07 [0.01, 0.14]). Nor did the participants in the feedback and no feedback conditions appear to differ in the extent to which they took advice (F(1,60) = 1.79, p = .186; $\mathrm{M}_{\text{NoFeedback}}$ = 0.36 [0.26, 0.47], $\mathrm{M}_{\text{Feedback}}$ = 0.29 [0.22, 0.35]).

These influence measurements are calculated on the Familiarisation phase trials in which participants are not offered a choice of advisor. It is during this phase that participants are learning about the value of the advice (especially in the Feedback condition), and thus any influence on later trials may be diluted by low influence on trials which occur before an advisor has had time to develop a reputation as reliable. This means that influence cannot be used as a reliable outcome measure for this experimental design, but it is nevertheless useful to explore to get a sense of how participants responded to the advice. An inspection of the individual participants' data shows that very few participants had large influence differences between advisors (Figure 3.10).

Hypothesis test The key analysis in this experiment explores the participants' preferences for picking the High accuracy advisor over the Low accuracy advisor. In the No feedback condition the mean of the distribution of participant picking preferences between the advisors was equivalent to chance (t(27) = -0.93, p = .363, d = 0.18, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/3.37; $\mathrm{M}_{\text{NoFeedback}}$ = 0.45 [0.33, 0.57], μ = 0.5). This is a different result to that observed in the Dots task§3.1.1.3, which also had no feedback. Preferences were quite evenly distributed across the full range of directions and strengths, with a slight numerical advantage for the Low accuracy advisor (Figure 3.12).

In the Feedback condition the mean of the distribution of selection rates was clearly different from chance. The High accuracy advisor was preferred by more participants, and preferred more strongly (t(33) = 3.41, p = .002, d = 0.58, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 19.7; $\mathrm{M}_{\text{Feedback}}$ = 0.67 [0.57, 0.78], μ = 0.5). The modal selection strategy was to select the High accuracy advisor at every opportunity. This indicates that participants could identify the more accurate advisor when feedback was provided

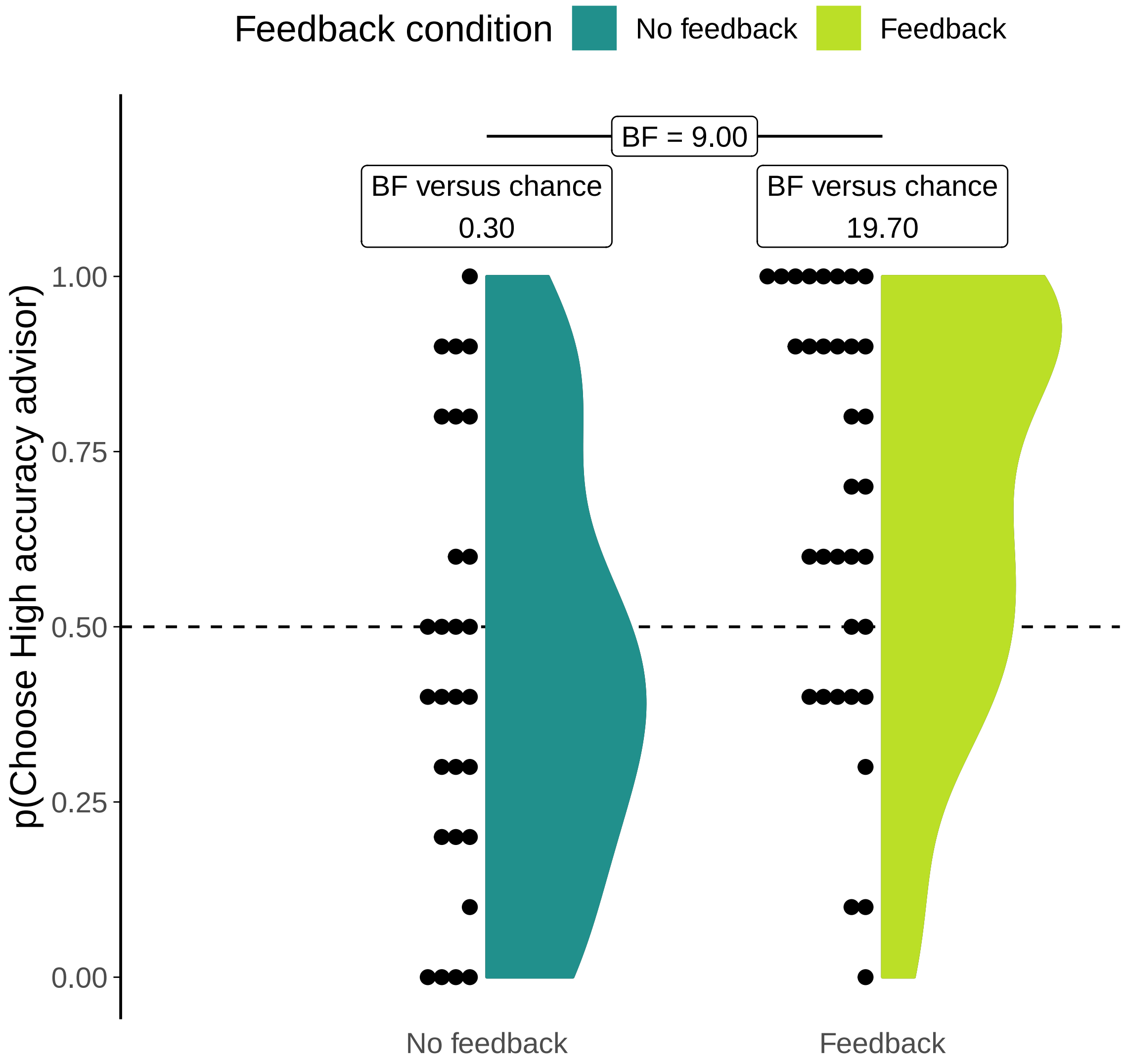


Figure 3.12: Dates task advisor choice for in/accurate advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarisation phase, but not during the Choice phase.

and preferred to receive advice from that advisor. Interestingly, this meant that there was a difference in participants' preference for picking the High accuracy advisor according to their experimental condition ($t(57.02) = 2.95$, $p = .005$, $d = 0.75$, $\mathrm{BF}_{\mathrm{H1:H0}} = 8.99$; $\mathrm{M}_{\mathrm{Feedback}} = 0.67$ [0.57, 0.78], $\mathrm{M}_{\mathrm{NoFeedback}} = 0.45$ [0.33, 0.57]).

It was discovered after the completion of experiments that the advisor position (whether the advisor appears on the top or bottom of the advisor choice panel) was not counterbalanced between advisors. This was true for all the Dates task experiments reported in this chapter. It had been decided during development of the tests to keep the advisors in the same position for every trial so that participants did not get mixed up between them. Together, this meant that the High accuracy advisor always appeared at the top and the Low accuracy advisor always appeared at the bottom, for every trial for every participant. We are thus unable to confirm that pick rate differences, or the absence of those differences, are caused by participants' preferences for the advice the advisor would provide or the position the advisor was in on the screen. Furthermore, it is possible that any genuine preference for one advisor over the other was *induced* by the position, rather than independent of it (Zajkowski and Zhang 2021).

Follow-up tests

Ability of participants It is plausible that participants who were better at the task had more insight into which of their advisors was more accurate. There was not enough evidence to determine whether participants in the No feedback condition selected the High accuracy advisor more frequently where they were more accurate themselves ($\mathrm{r}(26) = -.108$, $p = .586$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.14$) or more well calibrated (as measured by area under the Receiver Operator Characteristics curve for initial estimates; $\mathrm{r}(26) = .157$, $p = .424$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/1.85$).

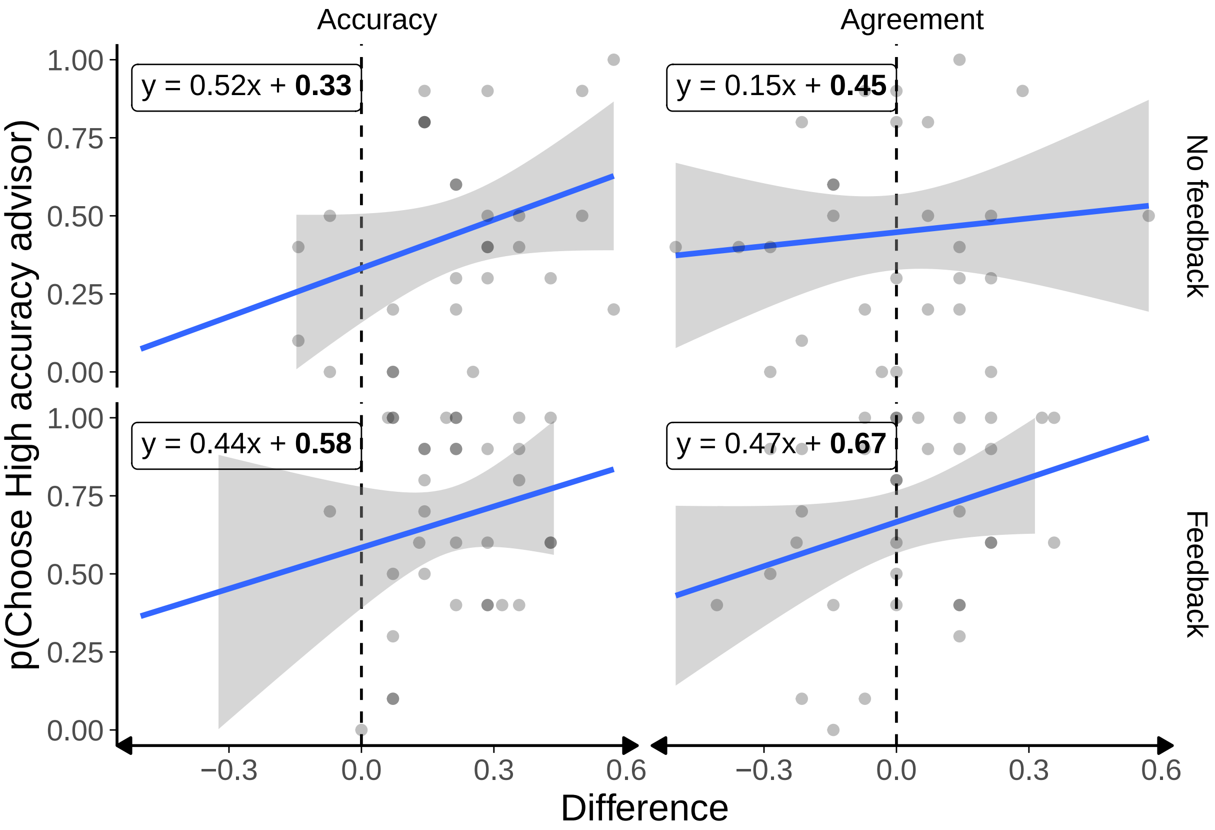


Figure 3.13: Preference predictors in the Dates task with in/accurate advisors. Scatter plots of participants' experience with advisors in terms of agreement or accuracy rates. Differences are expressed as the experienced rate for the High accuracy advisor minus the experienced rate for the Low accuracy advisor during the Familiarisation phase. Numbers in bold in the regression equations are significant at p < .05.

Table 3.8: Experienced accuracy difference effects in the Dates task with in/accurate advisors

Effect	Estimate	SE	t	p	
(Intercept)	0.33	0.09	3.91	$<$ .001	$*$
Accuracy	0.52	0.29	1.78	.080	
FeedbackFeedback	0.25	0.13	1.97	.053	
Accuracy:FeedbackFeedback	-0.08	0.49	-0.16	.875	

Model fit: $F(4.5, 3) = 58$; p .007; $R^2_{adj} = .147$

Table 3.9: Experienced agreement difference effects in the Dates task with in/accurate advisors

Effect	Estimate	SE	t	p	
(Intercept)	0.45	0.06	7.96	$<$ .001	$*$
Agreement	0.15	0.26	0.58	.565	
FeedbackFeedback	0.22	0.08	2.88	.006	$*$
Agreement:FeedbackFeedback	0.32	0.37	0.88	.385	

Model fit: $F(4.1, 3) = 58$; p .010; $R^2_{adj} = .134$

Experience of advisors The stochastic nature of the advisors' advice meant that there was some variation in the participants' experience of the advisors. Linear models were run predicting advisor choice behaviour based on experienced differences in accuracy (Table 3.8) and agreement (Table 3.9). Bayesian linear models were run obtaining Bayes Factors of leaving each component out of a model containing experienced agreement or accuracy difference, feedback condition, and their interaction, as well as a random factor for the participant's identity.

In aggregate, the models indicated that participants in the Feedback group had a stronger preference for the High accuracy advisor than participants in the No feedback group, regardless of their actual experience of advisor agreement ($\text{BF}_{\text{+Feedback:-Feedback}} = 13.5$) or accuracy ($\text{BF}_{\text{+Feedback:-Feedback}} = 8.87$). There was no evidence of a relationship between participants' advisor preference and their experience of either advisor accuracy ($\text{BF}_{\text{+Accuracy:-Accuracy}} = 1.60$) or advisor agreement ($\text{BF}_{\text{+Agreement:-Agreement}} = 1.16$), but neither of these had evidence strong enough to suggest the absence of such a relationship. The strongest evidence for the absence of an effect was for the interaction between feedback condition and experienced accuracy ($\text{BF}_{\text{+Interaction:-Interaction}} = 1/2.94$) or agreement ($\text{BF}_{\text{+Interaction:-Interaction}} = 1/2.09$), but this was also not beyond the stated threshold of 1/3.

The vagueness of the results is is not entirely surprising because, as with the accuracy correlations discussed above, there was relatively little variation in experience of advisors so effects might be expected to be small and difficult to detect. These uninformative results are typical of those across all Dots task experiments in this chapter. To save space, these analyses are not reported for subsequent Dots task experiments.

3.1.3 Discussion

We investigated whether more accurate advisors would be preferentially selected by participants when participants were unable to use feedback to evaluate the quality of advisors. We performed two experiments using different tasks, and found mixed results. In the branch of the Dates task where feedback was provided, participants had a clear preference for the more accurate advisor. This preference was also seen in the Dots task, in which no participants received feedback. Contrary to these results, however, participants in the Dates task who did not receive feedback did not show a systematic preference for either advisor.

The difference between the task results where feedback was denied to participants is probably due to the Dates task being a generally more difficult task for participants than the Dots task. This extra difficulty likely meant that participants in the Dates task were unable to tell the advisors apart. If the Pescetelli and Yeung (2021) model of metacognitive evaluation of advice is accurate, participants may have been subjectively very unsure of whether their answer was correct or incorrect, and thus in the absence of feedback they cannot glean insight into the accuracy of advice by attending to whether or not the advice contradicts their initial estimate. Alternatively, if this process is not driving performance in the Dates task, the additional difficulty may simply have meant that participants' strong desire to see advice (Gino and Moore 2007) may have rendered the relative quality of the advice unimportant.

In both tasks the advisor preferences included many participants whose preference was either neutral (all tasks) or in favour of the Low accuracy advisor (Dates task). This variability in pick rates from the participants in the No feedback condition of the Dates task, suggests that preferences are diverse both in terms of direction and strength in the absence of any systematic effects. There is substantial variability in pick rates for participants in the Feedback condition of the Dates task, too, indicating that, compared to the No feedback condition, everyone might have nudged their preference a bit towards the High accuracy advisor.

Pick rates in the Dots task were also varied, but rather than being evenly spread (as in the No feedback Dates task participants) or massed in favour of the High accuracy advisor with a long, fat tail including exclusive selection of the Low accuracy advisor, Dots task participants were massed in the centre with a long tail out to strong preference for the High accuracy advisor. The Dots task data may have reduced variability because there were more trials that offered the participants a choice of advisor, and the novelty value of the ignored advisor may have increased relative to the chosen advisor as the test phase progressed. Alternatively, participants making repeated choices may have eventually felt that continuing to ignore one advisor was unfair, and that pragmatic reasons for including the opinion of a less expert voice outweighed the performance-maximisation reasons for not including that voice (Mahmoodi et al. 2015).

Another explanation for the difference may be the level of engagement with the tasks. If participants in the Dates task were more engaged with the more challenging and (subjectively but consistent with participant feedback) more enjoyable task, picking of advisors may have been more deliberative than in the Dots task, where the repetitive nature of the trials could have led to disengagement and random advisor choice behaviour for some participants.

The results of these studies were mixed in terms of supporting our hypothesis that more expert advisors would be discriminated and preferentially picked by participants even in the absence of feedback. The underlying mechanism we believe to be responsible for evaluating advisors in the absence of feedback is agreement, and thus a more powerful test of the mechanism is to move from demonstrating the phenomenon (detection of accuracy differences without feedback) to demonstrating the mechanism (discrimination based on agreement differences). This investigation of agreement as a mechanism for driving advisor evaluation in the absence of feedback is the subject of the next experiments.

3.2 Effects of advisor agreement on advisor choice

Experiments 1A§3.1.1 and 1B§3.1.2 revealed differences in how participants selected the advisors between the Dots task (which has no feedback) and the No feedback condition of the Dates task for High versus Low accuracy advisors. We may expect more pronounced effects in the absence of feedback when contrasting High versus Low agreement advisors, because we expect that agreement is the driving force behind the accuracy differences where feedback is not provided. Pescetelli and Yeung (2021) demonstrated that advisors who agree more frequently are more influential (regardless of the presence of feedback, but especially without it) in a lab-based perceptual decision-making task. Here we explored the impact of agreement on choice of advisor. Experiment 2A§3.2.1 looks at this effect in the Dots task, while Experiment 2B§3.2.2 does the same for the Dates task.

3.2.1 Experiment 2A: advisor agreement effects in the Dots task

3.2.1.1 Open scholarship practices

Due to an oversight, this experiment was not preregistered. The experiment data are available in the `esmData` package for R (Jaquiery 2021c), and also directly from https://osf.io/8cnpq/. A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html.

3.2.1.2 Method

68 participants each completed 368 trials over 7 blocks of a perceptual decision-making task. Each trial consisted of three phases: participants gave an initial estimate (with confidence) of which of two briefly presented boxes contained more dots; received advice on their decision from an advisor; and made a final decision (again, with confidence).

Table 3.10: Advisor advice profiles for Dots task with dis/agreeing advisors

	Probability of agreement			
Advisor	Participant correct	Participant incorrect	Overall	Overall accuracy
High agreement	.840	.610	.773	.709
Low agreement	.660	.170	.518	.709

Participants started with 2 blocks of 60 trials that contained no advice. The first 3 trials were introductory trials that explained the task. All trials in this section included feedback indicating whether or not the participant's response was correct.

Participants then did 5 trials with a practice advisor. They were informed that they would "get **advice** from an advisor to help you make your decision [original emphasis]", and that "advice is not always correct, but it is there to help you: if you use the advice you will perform better on the task."

Participants then performed 2 sets of 2 blocks each. These sets consisted of 1 Familiarisation block of 60 trials in which participants were assigned one of two advisors. The Familiarisation block was followed with a Test block of 60 trials in which participants could choose between the advisors they encountered in the Familiarisation block. The participants saw different pairs of advisors in each set, with each pair consisting of one advisor with each of the advice profiles.

Advice profiles The two advisor profiles (Table 3.10) used in the experiment were High agreement and Low agreement. These advisors were defined in terms of their likelihood of agreement with participants' correct and incorrect initial estimates, while being matched for objective accuracy. The High agreement advisor gave advice that endorsed the same answer side as the participant's initial estimate 77.3% of the time while the Low agreement advisor agreed with the participant 51.8% of the time. These overall agreement rates were split based on the target accuracy rates for participants' initial estimates to achieve balanced overall accuracy rates between advisors.

Table 3.11: Participant exclusions for Dots task with dis/agreeing advisors

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	7
Skewed confidence categories	12
Too many participants	0
Total excluded	**18**
Total remaining	**50**

3.2.1.3 Results

Exclusions In line with the preregistration, participants' data were excluded from analysis where they had an average accuracy below 0.6 or above 0.85, did not have choice trials in all confidence categories (bottom 30%, middle 40%, and top 30% of prior confidence responses), had fewer than 12 trials in each confidence category, or finished the experiment after 50 participants had already submitted data which passed the other exclusion tests. Overall, 18 participants were excluded, with the details shown in Table 3.11.

Task performance Basic behavioural performance was similar to that observed with the same Dots task in Experiment 1A§3.1.1.3. Initial estimate accuracy converged on the target 71%, and participants may have benefited from advice in terms of their final decisions being more accurate than their initial estimates (F(1,49) = 5.48, p = .023; $\mathrm{M}_{\mathrm{Final}}$ = 0.73 [0.72, 0.74], $\mathrm{M}_{\mathrm{Initial}}$ = 0.72 [0.71, 0.73]; Figure 3.14). There was no evidence of a general difference in participants' overall accuracy between advisors (F(1,49) = 0.66, p = .420; $\mathrm{M}_{\mathrm{LowAgreement}}$ = 0.72 [0.71, 0.73], $\mathrm{M}_{\mathrm{HighAgreement}}$ = 0.73 [0.72, 0.74]), nor was there evidence of a difference in participants' improvement in accuracy between advisors (F(1,49) = 1.33, p = .255; $\mathrm{M}_{\mathrm{Improvement|LowAgreement}}$ = 0.02 [0.00, 0.03], $\mathrm{M}_{\mathrm{Improvement|HighAgreement}}$ = 0.00 [-0.01, 0.02]).

Figure 3.15 and ANOVA indicated that participants were more confident in their answers when their initial estimate was correct as compared with incorrect (F(1,49) = 152.76, $p < .001$; $\mathrm{M}_{\mathrm{Correct}}$ = 28.78 [26.34, 31.21], $\mathrm{M}_{\mathrm{Incorrect}}$ = 21.30 [18.60, 24.00]),

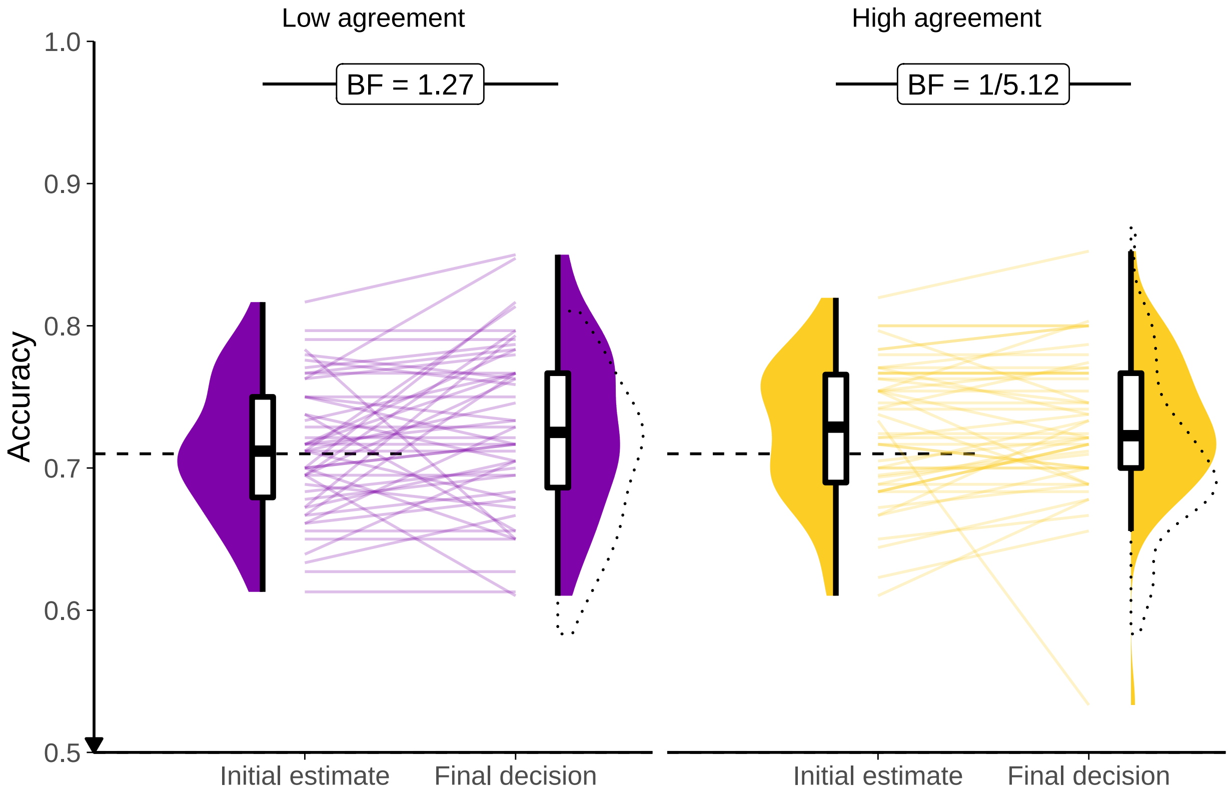


Figure 3.14: Response accuracy for the Dots task with dis/agreeing advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

and less confident in their final decisions than their initial estimates (F(1,49) = 6.44, p = .014; $\mathrm{M}_{\mathrm{Final}}$ = 24.01 [21.46, 26.57], $\mathrm{M}_{\mathrm{Initial}}$ = 26.06 [23.37, 28.75]). These two factors interacted, with confidence only decreasing for final decisions in trials where the initial estimate was incorrect (F(1,49) = 51.45, p < .001; $\mathrm{M}_{\mathrm{Increase|Correct}}$ = 0.81 [-0.62, 2.23], $\mathrm{M}_{\mathrm{Increase|Incorrect}}$ = -4.90 [-7.03, -2.78]).

Perhaps surprisingly, there was no correlation between initial estimate accuracy and confidence (1/3.06), and no evidence for a correlation between final decision accuracy and confidence (1/1.06).

Advisor performance The advice is generated probabilistically from the rules described previously in Table 3.10. It is thus important to get a sense of the

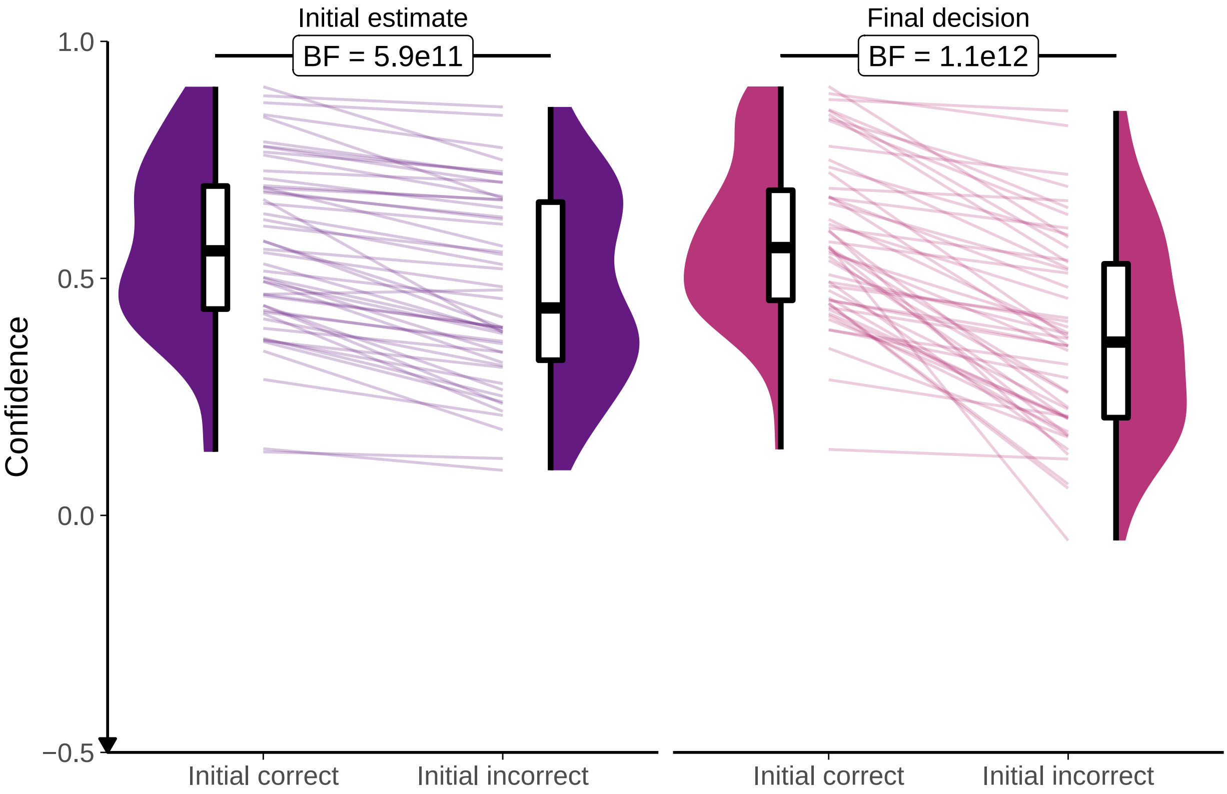


Figure 3.15: Confidence for the Dots task with dis/agreeing advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

Table 3.12: Advisor agreement for Dots task with dis/agreeing advisors

Advisor	Target|correct	Actual|correct	Target|incorrect	Actual|incorrect
High agreement	.840	.832	.610	.631
Low agreement	.660	.651	.170	.166

actual advice experienced by the participants.

The advisors agreed with the participants' initial estimates at close to target rates (Table 3.12), and were as accurate on average as expected (Table 3.13). Nevertheless, some participants experienced in practice 10-20% differences in advisor accuracy

Table 3.13: Advisor accuracy for Dots task with dis/agreeing advisors

Advisor	Target accuracy	Mean accuracy
High agreement	.709	.705
Low agreement	.709	.704

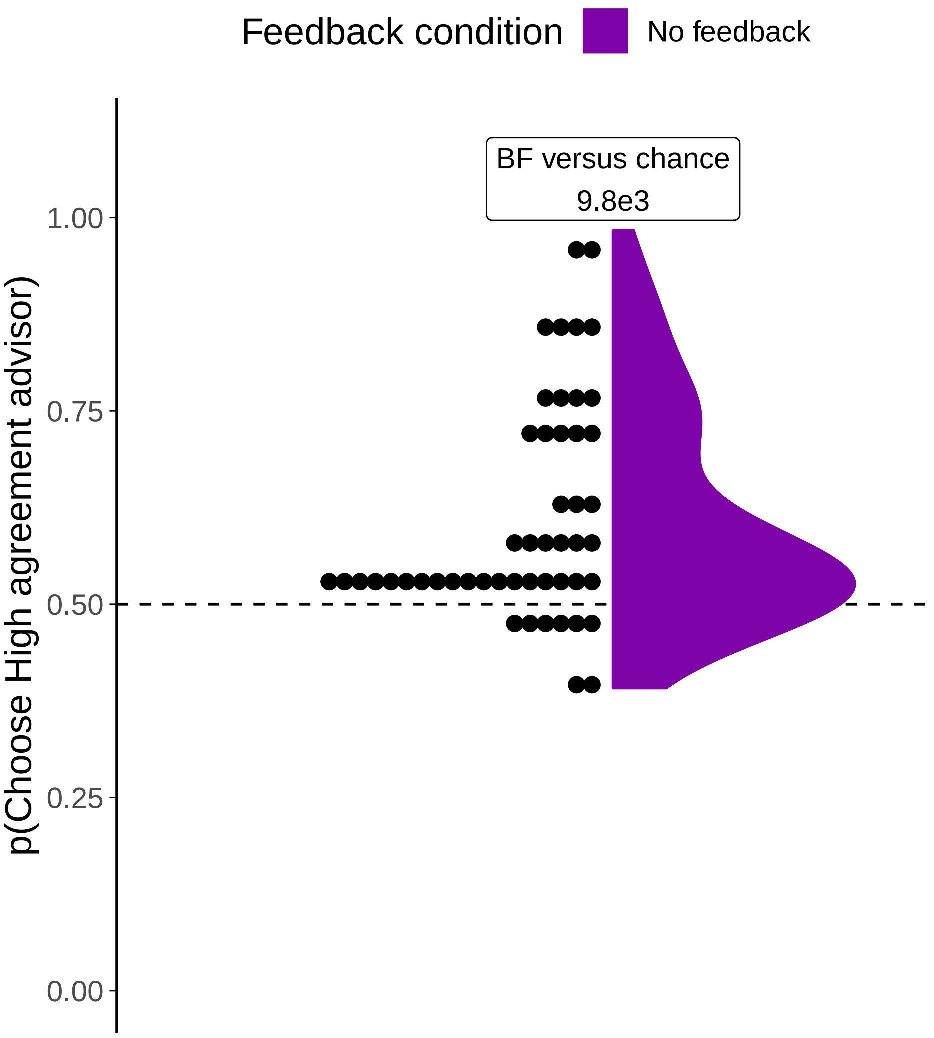


Figure 3.16: Dot task advisor choice for dis/agreeing advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

(although neither advisor was systematically more accurate across participants). All participants experienced the intended relationship wherein the High agreement advisor agreed with them more than the Low agreement advisor.

Hypothesis test Our key analysis concerned whether participants would have a systematic preference for choosing the High agreement advisor when they were given a choice of advisor. Consistent with the key prediction of this experiment, advisor choice varied significantly as a function of advisor agreement rate (Figure 3.16): The High agreement advisor was preferred at a rate greater than that expected by chance ($t(49) = 5.43$, $p < .001$, $d = 0.77$, $\text{BF}_{\text{H1:H0}} = 9.8\text{e}3$; $M = 0.61$ [0.57, 0.65], $\mu = 0.5$). The modal preference remained at chance, but almost all participants who manifested a preference preferred the High agreement advisor.

While this effect is interesting, it is substantially smaller than participants' preference for picking the top advisor regardless of identity ($t(49) = 7.26$, $p < .001$, $d = 1.03$, $\text{BF}_{\text{H1:H0}} = 4.4\text{e}6$; $\text{M}_{\text{P(PickFirst)}} = 0.66$ [0.62, 0.71], $\mu = 0.5$), an effect that we would hope would be random and even out across participants. Note that because the advisor position is well balanced across advisors ($\text{BF}_{\text{H1:H0}} = 1/2.70$; $\text{M}_{\text{P(HighAgreementFirst)}} = 0.51$ [0.50, 0.52], $\mu = 0.5$) the presence of a preference for advisor by position would not cause a preference for an individual advisor.

Summary/Discussion Participants who had a preference for one of the two advisors almost universally preferred the High agreement advisor. These results are in line with the effects of advisor accuracy in the same task as found in Experiment 1A§3.1.1. They are also consistent with our hypothesis that agreement is used as a proxy for feedback when objective feedback is unavailable. Pescetelli and Yeung (2021) found a similar pattern using the same perceptual decision-making task and measuring the influence of advice rather than the choice of advisor. We next explored whether this pattern would also be apparent in the Dates task.

3.2.2 Experiment 2B: advisor agreement effects in the Dates task

As with Experiment 1B§3.1.2, we attempted to replicate the result using the Dates task. Participants in this task were split into conditions depending upon whether or not they received feedback, allowing a direct exploration of the effect of feedback on advisor preference.

3.2.2.1 Open scholarship practices

This experiment was preregistered at https://osf.io/8d7vg. The experiment data are available in the `esmData` package for R (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html.

3.2.2.2 Method

76 participants each completed 52 trials over 4 blocks of the binary version of the Dates task§2.1.3.2. Participants started with 1 block of 10 trials that contained no advice. All trials in this section included feedback for all participants indicating whether or not the participant's response was correct.

Participants then did 2 trials with a practice advisor. They also received feedback on these trials. They were informed that they would "receive advice from advisors" to "help you complete the task". They were told that the "advisors aren't always correct, but they are quite good at the task", and informed that they should "identify which advisors are best" and "weigh their advice accordingly".

Participants then performed 3 blocks of trials that constituted the main experiment. The first two of these were Familiarisation blocks where participants had a single advisor in each block for 14 trials, plus 1 attention check.

Participants were split into four conditions that produced differences in their experience of these Familiarisation blocks. These conditions were whether or not they received feedback, and which of the two advisors they were familiarised with first.

Finally, participants performed a Test block of 10 trials that offered them a choice on each trial of which of the two advisors they had encountered over the last two blocks would give them advice. No participants received feedback during the test phase.

Table 3.14: Advisor advice profiles for Dates task Agreement experiment

	Probability of agreement (%)			
Advisor	Participant correct	Participant incorrect	Overall[a]	Overall accuracy[a]
High agreement	.900	.650	.775	.625
Low agreement	.750	.350	.550	.700

[a] Where participants' initial estimate accuracy is 50%

Advice profiles The High agreement and Low agreement advisor profiles issued binary advice (endorsing either the 'before' or 'after' column) probabilistically based on which column the participant had selected in their initial estimate and whether that was the correct answer (Table 3.14). Unlike in the Dots task above (Experiment 2A§3.2.1), the accuracy of the advisors was not controlled because we were unable to control the participants' accuracy, and advisor accuracy depends upon participant accuracy when agreement rates are fixed.

3.2.2.3 Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. 3 participants had a total of 3 trials removed in this way, representing 0.11% of all trials. Participants were then excluded for having fewer than 11 trials remaining, fewer than 10 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. These criteria led to no participants being excluded from this experiment.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor choice behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarisation phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor. As before (Experiment 1B§3.1.2), the conditions are pooled together while exploring participants' performance on the task.

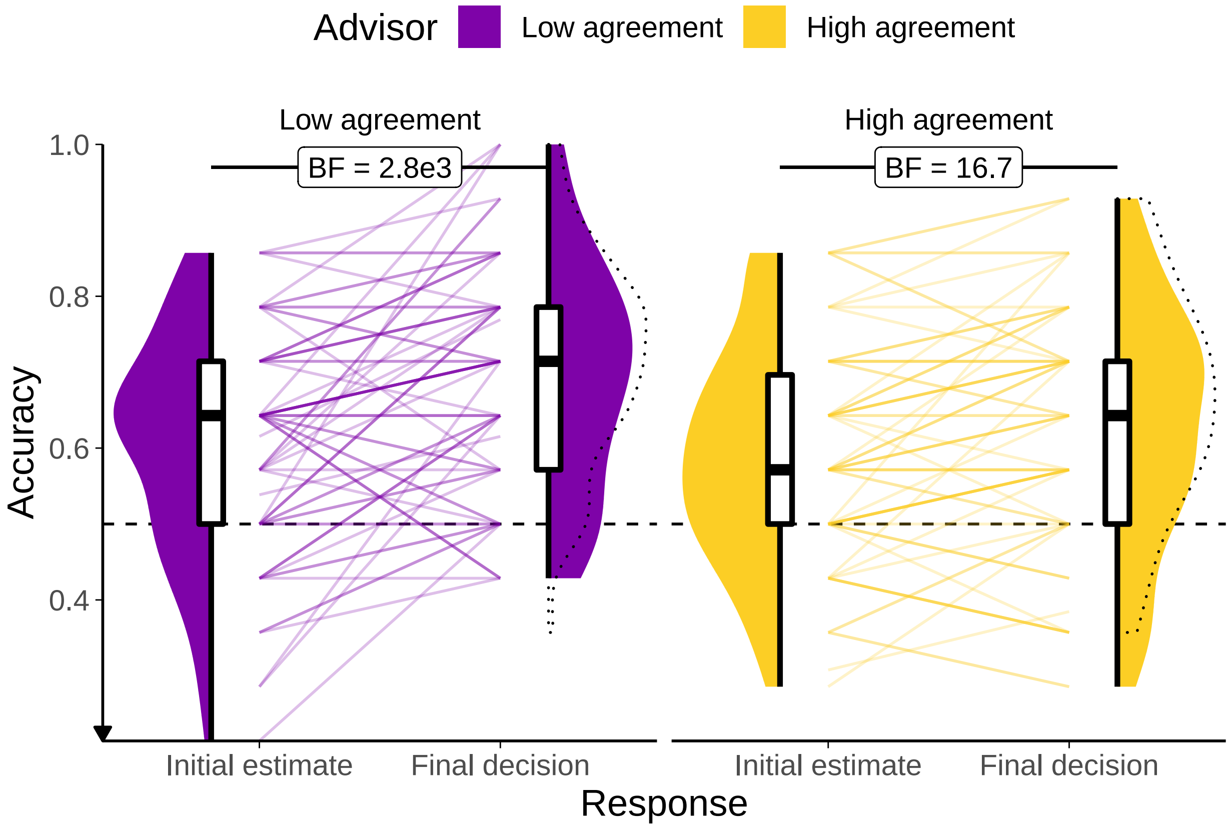


Figure 3.17: Response accuracy for the Dates task with dis/agreeing advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy. Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

There were some similarities to and some differences from the basic behavioural performances compared to the same Dates task in Experiment 1B§3.1.2.3. Participants' accuracy (Figure 3.17), which was uncontrolled in this task, was greater on final decisions than on initial estimates ($F(1,73) = 32.19$, $p < .001$; $M_{\text{FinalDecision}} = 0.66$ [0.63, 0.68], $M_{\text{InitialEstimate}} = 0.60$ [0.57, 0.62]). There was no significant difference between advisors ($F(1,73) = 3.28$, $p = .074$; $M_{\text{HighAgreement}} = 0.61$ [0.57, 0.64], $M_{\text{LowAgreement}} = 0.65$ [0.62, 0.68]), but the increase in final decision accuracy was greater for the Low agreement advisor than the High agreement advisor (F(1,73)

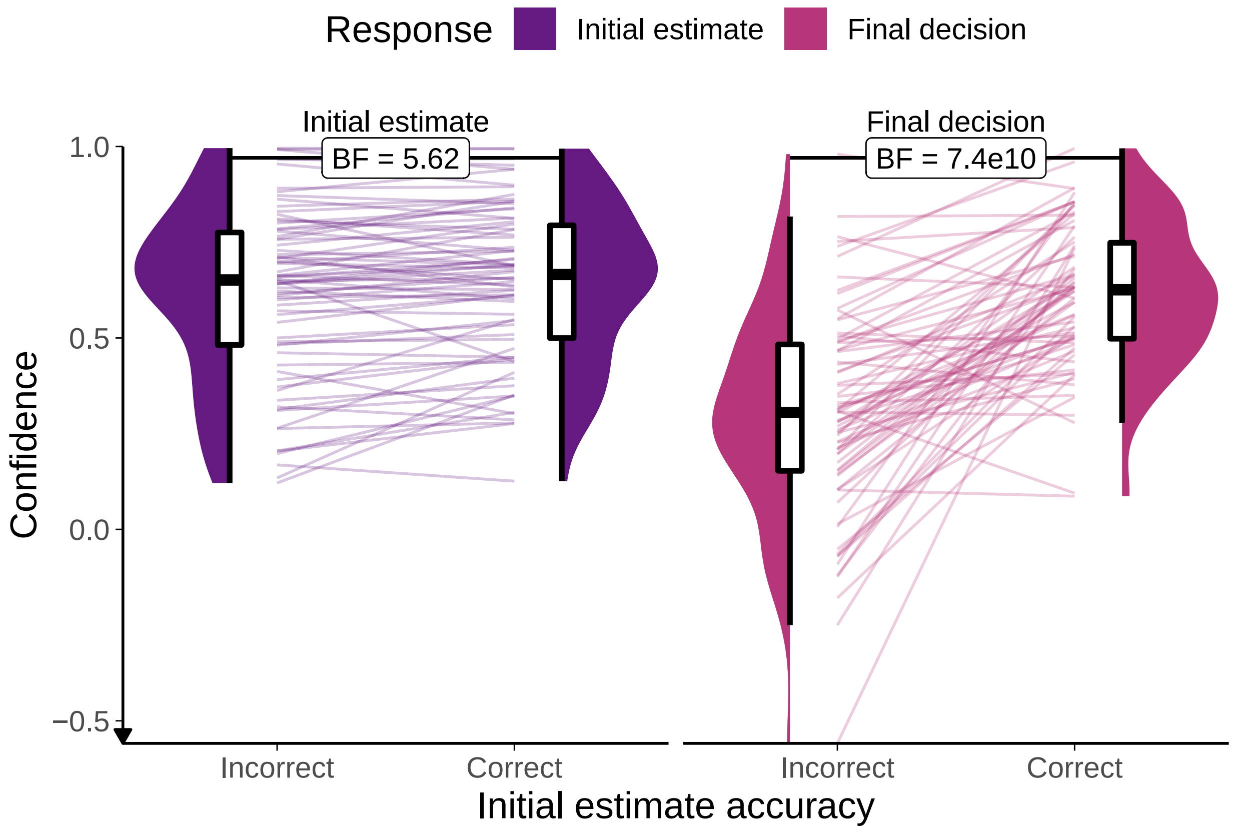


Figure 3.18: Confidence for the Dates task with dis/agreeing advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

= 5.30, p = .024; $\mathrm{M}_{\mathrm{Improvement|HighAgreement}}$ = 0.04 [0.02, 0.06], $\mathrm{M}_{\mathrm{Improvement|LowAgreement}}$ = 0.09 [0.05, 0.12]).

As expected, and as shown in Figure 3.18, participants were systematically more confident when their initial estimate was correct as compared to incorrect (F(1,73) = 90.28, p < .001; $\mathrm{M}_{\mathrm{Correct}}$ = 0.63 [0.58, 0.67], $\mathrm{M}_{\mathrm{Incorrect}}$ = 0.46 [0.41, 0.50]). Participants were less confident on final decisions than on initial estimates (F(1,73) = 61.19, p < .001; $\mathrm{M}_{\mathrm{FinalDecision}}$ = 0.46 [0.41, 0.50], $\mathrm{M}_{\mathrm{InitialEstimate}}$ = 0.63 [0.58, 0.68]), as expected given that the scale allows more scope for reducing than increasing confidence between initial estimate and final decision. This decrease in confidence was greater when the initial estimate was incorrect as compared to correct (F(1,73) = 67.07, p < .001; $\mathrm{M}_{\mathrm{Increase|Correct}}$ = -0.03 [-0.06, 0.00], $\mathrm{M}_{\mathrm{Increase|Incorrect}}$

Table 3.15: Advisor agreement for Dates task Agreement experiment

Advisor	Target|correct	Actual|correct	Target|incorrect	Actual|incorrect
High agreement	.900	.874	.650	.612
Low agreement	.750	.761	.350	.334

Table 3.16: Advisor accuracy for Dates task Accuracy experiment

Advisor	Target accuracy	Mean accuracy
High agreement	.625	.665
Low agreement	.700	.715

= -0.32 [-0.39, -0.24]).

Advisor performance The advice is generated probabilistically from the rules described previously (Advice profiles§3.1.2.2). The advisors agreed with participants contingent on the accuracy of the participants' initial estimates at close to the target rates (Table 3.15). This meant that advisors were distinguished by their overall agreement rates as they were intended to be in the Familiarisation phase. The accuracy of participants' initial estimates was not much above 50%, meaning that the overall accuracy rates of the advisors were similar to those projected (Table 3.16). Most (66/74, 89.19%) participants experienced the High agreement advisor as providing advice that agreed more frequently than the Low agreement advisor.

Hypothesis test Consistent with the result from the Dots task, the key analysis demonstrated that in the No feedback condition participants' preferences for receiving advice from the High agreement advisor were greater than chance ($t(34) = 2.62$, $p = .013$, $d = 0.44$, $\text{BF}_{\text{H1:H0}} = 3.39$; $\text{M}_{\text{NoFeedback}} = 0.63$ [0.53, 0.73], $\mu = 0.5$). The modal preference was to select the High agreement advisor on every Choice trial, and although some participants still showed a preference for hearing advice from the Low agreement advisor, preferences for the High agreement advisor were generally stronger and more frequent (Figure 3.19).

In the Feedback condition, the mean of the participants' selection rates was equivalent to random picking ($t(38) = 0.46$, $p = .648$, $d = 0.07$, $\text{BF}_{\text{H1:H0}} = 1/5.24$; $\text{M}_{\text{Feedback}} = 0.52$ [0.42, 0.62], $\mu = 0.5$). This is consistent with a strategy which

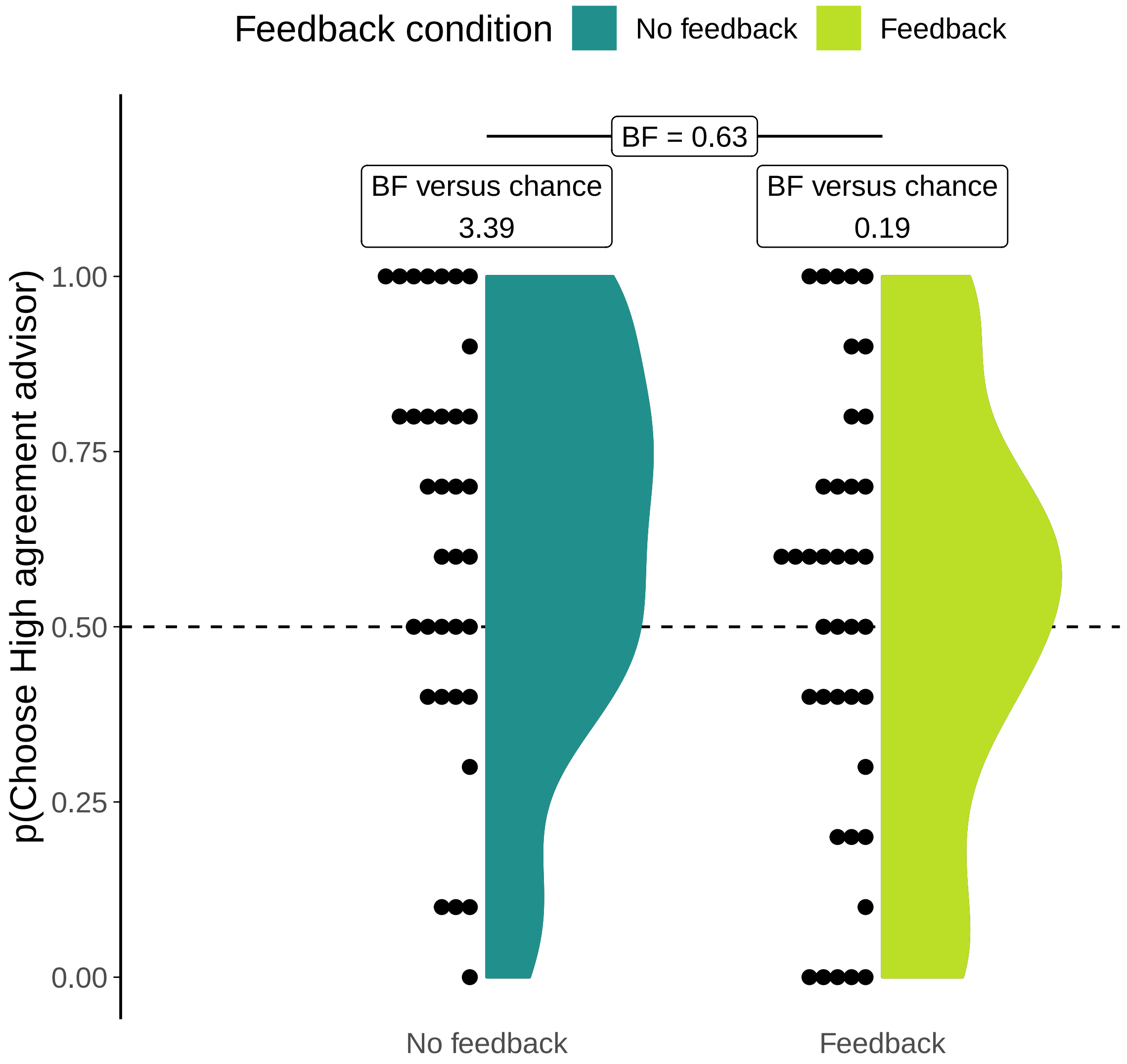


Figure 3.19: Dates task advisor choice for dis/agreeing advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarisation phase, but not during the Choice phase.

attempts to maximise the accuracy of final decisions, because neither advisor would help with this task systematically. The null result here does indicate, however, that there is no strong and clear preference for agreement over and above its accuracy benefits.

Interestingly, despite different patterns of preferences when compared to chance, there was not enough evidence to demonstrate whether the preference patterns for the two conditions were or were not different from one another. This may be a consequence of the variability of preferences: as with the Dates task using High accuracy and Low accuracy advisors (Experiment 1B§3.1.2), the participants in both Feedback and No feedback conditions spanned the entire gamut of preference strengths and directions. One participant in the No feedback group here, for example, never chose the High agreement advisor. Even where there were no systematic effects (No feedback condition in Experiment 1B, Feedback condition here), participants still had a range of preferences with some picking one or the other advisor exclusively. A substantial minority (33.8%) of participants had preference strengths beyond those expected by chance when picking randomly.[5] As noted previously (Experiment 1: Discussion§3.1.3), this is in contrast to the behaviour in the Dots task, where preferences tend to mass towards even pick rates with a long tail differentiating the population preferences from chance.

3.2.3 Discussion

These two experiments investigated the impact of advisor agreement on choice of advice. In both Dots and Dates tasks, where the absence of feedback means advisors' performances cannot be evaluated objectively, participants preferred High agreement advisors to Low agreement advisors. This is consistent with our underlying theory that agreement is used as a mechanism for evaluating advisors in the absence of objective feedback.

When feedback was provided in the Dates task, advisors were selected at rates equivalent to chance overall. Despite this overall chance level of preference in

[5] 5% would be expected to have had strengths of these magnitudes if picking randomly.

the sample, individual participants had a wide range of preference strengths and directions, and this was true for both the experimental conditions. This contrasts with what we see in the Dots task data, for both Experiment 1A§3.1.1 and 2A§3.2.1, where the majority of participants' preferences are moderate, with a minority of participants' more marked preferences producing the systematic effects.

In the Dates task, as we saw in Experiment 1B§3.1.2, systematic effects show up as a general nudging of preferences in one direction rather than as every participant developing similar preferences. This wide variability with some systematic nudging is explicable in terms of the heterogeneity of the Dates task. Participants will have a different level of knowledge on different questions, and we might expect advisors to occasionally offer implausible advice to questions participants consider easy. If this were to happen, participants might weight these occurrences highly, in line with the Pescetelli and Yeung (2021) theory of metacognitive evaluation of advice. More frequently disagreeing advisors will be more likely to disagree on these subjectively easy questions and suffer the reputational consequences. This greater likelihood could mean that the wide spread of behaviour from participants is due to the greater frequency of these important events for the Low accuracy advisor, rather than a consistent effect of agreement alone.

A challenge for this explanation is that in Experiment 1B§3.1.2 we observed participants in the No feedback condition demonstrating a range of preferences that were not systematically different from chance. We would expect, provided participants' performance on subjectively easy questions is above chance, that the High accuracy advisor would have a higher probability of agreement on those questions, just as the High agreement advisor had a higher probability of agreement in Experiment 2B§3.2.2. It is unlikely, but not impossible, that participants' accuracy even on subjectively easy questions was sufficiently bad as to render these effects undetectable.

Another reason why advisor preferences are so varied is that there are reasons why participants might prefer to see advice from the disagreeing advisor. In both Dots and Dates tasks, it was harder to predict what the Low agreement advisor would say, potentially making the advice more interesting. Furthermore, if participants thought they were unlikely to be correct, the Low agreement advisor's advice was more diagnostic of the correct answer because the Low agreement advisor was far less likely to endorse incorrect responses. It may be, therefore, that people evaluate advisors on the basis of agreement, but that what they do on the basis of that evaluation is a matter of personal preference.

In the previous two studies we looked at the effects of advisor accuracy and agreement on the preference for picking those advisors when offered a choice. In order to isolate the effects, agreement was balanced in the accuracy experiments, and accuracy was balanced (as well as we were able) in the agreement experiments. Next, we directly contrast these domains by providing participants with a pairing consisting of a High accuracy advisor and a High agreement advisor, introducing a clear performance cost of preferring to hear advice from the High agreement advisor.

3.3 Effects of accuracy versus agreement

The High versus Low agreement advisor experiment showed that participants tended to prefer to receive agreeing advice when they did not receive feedback. The advisors did not differ in their accuracy (by design), which meant that participants could not increase their performance by selecting one advisor over another. Here we introduce a discrepancy between advisors' objective performance (accuracy) and their subjective performance from the judge's perspective (agreement). By playing off accuracy against agreement we can explore whether participants continue to prefer agreement when there is a cost associated with agreement through a reduction in overall accuracy. We expect that participants will gravitate towards the agreeing advisor who, from their perspective, should appear more accurate, despite the poorer objective performance of that advisor.

3.3.1 Experiment 3A: accuracy versus agreement effects in the Dots task

3.3.1.1 Open scholarship practices

This experiment was preregistered at https://osf.io/f3k4x. The experiment data are available in the esmData package for R (Jaquiery 2021c), and also directly from https://osf.io/y47ec/. A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/c18c26b5da3622988e2261433cf256aae4d19f39/AdvisorChoice/ava.html.

3.3.1.2 Unanalysed data

An initial version of this study was conducted and run as a proper experiment in which participants learned about both advisors simultaneously (preregistered at https://osf.io/5z2fp), but there were no effects in the data. We suspected the absence of effects was because participants had difficulty distinguishing the advisors when they were presented together. The version of the experiment reported here presented one advisor per block in the Familiarisation phase. Data for the unreported null study can be found in the esmData R package (Jaquiery 2021c) and at https://osf.io/26yut/.

3.3.1.3 Method

89 participants each completed 277 trials over 6 blocks of the Dots task. Participants started with 2 blocks of 60 trials that contained no advice. The first 3 trials were introductory trials that explained the task. All trials in this section included feedback indicating whether or not the participant's response was correct.

Participants then did 4 trials with a practice advisor. They were informed that they would "get **advice** from advisors to help you make your decision [original emphasis]", and that "advice is not always correct, but it is supposed to help you perform better on the task."

Table 3.17: Advisor advice profiles for Dots task with accurate versus agreeing advisor

	Probability of agreement			
Advisor	Participant correct	Participant incorrect	Overall	Overall accuracy
High accuracy	.800	.200	.626	.800
High agreement	.800	.800	.800	.626

Table 3.18: Participant exclusions for Dots task with accurate versus agreeing advisors

Reason	Participants excluded
Accuracy too low	1
Accuracy too high	0
Missing confidence categories	6
Skewed confidence categories	18
Too many participants	14
Total excluded	**39**
Total remaining	**50**

Participants then performed 3 blocks of trials that made up the core experiment. There were 2 Familiarisation blocks of 60 trials, with participants seeing one of the two advisors for an entire block in random order. The Familiarisation block was followed with a Test block of 30 trials in which participants could choose between the advisors they encountered in the Familiarisation blocks.

Advice profiles The two advisor profiles used in the experiment were High accuracy and High agreement. The advisors' advice was stochastically generated according to the participant's response. The advisor profiles were not balanced for overall agreement or accuracy rates.

The High accuracy advisor predominantly agreed with correct participant responses and disagreed with incorrect ones. The High agreement advisor agreed with the participant at the same rate regardless of the accuracy of the participant's initial estimate. In practical terms, the difference between the advisors is that the High agreement advisor continued to agree with participants where their initial estimates were incorrect, while the High accuracy advisor did not (Table 3.17).

3.3.1.4 Results

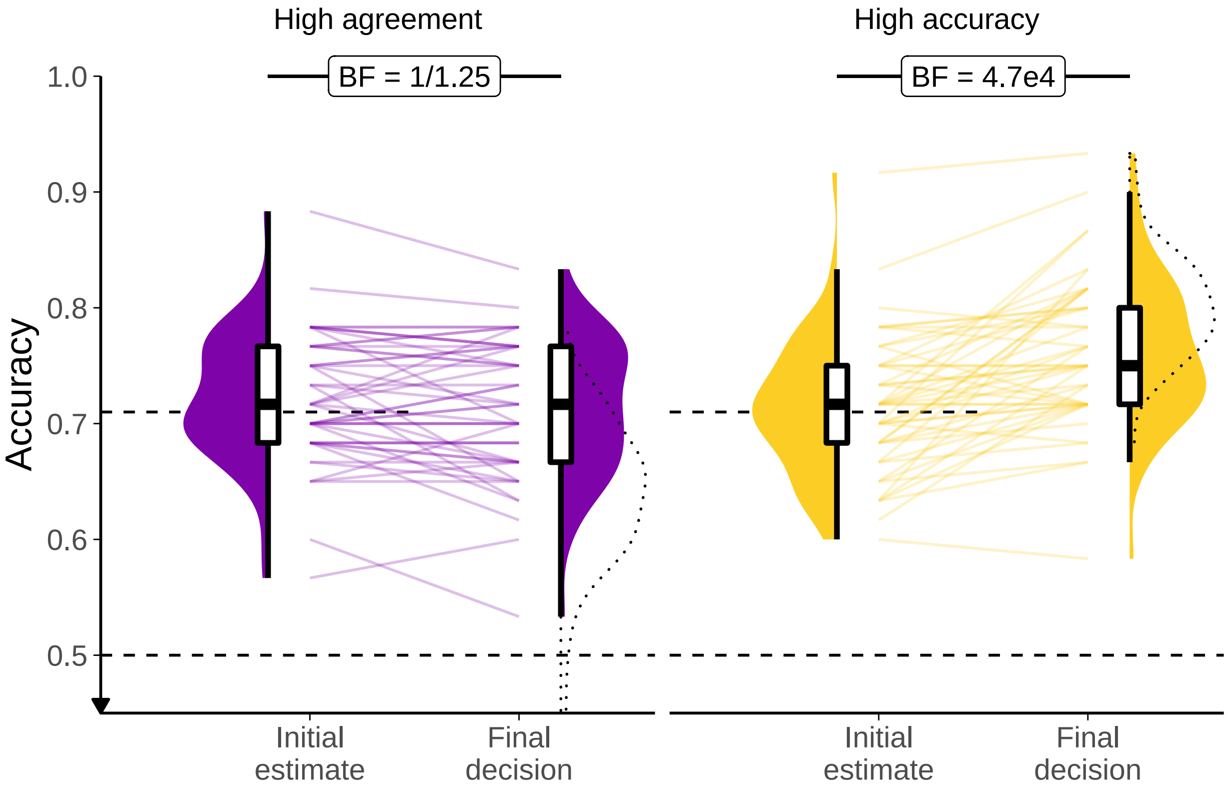


Figure 3.20: Response accuracy for the Dots task with agreeing versus accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

Exclusions In line with the preregistration, participants' data were excluded from analysis where they had an average accuracy below 0.6 or above 0.85, did not have choice trials in all confidence categories (bottom 30%, middle 40%, and top 30% of prior confidence responses), had fewer than 12 trials in each confidence category, or finish the experiment after 50 participants have already submitted data which passed the other exclusion tests. Overall, 39 participants were excluded, with the details shown in Table 3.18. This number is somewhat higher than in the previous experiments, but this is largely due to collecting data in larger batches than previously: a large number of participants were excluded because their data were in excess of the preregistered sample size.

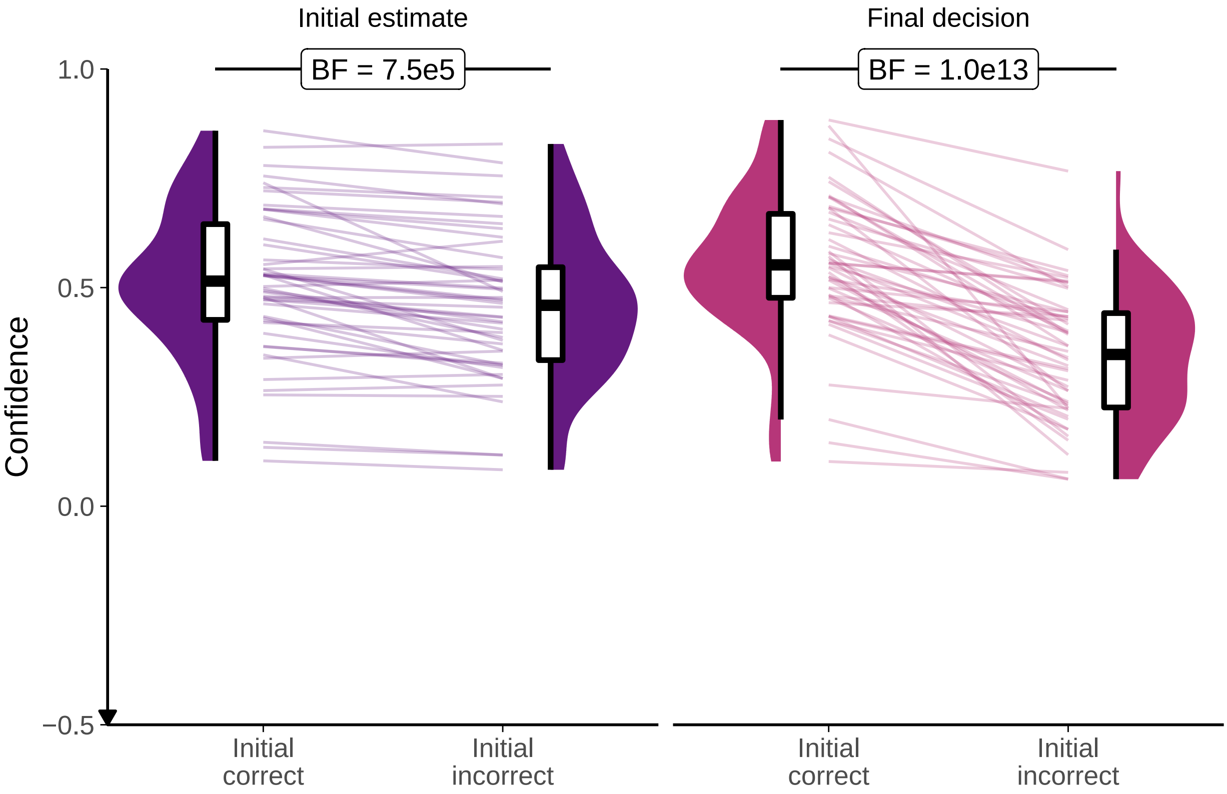


Figure 3.21: Confidence for the Dots task with agreeing versus accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

Task performance Basic behavioural performance was similar to that observed with the same Dots task in Experiments 1A§3.1.1.3 and 2A§3.2.1.3. Initial estimate accuracy converged on the target 71%, and, as shown in Figure 3.20, participants benefited from advice in terms of their final decisions being more accurate than their initial estimates (ANOVA main effect of Time: F(1,49) = 16.63, p < .001; $\mathrm{M}_{\mathrm{Final}}$ = 0.74 [0.72, 0.75], $\mathrm{M}_{\mathrm{Initial}}$ = 0.72 [0.71, 0.73]), especially following advice from the High accuracy advisor (interaction of Time and Advisor: F(1,49) = 33.88, p < .001; $\mathrm{M}_{\mathrm{Improvement|HighAgreement}}$ = -0.01 [-0.02, 0.00], $\mathrm{M}_{\mathrm{Improvement|HighAccuracy}}$ = 0.05 [0.03, 0.06]). There was no main effect of Advisor (F(1,49) = 3.42, p = .070; $\mathrm{M}_{\mathrm{HighAgreement}}$ = 0.72 [0.70, 0.73], $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 0.74 [0.72, 0.75]).

Figure 3.21 and ANOVA indicated that participants were more confident in

Table 3.19: Advisor agreement for Dots task with accurate versus agreeing advisors

Advisor	Target|correct	Actual|correct	Target|incorrect	Actual|incorrect
High accuracy	.800	.804	.200	.208
High agreement	.800	.812	.800	.804

Table 3.20: Advisor accuracy for Dots task with accurate versus agreeing advisors

Advisor	Target accuracy	Mean accuracy
High accuracy	.800	.800
High agreement	.626	.640

their correct answers than their incorrect ones (F(1,49) = 139.32, $p < .001$; $\mathrm{M}_{\mathrm{Correct}}$ = 26.60 [24.28, 28.92], $\mathrm{M}_{\mathrm{Incorrect}}$ = 19.94 [17.81, 22.08]), and less confident in their final decisions than their initial estimates (F(1,49) = 10.64, $p = .002$; $\mathrm{M}_{\mathrm{Final}}$ = 22.36 [20.31, 24.42], $\mathrm{M}_{\mathrm{Initial}}$ = 24.18 [21.80, 26.57]), and that these two factors interacted (F(1,49) = 90.55, $p < .001$; $\mathrm{M}_{\mathrm{Increase|Correct}}$ = 2.05 [0.88, 3.22], $\mathrm{M}_{\mathrm{Increase|Incorrect}}$ = -5.70 [-7.28, -4.12]). There was no evidence of a correlation between initial estimate accuracy and confidence (1/1.69), and no evidence for a correlation between final decision accuracy and confidence (1/2.30).

Advisor performance The advice is generated probabilistically from the rules described previously in Table 3.17. It is thus important to get a sense of the actual advice experienced by the participants.

The advisors agreed with the participants' initial estimates at close to target rates (Table 3.19), and were as accurate on average as expected (Table 3.20). 49/50 participants experienced the intended relationship wherein the High agreement advisor agreed with them more than the High accuracy advisor and the High accuracy advisor gave more accurate advice than the High agreement advisor.

Hypothesis test Despite the influence differences observed above, and counter to our predictions, Figure 3.22 shows that there was no consistent picking preference in favour of either the High accuracy or the High agreement advisor. While several participants did develop very strong preferences, picking one or the other advisor nearly all the time, these preferences were not systematically oriented

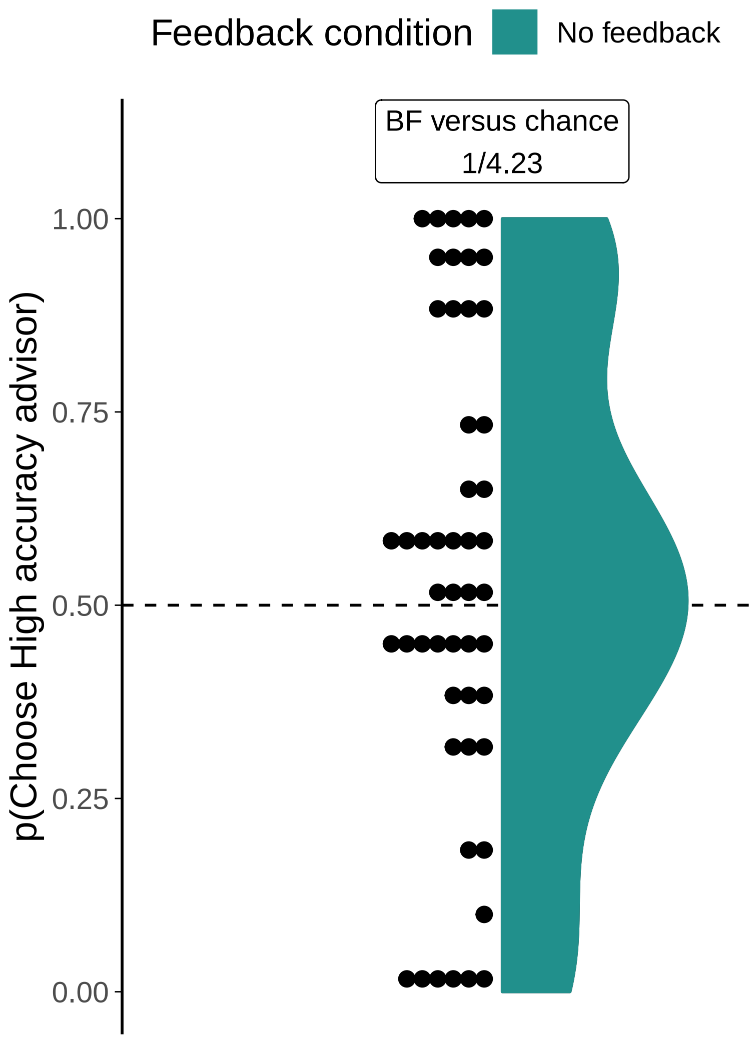


Figure 3.22: Dot task advisor choice for accurate versus agreeing advisors. Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

towards either advisor ($t(49) = 0.95$, $p = .345$, $d = 0.13$, $\text{BF}_{\text{H1:H0}} = 1/4.23$; $M = 0.54$ [0.45, 0.63], $\mu = 0.5$).

For context, note that we did see a significant effect of picking the advisor in the first position on the screen ($t(49) = 2.65$, $p = .011$, $d = 0.37$, $\text{BF}_{\text{H1:H0}} = 3.49$; $\text{M}_{\text{P(PickFirst)}} = 0.54$ [0.51, 0.56], $\mu = 0.5$), an effect that we would hope would be random and even out across participants. In this experiment, as a function of chance, the High accuracy advisor appeared in the favoured top position less frequently than we would expect ($\text{BF}_{\text{H1:H0}} = 3.49$; $\text{M}_{\text{P(HighAccuracyFirst)}} = 0.46$ [0.44, 0.49], $\mu = 0.5$). If there were a general preference for the High agreement advisor, as per our prediction, this would be *increased* by that advisor appearing more often in the top position. Thus, even if the position of the advisors affected the results, it would not be confounding results in the direction of our prediction.

Follow-up tests

Ability of participants Although participants did not have a clear preference for the agreeing over the accurate advisor, it may be the case that participants who were well-calibrated were able to detect the usefulness of the accurate advisor, and therefore tended to prefer to hear that advisor's advice. The evidence was not sufficient to draw firm conclusions, but there was little indication of a correlation between preference for the High accuracy advisor and participant accuracy (r(48) = -.071, $p = .623$, $\text{BF}_{\text{H1:H0}} = 1/2.81$) or confidence calibration (r(48) = .106, $p = .463$, $\text{BF}_{\text{H1:H0}} = 1/2.46$).

Experience of advisors The lack of systematic preference from the participants was surprising. Each participant's data were tested against a null hypothesis that their picking was random, and 52.0% of participants demonstrated a statistically significant preference. As seen in Figure 3.22, however, these preferences were quite evenly split between the High accuracy and the High agreement advisors, both in terms of frequency and strength (although with a slight advantage for the High accuracy advisor).

The wide variation in preferences was not significantly correlated with either experienced agreement (r = .035, p = .810) or experienced accuracy (r = .162, p = .262).

Summary Contrary to expectations, participants did not appear to use advice agreement as a proxy for advice accuracy when objective feedback was not available. Whereas in the previous experiment (2A§3.2.1.3) this did happen, in the current experiment there was a cost to seeking advice from the High agreement advisor: the advice was less accurate.

It is not clear whether some participants were aware of the accuracy difference by some mechanism other than agreement. It is possible that, as suggested by Pescetelli and Yeung (2021), participants were able to use a combination of their subjective confidence and advice agreement to determine advisors' accuracy. This is explored more in Experiment 4A§3.4.2.

The preferences for advisors in the previous Dots task experiments, Experiment 1A§3.1.1.3 and Experiment 2A§3.2.1.3, showed clustering of preferences around the midpoint; participants generally had mild preferences. In contrast, in the present experiment, there was a wide range of preference strengths, but these were distributed evenly between the two advisors. It seems probable that participants in the current experiment were sensitive to differences between the advisors, but that their response to these differences was not uniform.

The next experiment explores the contrasted High agreement and High accuracy advisors in the Dates task.

3.3.2 Experiment 3B: accuracy versus agreement effects in the Dates task

This experiment explored preferences for accuracy versus agreement using the continuous version of the Dates task introduced in Experiment B.1. The continuous task was used because we aimed to investigate both advisor influence and advisor choice as dependent variables. Once again, participants were separated into Feedback and No feedback conditions.

3.3.2.1 Open scholarship practices

This experiment was preregistered at https://osf.io/nwmx5. This is a replication of a study of identical design. The data for this and the original study can be obtained from the esmData R package (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/ed13951c488e1996df7ff53d48629843bacfd074/ACv2/ac.html.

3.3.2.2 Method

49 participants each completed 52 trials over 4 blocks of the continuous version of the Dates task§2.1.3.2. On each trial, participants were presented with an historical event that occurred on a specific year between 1900 and 2000. They were asked to drag one of three markers onto a timeline to indicate the date range within which they thought the event occurred. The three markers each had different widths, and each marker had point value associated with it, with wider markers worth fewer points. The markers were 7, 13, and 21 years wide, being worth 25, 10, and 5 points respectively. Participants then received advice indicating a region of the timeline in which the advisor suggested the event occurred. Participants could then mark a final response in the same manner as their original response, and could choose a different marker width if they wished.

Participants started with 1 block of 10 trials that contained no advice to allow them to familiarise themselves with the task. All trials in this section included feedback for all participants indicating whether or not the participant's response was correct.

Participants then did 2 trials with a practice advisor to get used to receiving advice. They also received feedback on these trials. They were informed that they would "get advice on the answers you give" and that the feedback they received would "tell you about how well the advisor does, as well as how well you do". Before starting the main experiment they were told that they would receive advice from multiple advisors and that "advisors might behave in different ways, and it's up to you to decide how useful you think each advisor is, and to use their advice accordingly".

Participants then performed 3 blocks of trials that constituted the main experiment. The first two of these were Familiarisation blocks where participants had a single advisor in each block for 14 trials, plus 1 attention check.

Participants were split into four conditions that produced differences in their experience of these Familiarisation blocks. These conditions were whether or not they received feedback, and which of the two advisors they were familiarised with first. For each advisor, participants saw the advisor's advice on 14 trials. On most of these trials, including the first two, the advisor gave advice according to the advice profile (detailed below). On between 2 and 3 trials, the advisors issued the same kind of advice as one another, chosen to neither agree with the participant's answer nor indicate the correct answer. This "Off-brand" advice was used to control for the effects of advice when the influence of advice was the dependent variable.

Finally, participants performed a Test block of 10 trials that offered them a choice on each trial of which of the two advisors they had encountered over the last two blocks would give them advice. No participants received feedback during the test phase, and all advisors gave on-brand advice according to their advice profile.

Advice profiles The High accuracy and High agreement advisor profiles defined marker placements based on the timeline based on the correct answer and the participant's initial estimate respectively. Both advisors used markers that spanned 7 years, and both placed the markers in a normal distribution around the target point with a standard deviation of 5 years. The target point for the High accuracy advisor was the correct answer, and the target point for the High agreement advisor was the participant's initial estimate. Neither advisor ever placed their marker exactly on the midpoint of the participant's marker (because doing so means the Weight on Advice statistic is undefined).

Table 3.21: Participant exclusions for Dates task with accurate versus agreeing advisors

Reason	Participants excluded
Too few trials	0
Insufficient advice-taking	0
Too few choice trials	0
Wrong markers	2
Non-numeric advice	0
Total excluded	**2**
Total remaining	**33**

Note that just as the task is continuous rather than binary, so agreement is continuous rather than binary. There is no objective threshold at which to classify advice as 'agreement', although we can classify accuracy in a binary way as whether or not a marker includes the correct answer.

On Off-brand advice trials, of which there were between 2 and 3 per Familiarisation block, advisors neither indicated the correct answer nor agreed with the participant. This was achieved by picking a target point of the participant's answer reflected around the correct answer. A detailed example was given in Appendix B.1.0.2.

3.3.2.3 Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. Participants were then excluded for having fewer than 11 trials remaining, fewer than 8 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. Participants were also excluded for technical problems with the experiment and data: sometimes the widths of the markers placed by the participants had unrecognised values, and sometimes the values for the advisors' advice were corrupted. Overall, 2 participants were excluded, with the details shown in Table 3.21.

Task performance In this section, task performance is explored during the Familiarisation phase of the experiment where participants received advice from

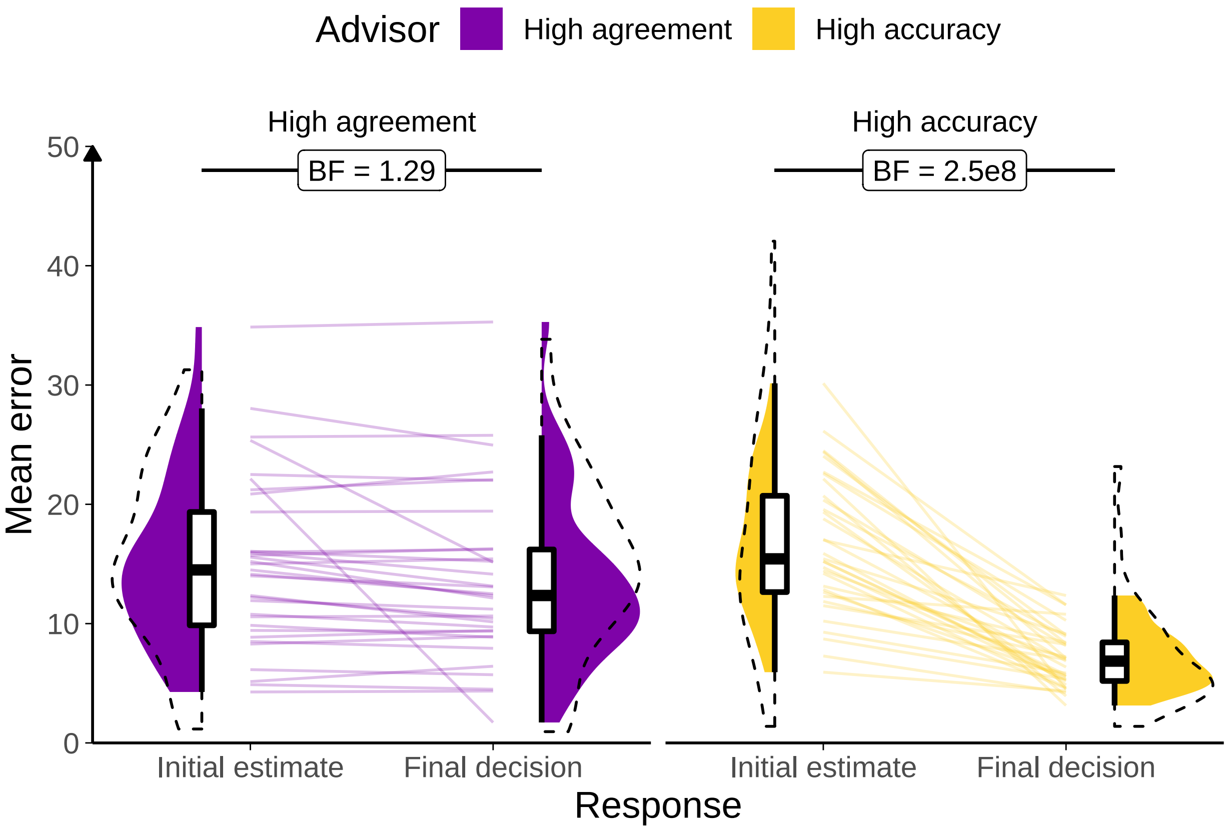


Figure 3.23: Response error for the Dates task with accurate versus agreeing advisors. Faint lines show individual participant mean error (the absolute difference between the participant's response and the correct answer), for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of participant means on the original study which this is a replication. The dependent variable here is error, the distance between the correct answer and the participant's answer, and consequently lower values represent better performance. The theoretical limit for error is around 100.

a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Participants generally improved their response accuracy following advice; they had lower error on final decisions than on their initial estimates (F(1,32) = 69.02, p < .001; $\text{M}_{\text{Initial}}$ = 15.84 [13.80, 17.89], M_{Final} = 10.28 [8.87, 11.69]). They also had lower error on their answers with the High accuracy advisor (F(1,32) = 5.73, p = .023; $\text{M}_{\text{HighAgreement}}$ = 14.28 [11.86, 16.71], $\text{M}_{\text{HighAccuracy}}$ = 11.84 [10.61, 13.07]). As expected, there was an interaction: participants reduced their error much more following advice from the High accuracy advisor (F(1,32) = 60.26,

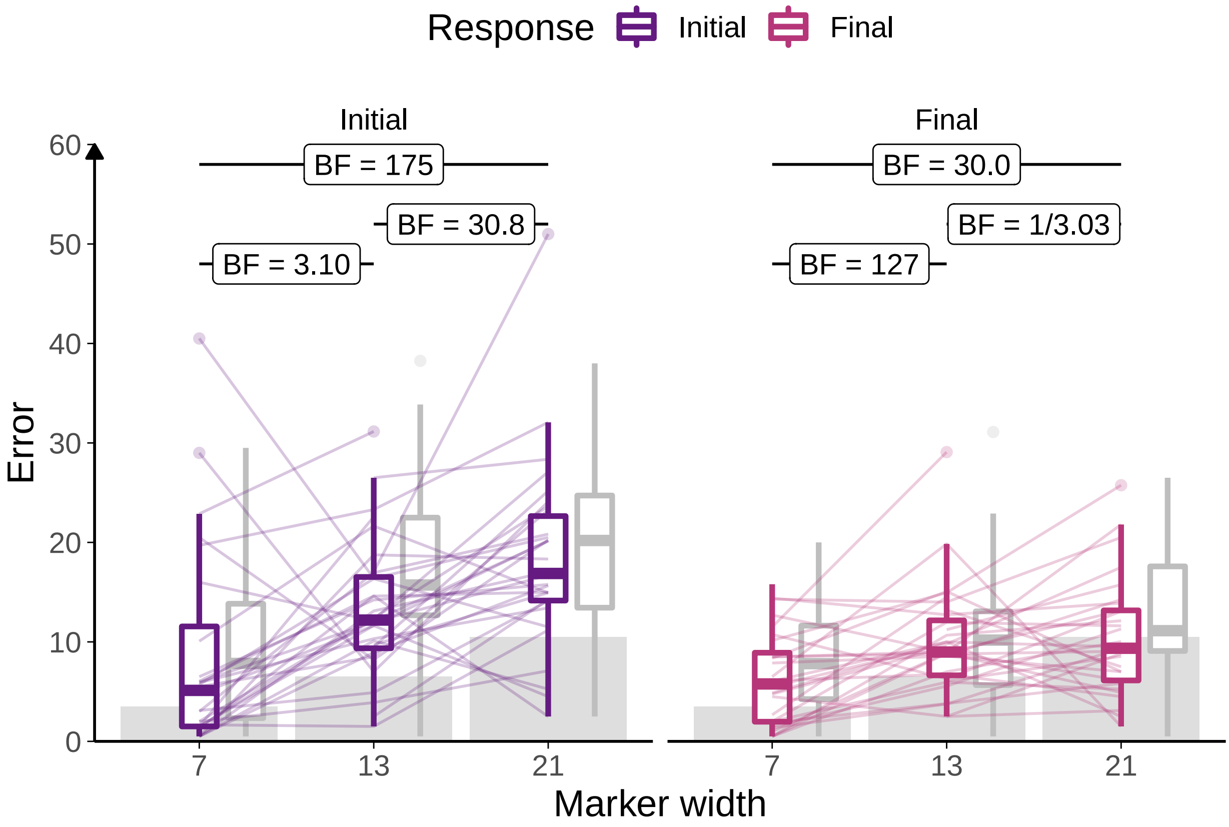


Figure 3.24: Error by marker width for the Dates task with accurate versus agreeing advisors.
Faint lines show individual participant mean error (distance from the centre of the participant's marker to the correct answer) for each width of marker used, and box plots show the distributions. Some participants did not use all markers, and thus not all lines connect to each point on the horizontal axis. The dashed box plots show the distributions of participant means in the original experiment of which this is a replication. The faint black points indicate outliers. Grey bars show half of the marker width: mean error scores within this range mean the marker covers the correct answer.

p < .001; $M_{\text{Reduction}|\text{HighAgreement}} = 1.46$ [0.05, 2.87], $M_{\text{Reduction}|\text{HighAccuracy}} = 9.67$ [7.66, 11.68]; Figure 3.23).

Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Confidence can be measured by the width of the marker selected by the participant. Where participants are more confident in their response, they can maximise the points they receive by selecting a thinner marker. Where participants are unsure, they can maximise their chance of getting the answer correct by selecting a wider marker. Participants' error was lower for each marker width in final decisions than initial estimates (Figure 3.24). For both initial estimates and final decisions, error was higher for wider markers than for narrower ones.

Advisor performance The advice is generated probabilistically so it is important to check that the advice experienced by the participants matched the experience we designed. On average, the High accuracy advisor had lower error than the High agreement advisor ($t(32)$ = -8.75, p < .001, d = 2.22, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 2.0e7; $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 6.02 [5.55, 6.48], $\mathrm{M}_{\mathrm{HighAgreement}}$ = 15.30 [13.25, 17.35]), and their advice was further away from the participants' initial estimates than the High accuracy advisor's ($t(32)$ = 12.47, p < .001, d = 2.24, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 9.0e10; $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 20.28 [18.21, 22.35], $\mathrm{M}_{\mathrm{HighAgreement}}$ = 9.59 [8.39, 10.80]). 32/33 (96.97%) participants experienced the High accuracy advisor as having lower average error than the High agreement advisor, and 33/33 (100.00%) participants experienced the High agreement advisor as offering advice closer to their initial estimates than the High accuracy advisor. Overall, this indicates that the manipulation was implemented as planned.

Hypothesis test Consistent with the result from the Dots task (Experiment 3A§3.3.1), in the No feedback condition participants' preferences for receiving advice from the High accuracy advisor were not different from chance ($t(13)$ = -0.16, p = .879, d = 0.04, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/3.66; $\mathrm{M}_{\mathrm{NoFeedback}}$ = 0.49 [0.29, 0.68], μ = 0.5) on average, and varied widely across individual participants: participant preferences in the No feedback condition were almost perfectly evenly distributed, both in terms of which advisor was preferred and the strength of that preference, in both the original study and the replication (Figure 3.25).

In the Feedback condition, in contrast, the mean of the participants' selection rates clearly favoured the High accuracy advisor ($t(18)$ = 5.00, p < .001, d = 1.15, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 297; $\mathrm{M}_{\mathrm{Feedback}}$ = 0.81 [0.68, 0.94], μ = 0.5). This is consistent with a strategy which attempts to maximise the accuracy of final decisions. This qualitative difference from the No feedback condition also translated into a statistical

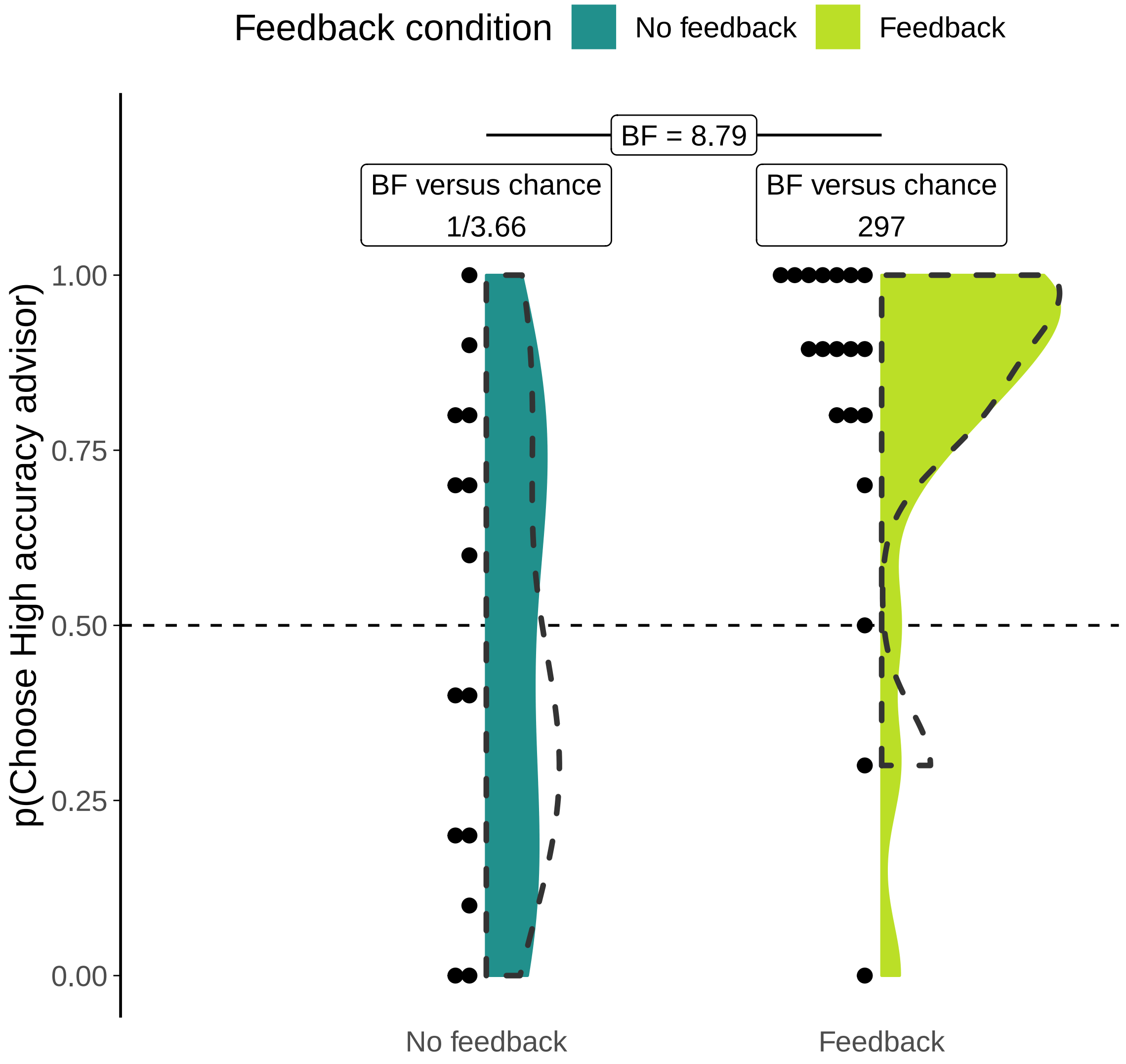


Figure 3.25: Dates task advisor choice for accurate versus agreeing advisors. Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarisation phase, but not during the Choice phase. The dotted outline indicates the distribution of participant means in the original experiment of which this experiment is a replication.

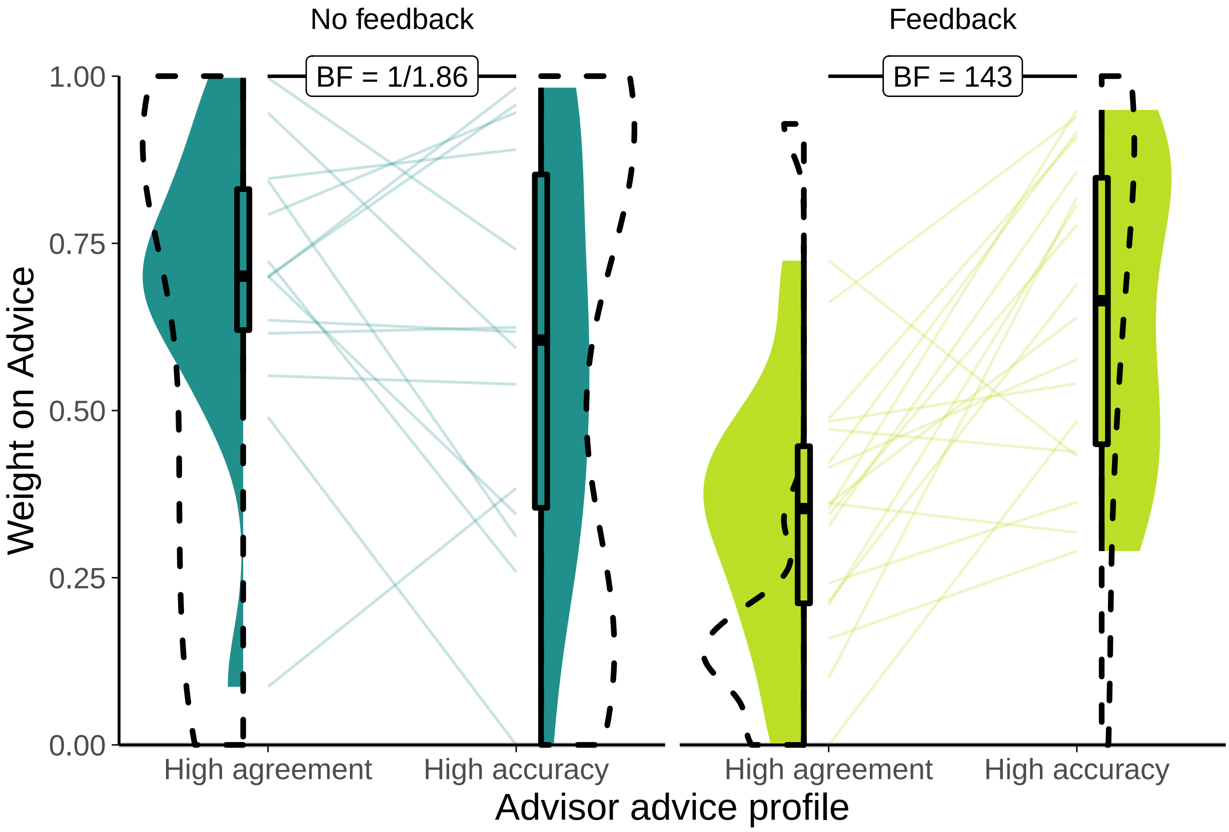


Figure 3.26: Date task advisor WoA for accurate versus agreeing advisors. Participants' weight on the advice for advisors in the Familiarisation phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The dotted outline indicates the distribution of participant means in the original experiment of which this experiment is a replication.

difference: the two preference distributions were clearly different from one another ($t(23.95) = 2.92$, $p = .007$, $d = 1.07$, $\mathrm{BF}_{\mathrm{H1:H0}} = 8.79$; $\mathrm{M}_{\mathrm{Feedback}} = 0.81$ [0.68, 0.94], $\mathrm{M}_{\mathrm{NoFeedback}} = 0.49$ [0.29, 0.68]).

Advisor influence We included in our design a subset of trials on which advisors offered the same kind of advice.[6] This meant that we could investigate the influence of the advisors while controlling for differences in their advice. Examining the influence of these Off-brand trials indicated that the High accuracy advisor was more influential than the High agreement advisor (F(1,30) = 6.08, p = .020;

[6] 1 participant was dropped from this analysis because their data did not have any remaining Off-brand trials for one advisor.

$M_{\text{HighAccuracy}} = 0.62$ [0.53, 0.72], $M_{\text{HighAgreement}} = 0.50$ [0.40, 0.59]). This difference was more pronounced in the Feedback condition (F(1,30) = 15.82, $p < .001$; $M_{\text{Accuracy-Agreement|Feedback}} = 0.30$ [0.17, 0.44], $M_{\text{Accuracy-Agreement|NoFeedback}} = -0.10$ [-0.28, 0.07]). There was no evidence of a main effect of condition, however (F(1,30) = 4.12, $p = .051$; $M_{\text{Feedback}} = 0.50$ [0.42, 0.58], $M_{\text{NoFeedback}} = 0.64$ [0.51, 0.76]).

3.3.3 Discussion

Whereas previous experiments assessed the separate effects of advisor accuracy (1A§3.1.1 and 1B§3.1.2) and agreement (2A§3.2.1 and 2B§3.2.2), in these two experiments these factors were set in opposition. The results were clear when feedback was provided (in the Dates task), indicating that people are capable of attending to challenging but useful information provided they have a chance to learn that the information is actually useful. In contrast, when people are not given objective feedback against which to evaluate the advice they receive, there does not seem to be a systematic response: some people seek agreement while others seek alternate perspectives, and the extent to which each strategy is pursued to the exclusion of the other is also highly variable. It is an open question whether a person's strategy choice in the absence of useful cues as to the utility of the information they receive is due to random selection or related in a meaningful way to their personality or cognitive style.

The distributions of preferences were compatible with but slightly different to those seen in the previous experiments. In neither Experiment 1A§3.1.1 nor 2A§3.2.1 did we see a null effect of choice, so it is not clear whether that would look like random picking or, as it does here, a similar range of preference directions and strengths to those seen in the Dates task. The Dots task distribution seen here might be expected, given no systematic preferences, to cluster around the middle, representing the drives for fairness, novelty, etc. as discussed in previous chapters. The Dates task distribution is more in keeping with previous experiments: the expected range of preference directions and strengths is there in the absence of a systematic preference. In the Feedback condition where systematic differences are seen, we still see the entire range of preference directions and strengths, as we did before, although the clustering towards the High accuracy advisor is much more pronounced than in previous experiments. This clustering of the distribution suggests that the manipulation worked more effectively in this experiment, perhaps because of the use of the continuous Dates task: participants may have experienced the visual depiction of a distance between their estimate and the advice as a stronger signal than binary agreement or disagreement.

The results of the advisor choice behaviour in this experiment conceptually replicate the advisor influence results in Experiment B, and the influence results in the Dates task similarly replicate those findings. While this study was not set up to examine influence rigorously, the inclusion of trials where the advisors offered equivalent advice allowed us to explore advisor influence without the confound of differences in the nature of the advice itself. We found, in two preregistered studies, that participants in the Feedback condition discriminated between their advisors, being more influenced by the accurate rather than agreeing advice. Contrary to our hypothesis, however, we found that participants in the No feedback condition either did not discriminate (original) or did not provide evidence of discrimination (replication).

The similarity of these results to those of the previous study is encouraging, although once again the methodology does not entirely equate the advice between the advisors. Although advice on the Off-brand trials is constructed using the same rules for each advisor, Off-brand advice from the High agreement advisor will be much more surprising than Off-brand advice from the High accuracy advisor, because participants will usually have experienced the High accuracy advisor offering advice that deviates dramatically from their initial estimates whereas they will not have experienced this from the High agreement advisor. Assuming that participants in the No feedback condition have no better insight into the correct answer than they offer with their estimate, advice from the High agreement advisor may be discounted simply because it deviates more from the perceived correct answer than usual. Participants in the No feedback condition may also feel reassured by the High agreement advisor that they are good at the task, and thus reduce their reliance on advice. It is possible, in sum, for participants in the No feedback condition to suffer suppression of influence on Off-brand trials specifically for the High agreement advisor, which would lead to genuine influence effects being difficult to detect.

Another consideration is that the Off-brand advice trials that the influence analysis is based on occur during the Familiarisation phase when participants are learning about the advisors. Participants in the Feedback condition appear to have learned rapidly that the advice of the High accuracy advisor is worth following, and the advice of the High agreement advisor is not informative. There is a slight suggestion from the individual participant data that many of the participants in the No feedback condition may have been creeping towards this conclusion, but the statistics are uninformative on the question. It is highly plausible that learning with feedback is far more rapid than learning without feedback, especially in noisy and heterogeneous tasks.

Whereas the previous experiments demonstrated that participants denied feedback would prefer to pick the High agreement advisor, that did not happen in this experiment. Together with the above, this might indicate that even without feedback participants were sensitive to the accuracy of advice over and above agreement. The Pescetelli and Yeung (2021) theory of metacognitive advice evaluation supports this because confidence is used to assess advice plausibility. Where people receive agreeing advice on questions that they feel confident about, they will consider the advice as highly likely to be accurate, whereas when they themselves are very unsure of the answer they will not learn very much about the advisor from the advice whether it agrees or not.

It is this confidence-based assessment of advice that we attempted to test in the next experiments using advisors whose advice was generated depending on the participant's confidence in their initial estimate.

3.4 Effects of confidence-contingent advice

The results of the experiments described previously (1A§3.1.1, 1B§3.1.2, 2A§3.2.1, 2B§3.2.2, 3A§3.3.1, 3B§3.3.2) were mixed, but indicated overall (and in combination with Pescetelli and Yeung 2021) that people prefer accurate over inaccurate advisors when they can assess the advisors based on feedback, but prefer agreeing over disagreeing advisors when feedback is absent. Here we test whether advice selection can depend on a more sophisticated mechanism when feedback is absent.

This mechanism, proposed by Pescetelli and Yeung (2021), weights the updating of trust in an agreeing advisor using the confidence of the judge's initial estimate. Where the initial estimate is made with high confidence, agreement is taken to indicate that the advisor has a high probability of being correct, while disagreement is taken to indicate that the advisor has a low probability of being correct. Where the initial estimate is low confidence, however, neither agreement nor disagreement is a good indicator of correctness. This intuition accords neatly with a Bayesian approach: the greater the uncertainty around whether or not the advice is correct, the lower the strength of the updating of trust in the advisor.

This mechanism can be directly tested by using advisors who have different agreement rates contingent upon the judge's confidence in their initial estimate, and Pescetelli and Yeung (2021) used just such an approach to provide evidence for the mechanism in the domain of advisor influence. Here, a very similar design is used to explore this effect in the domain of advisor choice.

Unlike the previous experiments, in which the same approach was implemented in both the Dots and the Dates tasks, this experiment uses only the Dots task. To balance the advisors' accuracy and agreement rates (such that only their *confidence-contingent agreement* is varied), precise control is required over the participants' initial estimate accuracy. Such control is achieved using a staircasing procedure in the Dots task, and cannot be done in the Dates task (because there are too few questions to choose from, the questions have too wide a range of difficulty, and a question's difficulty varies dramatically and unpredictably between participants).

Instead of a Dots and a Dates task, this experiment is repeated in two versions of the Dots task. The first version, directly adapted from Pescetelli and Yeung (2021), was performed in the lab with a correspondingly low sample size and high number of trials per participant. The second version was a replication run on-line and with a larger sample size and lower number of trials per participant. For ease of reference, the lab study is referred to as 'Lab study', while the on-line study retains the numbering and lettering used previously, meaning it is designated Experiment 4A. There is no Experiment 4B, because it was not possible to use the Dates task with this design of advisors.[7]

3.4.1 Lab study of confidence-contingent advice

Pescetelli and Yeung (2021) used a Judge-Advisor System to demonstrate that judges are influenced to a greater extent by advisors who share their biases. Participants played the role of judge in a Judge-Advisor System, while the advisors were virtual agents whose advice-giving was dependent upon the confidence and correctness of the judges' initial estimates. The advisors were balanced for both overall agreement with the judge and objective correctness of advice. This was achieved by varying the agreement rates for the advisors contingent on the confidence of the judge's initial estimate. The **Bias-sharing** advisor would agree more frequently when the participant expressed high confidence in their initial estimate and less frequently when the participant expressed low confidence in their initial estimate. The **Anti-bias** advisor did the opposite. Crucially, the advisors were matched in agreement when participants were moderately confident in their initial estimates, allowing the advisors to be compared directly on these trials (Pescetelli and Yeung 2021). We place participants in a similar paradigm in which they are given a choice between advisors, and hypothesise that they will more frequently seek advice from the Bias Sharing advisor than from the Anti Bias advisor. We predict that, given a choice,

[7] The lab version of this study was the first one performed in chronological order. The results were not as anticipated, so a replication was run on-line (Experiment 4A). These results were also odd, so other experiments were run to investigate the stability of the basic effects (Experiments 1A, 2A, and 3A).

judges will prefer to receive advice from a Bias Sharing advisor over receiving advice from an advisor who does not share the judge's bias.

3.4.1.1 Open scholarship practices

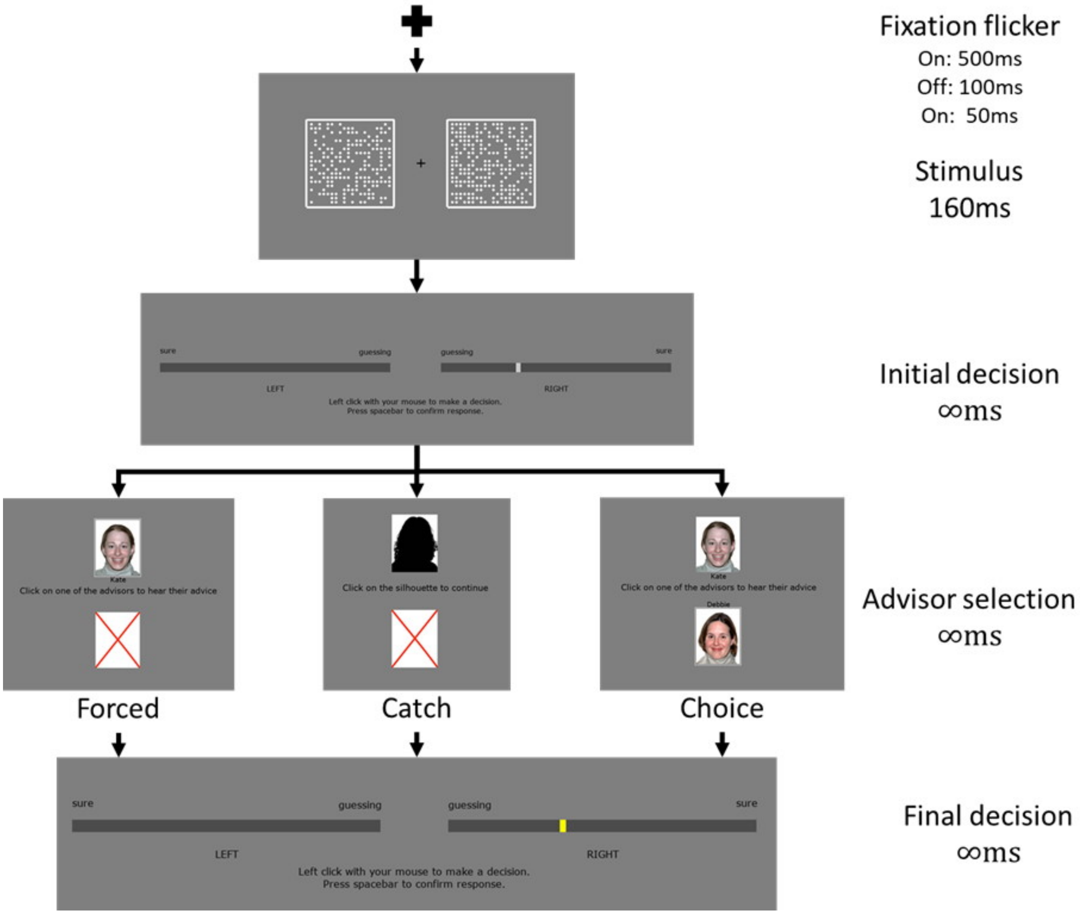
This experiment was preregistered at `https://aspredicted.org/ze3tn.pdf`. One analysis in the preregistration is not reported here because the results are non-significant and they represented a branch of analysis that we sidelined in other experiments. This analysis explored participants' subjective assessments of advisors. The experiment data are available in the `esmData` package for R (Jaquiery 2021c), and also directly from `https://osf.io/vgcnb/`. The code for running the experiment can be obtained from `https://github.com/mjaquiery/nofeedback_trust`.

3.4.1.2 Method

The method for the lab study was different in many small ways from the general method used for the Dots task on-line. The basic trial experience was the same: participants saw two boxes of dots flashed briefly on the screen, and were asked to indicate which box had more dots, along with how confident they were in this initial estimate. Participants then received advice and provided a final decision. On some trials participants were able to choose which of their advisors would provide the advice (Figure 3.27).

Overall, 26 participants who were recruited from the University of Oxford participant recruitment platforms took part in the experiment and attended experimental sessions. 1 participant was excluded because the preregistered sample size had already been met when their data were collected, and one participant's data was lost due to technical issues. The remaining participants had an average age of 21.75 (±SD 4.7). Their self-identified genders were given as 5 male, 19 female, 0 other. Participants were compensated for their time with either course credit for a psychology degree, or 10GBP.

Each participant completed 363 trials (51 practice trials over 2 blocks + 12 x 26-trial experimental blocks). Prior to the first experimental block, after the final



Left click with your mouse to make a decision.
Press spacebar to confirm response.

Click on one of the advisors to hear their advice

Click on the silhouette to continue

Click on one of the advisors to hear their advice

Left click with your mouse to make a decision.
Press spacebar to confirm response.

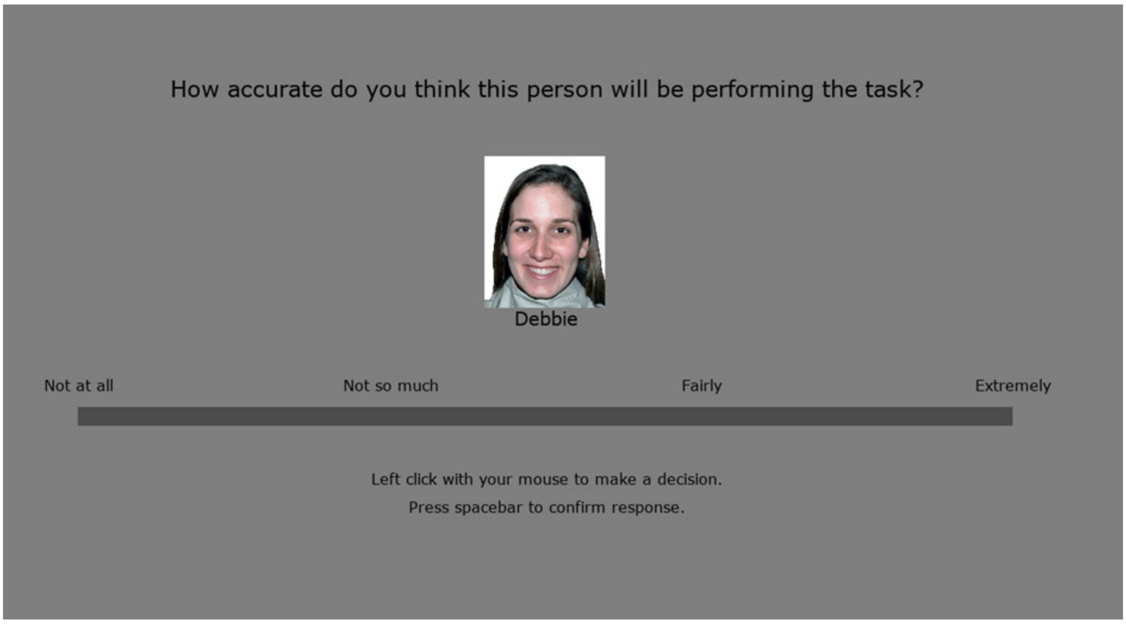
Figure 3.27: Experiment 1 procedure.
The task began with a blank grey screen containing only a fixation cross and progress bar. Momentarily prior to the onset of the stimuli the fixation cross flickered. The stimuli, two rectangles containing approximately 200 dots each, appeared for 0.16s, one on either side of the fixation cross. Once the stimuli disappeared, a response-collection screen appeared and prompted the participant to indicate their initial estimate and its confidence by selecting a point within one of two regions. 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.

experimental block, and after the 4th and 8th experimental blocks, participants were presented with a questionnaire (Figure 3.28). 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 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').

The blocks contained a mixture of choice, forced, and catch trials. On choice trials, participants had a choice between the two advisors giving advice on that block, and were able to select whichever they preferred by clicking the advisor's portrait. The relative frequency of selection on these trials provided our dependent variable of advisor choice. On forced trials, participants were faced with the advisor choice screen, but only one option was available. Participants could only continue when they selected the available portrait. These trials were included to allow measuring of the influence of advice without the confound of having just chosen the advisor over another advisor. Catch trials were included to encourage participants to respond accurately in their initial estimates: on these trials they were forced to select a blank advisor portrait and received no advice; their initial estimate became their final decision automatically.

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, Brainard, and Pelli 2007).



How accurate do you think this person will be performing the task?

Left click with your mouse to make a decision.
Press spacebar to confirm response.

Figure 3.28: Experiment 1 advisor questionnaire.
Participants rated advisors on a number of different dimensions.

Key differences from on-line version There were several differences from the on-line version of the task that are worth mentioning. First, there were more trials and the experiment took longer to complete. Second, when the participant was forced to have advice from one or other advisor, they were presented with the advisor choice screen and one of the options was a blank advisor that they were unable to select. This meant that advisor influence on the forced trials could be analysed as compared to choice trials. Third, 8% of trials were catch trials on which no advice was offered. Fourth, the questionnaires were more detailed and more numerous, and contained no free text fields. Fifth, participants had the opportunity to discuss the experiment with the experimenter after the experiment was complete, although none made use of this opportunity.

Advice profiles The two advisor profiles used in the experiment were Bias sharing and Anti-bias. The advisors are balanced for their overall accuracy and agreement rates, but the Bias sharing advisor agrees more frequently with participants when their initial estimate is correct and made with relatively high confidence. The Anti-bias advisor agrees more frequently with participants when their initial estimate

Table 3.22: Advisor advice profiles for Lab experiment

	Agreement						
	Initial estimate confidence[a]			... correctness			
Advisor	Low[b]	Medium[c]	High[b]	Correct[d]	Incorrect	Overall[e]	Accuracy[f]
Bias-sharing	60	70	80	70	30	**58.4**	70
Anti-bias	80	70	60	70	30	**58.4**	70

[a] Only correct trials received confidence-contingent advice
[b] 30% of trials
[c] 40% of trials
[d] Average correct agreement is the weigthed average of the previous three columns
[e] Overall agreement is p(correct) * p(agree|correct) + p(¬correct) * p(agree|¬correct)
[f] Overall accuracy is p(correct) * p(agree|correct) + p(¬correct) * p(¬agree|¬correct)

Table 3.23: Advisor agreement for Lab experiment

	Agreement						
	Initial estimate confidence			... correctness			
Advisor	Low	Medium	High	Correct	Incorrect	Overall	Accuracy
Anti-bias	78.7	70.3	63.6	71.2	31.6	59.4	70.5
Anti-bias (Target)	*80*	*70*	*60*	*70*	*30*	*58.4*	*70*
Bias-sharing	59.7	70.3	79.4	69.3	30.4	57.0	69.4
Bias-sharing (Target)	*60*	*70*	*80*	*70*	*30*	*58.4*	*70*

is correct and made with relatively low confidence (Table 3.22). Note that the overall correctness and agreement rates of the advisors are equivalent. Importantly, on a large minority of trials, the middle 40%, the advisors are exactly equivalent, meaning these trials can be compared directly without confounds arising from agreement rates and initial confidence.

3.4.1.3 Results

Exclusions Participants could be excluded for having initial estimate accuracy below 60% or above 90%, or for performing the experiment after the stated sample size had been reached. No participants were excluded for having accuracy out of range. 1 participant was excluded for being extraneous.

Advisor performance The advisors agreed with the participants' initial estimates at close to target rates in all confidence categories, and were as accurate on average as expected (Table 3.23).

Advisor choice 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 ($t(23) = 1.35$, $p = .190$, $d = 0.28$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.08$; $\mathrm{M}_{\mathrm{P(BiasSharing)}} = 0.55$ [0.47, 0.62], $\mu = 0.5$; Figure 3.29), although the Bayesian test indicated that the data were not sufficient to conclude that no effect was present. There was considerable variability across participants in the overall pick rate for the Bias Sharing advisor (range = [.10, .88]).

Unlike in other experiments, there was no systematic effect of the position of the advisors on the screen ($t(23) = 0.10$, $p = .920$, $d = 0.02$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/4.63$; $\mathrm{M}_{\mathrm{P(PickFirst)}} = 0.50$ [0.45, 0.55], $\mu = 0.5$). This placement is random, and has no relationship to the identity of the advisor ($\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.43$; $\mathrm{M}_{\mathrm{P(BiasSharingFirst)}} = 0.51$ [0.49, 0.52], $\mu = 0.5$), so we expect that it would be random across participants.

Advisor choice on medium-confidence trials The advisors differed in their advice-giving as a function of the judge's initial confidence. In trials where the judge's initial estimate was made with medium confidence, however, the advisors were equal on judge confidence and agreement rate. Comparing selection rates for these trials alone revealed a clear preference for the Bias Sharing advisor ($t(23) = 2.45$, $p = .022$, $d = 0.50$, $\mathrm{BF}_{\mathrm{H1:H0}} = 2.49$; $\mathrm{M}_{\mathrm{P(BiasSharing)}} = 0.58$ [0.51, 0.64], $\mu = 0.5$; Figure 3.29 "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 and Yeung 2021). Influence was examined with a 2x2x2 (Bias-sharing versus Anti-bias advisor; choice versus forced trials;

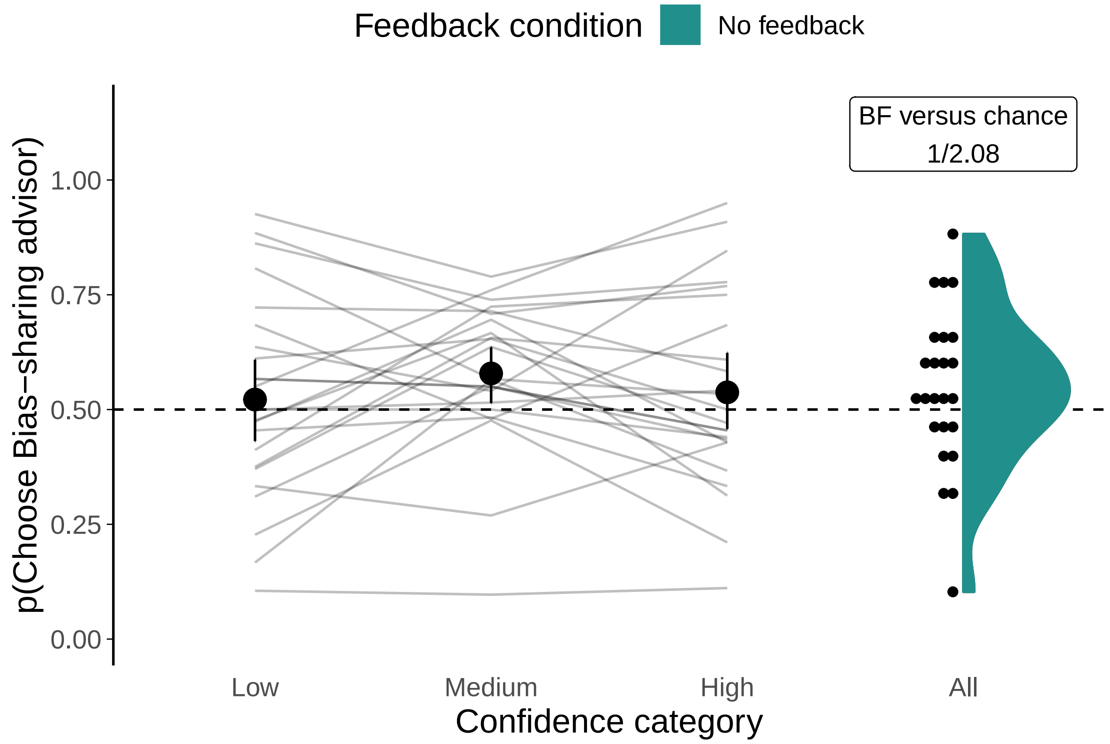


Figure 3.29: Advisor choice for Lab experiment.
Proportion of the time each participant picked the Bias Sharing advisor. Faint lines and dots indicate data from individual participants, while the large dot indicates the mean proportion across all participants. The dashed reference line indicates picking both advisors equally, as would be expected by chance. Error bars give 95% confidence intervals.

Table 3.24: ANOVA of Advisor influence for Lab experiment

Effect	$F(1, 23)$	p		η^2
Advisor	0.28	.602		.001
Trial type	4.23	.051		.001
Advisor agreement	13.88	.001	*	.175
Advisor:Trial type	0.04	.842		.000
Advisor:Advisor agreement	0.75	.395		.001
Trial type:Advisor agreement	1.99	.172		.000
Advisor:Trial type:Advisor agreement	0.01	.935		.000

Degrees of freedom: 1, 23

agreement versus disagreement trials) ANOVA (Figure 3.30). This was chronologically the first experiment we ran, and the analysis was preregistered prior to the development of the capped influence measure designed to allow agreement and disagreement to be compared more fairly. No main effect was found for advisor (F(1,23) = 0.28, p = .602, $M_{\text{Bias-sharing}}$ = 0.09 [0.06, 0.11], $M_{\text{Anti-bias}}$ = 0.08 [0.06, 0.11]), meaning that the previous finding was not replicated. As shown in Table 3.24, the only statistically significant effect was the main effect of agreement, with disagreement producing higher influence than agreement (F(1,23) = 0.04, p = .842, M_{Agree} = 0.06 [0.03, 0.08], M_{Disagree} = 0.13 [0.09, 0.17]).

When this analysis was repeated with the capped influence values (General Method - Capped influence§2.2.1.3) the effect of agreement was still significant, but less pronounced (F(1,23) = 1.28, p = .269, $M_{\text{Bias-sharing}}$ = 0.08 [0.06, 0.10], $M_{\text{Anti-bias}}$ = 0.08 [0.05, 0.10]), and a main effect of trial type also emerged (raw influence: F(1,23) = 4.23, p = .051, M_{Choice} = 0.09 [0.06, 0.12], M_{Force} = 0.08 [0.06, 0.11]; capped influence: F(1,23) = 5.38, p = .030, M_{Choice} = 0.08 [0.06, 0.11], M_{Force} = 0.08 [0.06, 0.10]).

✓ **Advisor influence on medium confidence trials** The agree-in-confidence and agree-in-uncertainty advisors 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. In a deviation from preregistration, this analysis was constrained to only forced trials because some participants had missing data for some advisor-trial type-agreement contingencies in the Medium confidence trials.

The results were qualitatively identical to those for the trials at all confidence levels: a main effect of agreement (raw influence: F(1,23) = 13.60, p = .001; M_{Disagree} = 0.12 [0.09, 0.16], M_{Agree} = 0.05 [0.02, 0.08]; capped influence: F(1,23) = 8.70, p = .007; M_{Disagree} = 0.10 [0.07, 0.13], M_{Agree} = 0.05 [0.02, 0.08]); and no significant effect of advisor (raw influence: F(1,23) = 0.10, p = .759; $M_{\text{Anti-Bias}}$

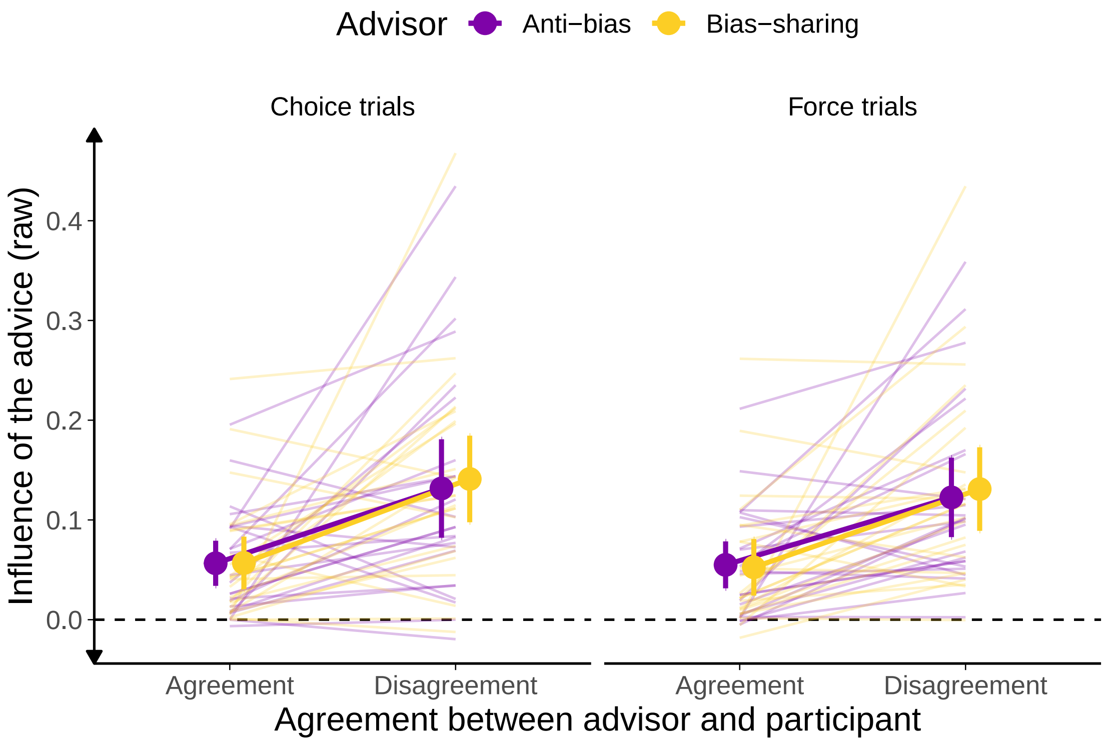


Figure 3.30: Advisor influence for Lab experiment.
Influence of advice from each advisor by advisor, agreement, and trial type. Faint lines and indicate data from individual participants, while the dots indicate the mean proportion across all participants. Error bars give 95% confidence intervals.
Note: vertical axis is truncated to show group differences more clearly, the theoretical maximum influence given the scale is 110. The minimum is -110 as shown.

= 0.08 [0.06, 0.11], $M_{\text{Bias-Sharing}}$ = 0.09 [0.06, 0.12]; capped influence: F(1,23) = 0.00, p = .989; $M_{\text{Anti-Bias}}$ = 0.08 [0.05, 0.10], $M_{\text{Bias-Sharing}}$ = 0.08 [0.05, 0.10]) or interaction (raw influence: F(1,23) = 1.36, p = .255; $M_{\text{Disagree-Agree|Anti-Bias}}$ = 0.05 [0.00, 0.10], $M_{\text{Disagree-Agree|Bias-Sharing}}$ = 0.09 [0.04, 0.14]; capped influence: F(1,23) = 1.91, p = .180; $M_{\text{Disagree-Agree|Anti-Bias}}$ = 0.04 [0.00, 0.08], $M_{\text{Disagree-Agree|Bias-Sharing}}$ = 0.06 [0.03, 0.10]).

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 advisors, 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 high- versus low-confidence trials. Pick rates did not differ ($t(23) = 0.46$, $p = .650$, $d = 0.07$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/4.23$; $\mathrm{M}_{\mathrm{HighConfidence}} = 0.54$ [0.45, 0.63], $\mathrm{M}_{\mathrm{LowConfidence}} = 0.52$ [0.43, 0.61]).

Follow-up tests Given the weak effects of advisor profile differences on advisor choice, we ran two follow-up analyses on pick rates. A first analysis showed that the participants' experiences of advisor agreement in the first block was correlated with the pick rate in later blocks. This suggested that initial exposure to the advisors may have overshadowed information in subsequent blocks. A second analysis showed that participants' answers on the questionnaire measure correlated with their pick preference strength for the questions on asking about how accurate (r(22) = .491 [.109, .747], $p = .015$), and trustworthy (r(22) = .446 [.052, .720], $p = .029$) advisors were. The higher a participant rated one advisor as compared to the other on the questionnaire scale, the heavier that participant's preference tended to be for picking the higher rated advisor. The same was not true for the questions asking how likeable (r(22) = .036 [-.373, .433], $p = .869$) and influential (r(22) = .345 [-.068, .657], $p = .099$) advisors were.

3.4.1.4 Discussion

The lab experiment, chronologically the first of all the experiments,[8] produced equivocal results. While many of the expected relationships were found between participants' subjective perceptions of the advisors and their behaviour towards them, these did not reliably translate into differential picking rates for the Bias-sharing and Anti-bias advisors. There was a significant difference in pick rates in medium confidence trials (where the advisors were equivalent to one another), but the Bayesian test indicated that the evidence in favour of differential pick rates was weak. Furthermore, there was no convincing reason we could determine why the preference (if it were a real effect) would not show up at other confidence levels. The early experience of advisors did appear to predict their relative pick rates, with

[8] Later experiments were conducted to investigate why the hypothesised effects were not found in this experiment.

the advisor who agreed more with the participant in the initial experimental block being more likely to be picked more frequently on subsequent blocks.

Overall, these results were underwhelming and difficult to interpret. We conducted an on-line replication of this study so that we could collect data from more participants and hopefully determine more accurately whether or not participants were sensitive to the differences in the advisors.

3.4.2 Experiment 4A: confidence-contingent advice effects in the Dots task

The results of the previous studies indicated that people update their preferences for advisors using agreement in place of feedback where objective feedback is unavailable. Previous results from our lab (Pescetelli and Yeung 2021) suggested that this capacity is modified by confidence. The results of the previous experiment, however, failed to demonstrate these effects with advisor choice as the outcome.

The previous experiment was tightly controlled but the sample size was small. Using insights from its data, we refined the design and recruited a larger number of participants for the replication. Refinements included shortening the experiment, making the advisor profiles more extreme, reducing the questionnaires, and changing advisor representations from real faces and names to colours and numbers.

3.4.2.1 Open scholarship practices

This experiment was preregistered at https://osf.io/h6yb5. The experiment data are available in the `esmData` package for R (Jaquiery 2021c), and also directly from https://osf.io/xb4kh/. A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/90c04ff21d3a2876beaddd9ee35c577a821e5727/AdvisorChoice/index.html.

3.4.2.2 Method

54 participants each completed 368 trials over 7 blocks of the Dots task. Participants started with 2 blocks of 60 trials that contained no advice. The first 3 trials were introductory trials that explained the task. All trials in this section included feedback indicating whether or not the participant's response was correct.

Participants then did 5 trials with a practice advisor. They were informed that they would "get **advice** from an advisor to help you make your decision [original emphasis]", and that "advice is not always correct, but it is there to help you: if you use the advice you will perform better on the task."

Participants then performed 2 sets of 2 blocks each. These sets consisted of 1 Familiarisation block of 60 trials in which participants were assigned one of two advisors on each trial. The Familiarisation block was followed with a Test block of 60 trials in which participants could choose between the two advisors they encountered throughout the Familiarisation block. The participants saw different pairs of advisors in each set, with each pair consisting of one advisor with each of the advice profiles.

Compared to the lab version of this experiment§3.4.1.2, the design was somewhat simplified. The forced and choice trials were grouped into discrete blocks, so participants had all the forced trials for a pair of advisors first, and then all the choice trials for those advisors. Secondly, the advisors' biases were more extreme, as shown in Table 3.25.

Advice profiles The two advisor profiles used in the experiment were Bias sharing and Anti-bias. The advisors are balanced for their overall accuracy and agreement rates, but the Bias sharing advisor agrees more frequently with participants when their initial estimate is correct and made with relatively high confidence. The Anti-bias advisor agrees more frequently with participants when their initial estimate is correct and made with relatively low confidence (Table 3.25).

Table 3.25: Confidence-contingent advisor advice profiles

	Agreement						
	Initial estimate confidence[a]			... correctness			
Advisor	Low[b]	Medium[c]	High[b]	Correct[d]	Incorrect	Overall[e]	Accuracy[f]
Bias-sharing	50	70	90	70	30	**58.4**	70
Anti-bias	90	70	50	70	30	**58.4**	70

[a] Only correct trials received confidence-contingent advice

[b] 30% of trials

[c] 40% of trials

[d] Average correct agreement is the weigthed average of the previous three columns

[e] Overall agreement is p(correct) * p(agree|correct) + p(¬correct) * p(agree|¬correct)

[f] Overall accuracy is p(correct) * p(agree|correct) + p(¬correct) * p(¬agree|¬correct)

Table 3.26: Participant exclusions for Dots task Confidence-contingent agreement experiment

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	3
Skewed confidence categories	1
Too many participants	0
Total excluded	**4**
Total remaining	**50**

3.4.2.3 Results

Exclusions In line with the preregistration, participants' data were excluded from analysis where they had an average accuracy below 0.6 or above 0.85, did not have choice trials in all confidence categories (bottom 30%, middle 40%, and top 30% of prior confidence responses), had fewer than 12 trials in each confidence category, or finished the experiment after 50 participants had already submitted data which passed the other exclusion tests. Overall, 4 participants were excluded, with the details shown in Table 3.26.

Task performance Basic behavioural performance was similar to that observed with the same Dots task in Experiments 1A§3.1.1.3 and 2A§3.2.1.3. Initial estimate accuracy converged on the target 71%, and, as shown in Figure 3.31, participants benefited from advice in terms of their final decisions being more accurate than their

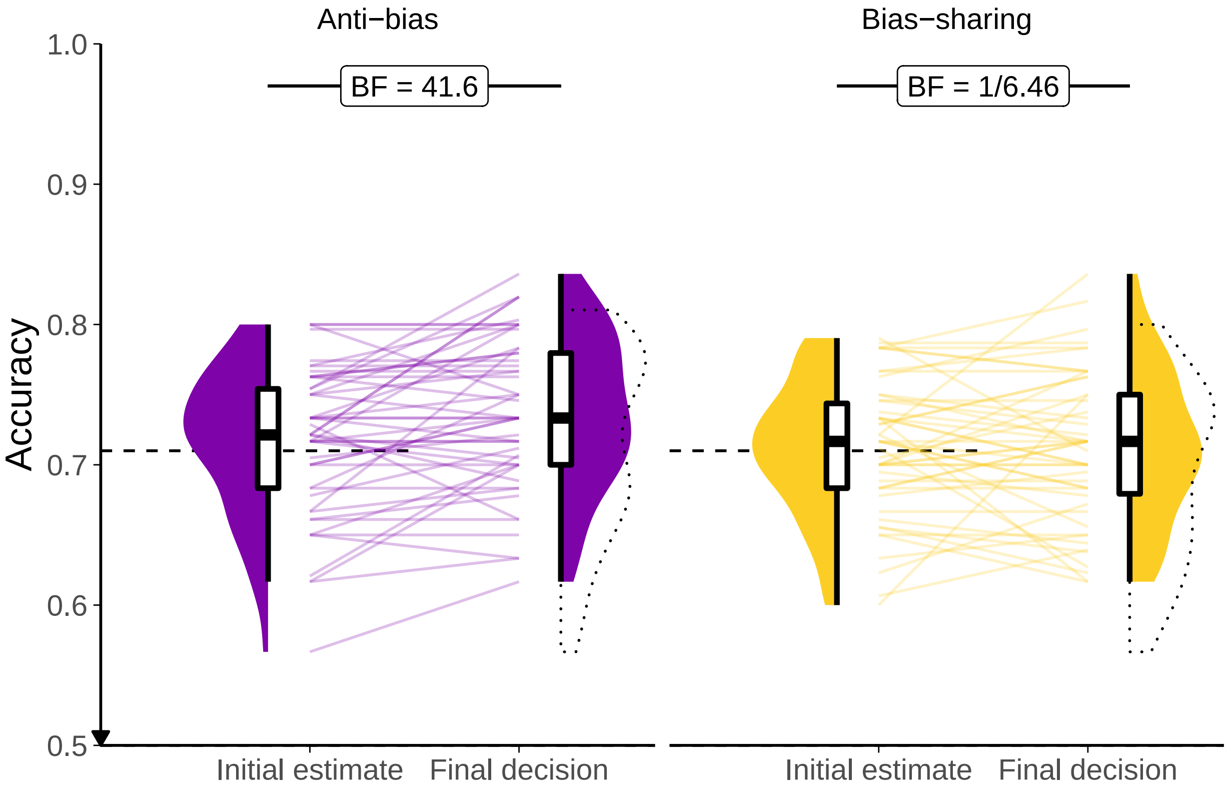


Figure 3.31: Response accuracy for the Dots task with confidence-contingent advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

initial estimates (ANOVA main effect of Time: F(1,49) = 5.78, p = .020; $\mathrm{M}_{\mathrm{Final}}$ = 0.72 [0.71, 0.73], $\mathrm{M}_{\mathrm{Initial}}$ = 0.71 [0.71, 0.72]), predominantly driven by following advice from the Anti-bias advisor (interaction of Time and Advisor: F(1,49) = 5.85, p = .019; $\mathrm{M}_{\mathrm{Improvement|Anti\text{-}Bias}}$ = 0.02 [0.01, 0.03], $\mathrm{M}_{\mathrm{Improvement|Bias\text{-}Sharing}}$ = 0.00 [-0.01, 0.01]). There was no main effect of Advisor (F(1,49) = 1.95, p = .169; $\mathrm{M}_{\mathrm{Anti\text{-}Bias}}$ = 0.73 [0.71, 0.74], $\mathrm{M}_{\mathrm{Bias\text{-}Sharing}}$ = 0.71 [0.70, 0.72]).

Figure 3.32 and ANOVA indicated that participants were more confident in their answers when they were correct compared to incorrect (F(1,49) = 187.32, p < .001; $\mathrm{M}_{\mathrm{Correct}}$ = 28.58 [26.49, 30.66], $\mathrm{M}_{\mathrm{Incorrect}}$ = 20.97 [18.67, 23.28]), that participants were less confident in their final decisions than their initial estimates (F(1,49) = 33.44, p < .001; $\mathrm{M}_{\mathrm{Final}}$ = 22.24 [19.87, 24.62], $\mathrm{M}_{\mathrm{Initial}}$ = 27.31 [25.08,

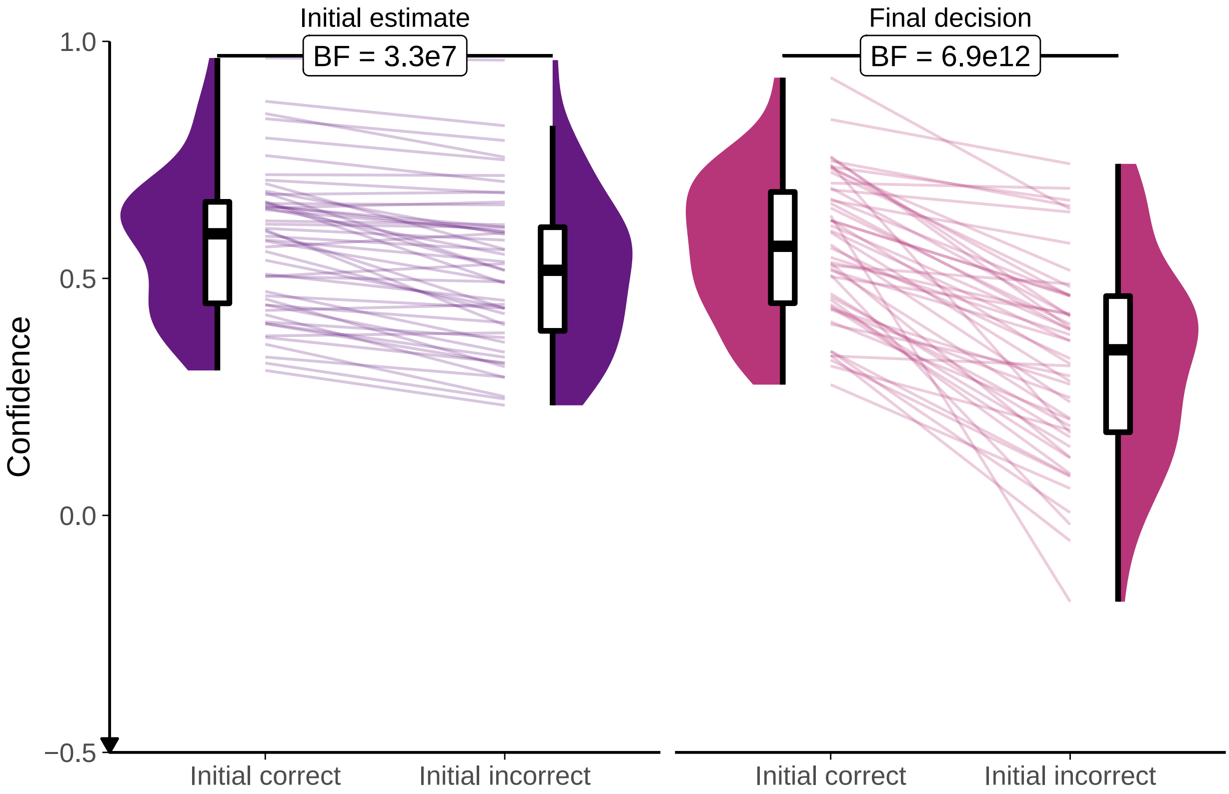


Figure 3.32: Confidence for the Dots task with confidence-contingent advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

29.53]), and that this decrease was larger for trials where the initial estimate was incorrect (F(1,49) = 65.10, p < .001; $\mathrm{M}_{\mathrm{Increase|Correct}}$ = -0.60 [-1.72, 0.52], $\mathrm{M}_{\mathrm{Increase|Incorrect}}$ = -9.52 [-12.24, -6.80]).

Advisor performance The advice is generated probabilistically from the rules described previously in Table 3.25. It is thus important to get a sense of the actual advice experienced by the participants.

Table 3.27 shows that advisors agreed with the participants' initial estimates at close to target rates in all confidence categories. There was a larger than expected difference in agreement on correct answers overall, but this was not a significant difference ($t(99) = 1.57$, $p = .118$, $d = 0.09$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.74$; $\mathrm{M}_{\mathrm{Anti\text{-}bias}}$ = 0.50 [0.46, 0.55], $\mathrm{M}_{\mathrm{Bias\text{-}sharing}}$ = 0.48 [0.44, 0.52]). The advisors were also further

Table 3.27: Advisor performance in Dots task with confidence-contingent advisors

	Agreement						
	Initial estimate confidence			... correctness			
Advisor	Low	Medium	High	Correct	Incorrect	Overall	Accuracy
Anti-bias	90.3	71.7	49.6	72.0	28.3	59.8	71.9
Anti-bias (Target)	*90*	*70*	*50*	*70*	*30*	*58.4*	*70*
Bias-sharing	49.0	69.0	90.3	67.8	28.8	56.5	68.7
Bias-sharing (Target)	*50*	*70*	*90*	*70*	*30*	*58.4*	*70*

apart in terms of overall accuracy than desired, because of the different agreement rates for correct answers, although once again this difference was not significant ($t(99) = 1.57$, $p = .118$, $d = 0.09$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.74$; $\mathrm{M}_{\text{Anti-bias}} = 0.50$ [0.46, 0.55], $\mathrm{M}_{\text{Bias-sharing}} = 0.48$ [0.44, 0.52]).

Hypothesis test There was a strong tendency for participants to express no, or slight preferences between advisors ($t(49) = \text{-}1.52$, $p = .134$, $d = 0.22$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/2.20$; $M = 0.47$ [0.44, 0.51], $\mu = 0.5$). Intriguingly, almost all participants who expressed a stronger preference expressed it towards the Anti-bias advisor: in the direction counter to that hypothesised (Figure 3.33).

In the Medium confidence trials, where the Lab experiment showed a glimmer of a difference, there was evidence against the existence of a difference ($t(49) = \text{-}0.91$, $p = .368$, $d = 0.13$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/4.40$; $M = 0.48$ [0.43, 0.53], $\mu = 0.5$).

In this experiment we saw an extremely strong effect of picking the advisor in the first position on the screen ($t(49) = 5.00$, $p < .001$, $d = 0.71$, $\mathrm{BF}_{\mathrm{H1:H0}} = 2.4\text{e}3$; $\mathrm{M}_{\mathrm{P(PickFirst)}} = 0.65$ [0.59, 0.71], $\mu = 0.5$), an effect that we would hope would be random and even out across participants. In this experiment, as a function of chance, the Bias-sharing advisor appeared in the favoured top position less frequently than we would expect ($\mathrm{BF}_{\mathrm{H1:H0}} = 18.3$; $\mathrm{M}_{\mathrm{P(BiasSharingFirst)}} = 0.48$ [0.47, 0.49], $\mu = 0.5$). The effect of preferring to pick the advisor in the first position, therefore, would enhance the pick rate of the Anti-bias advisor. This may to some extent explain why we did not find an effect of advisor (because a supposed preference for the Bias-sharing advisor was offset by a preference for picking the top advisor, i.e. favouring

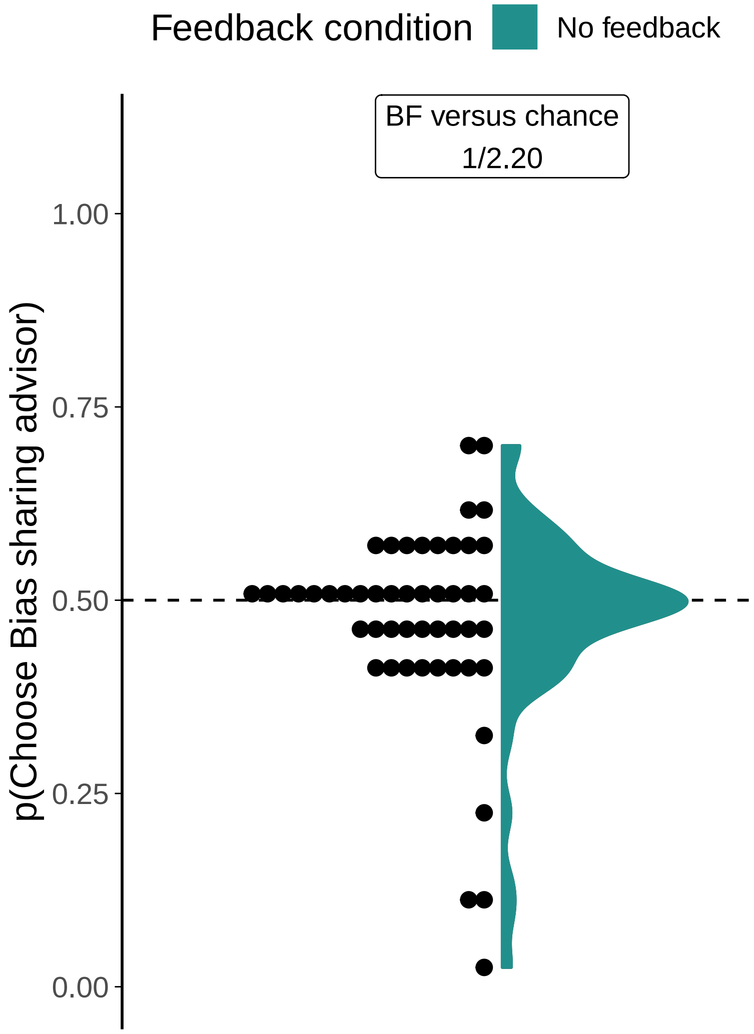


Figure 3.33: Dot task advisor choice for confidence-contingent advisors. Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

the Anti-bias advisor), although this explanation seems unlikely.

3.4.2.4 Discussion

While advisor choice and advice-taking are different domains, the previous experiments have shown strong similarities in the tendencies of participants: participants tend to be more influenced by the same kinds of advisors that they are more willing to hear from. On this basis, following Pescetelli and Yeung (2021), we would expect to see a preference for picking the Bias sharing advisor. We do not see this preference, and, insofar as we see any preference at all, we see the opposite.

3.4.3 Confidence-contingent advice effects in the Dates task

The Dates task was not used to study confidence-contingent advice because such advice requires both a precise control over the relative agreement and accuracy rates of the advisors and the ability to estimate confidence in responses. The advisors' agreement (and hence accuracy) profiles depend on the participant's performance, and this is unknown a priori in the Dates task whereas it is controlled in the Dots task using a staircase procedure. Different approaches to estimating participants' confidence were trialled, including a pilot experiment in which the width of marker used by participants was used as a proxy for confidence, but none of the approaches produced any discernible effect of confidence on advisor agreement.

We did attempt to design a version of the Dates task that would allow confidence-contingent advice, but pilot studies were largely unsuccessful and the time, cost, and effort required to refine the study were not deemed worthwhile in light of the null results from the Dots task studies.

3.4.4 Discussion

Across two tasks investigating whether people preferentially selected advisors who shared their biases we found scant evidence in favour of the effect. The two advisors, balanced for overall agreement and accuracy to eliminate the effects seen in previous experiments, differed in their likelihood of agreement based on the participants' initial estimate accuracy (following Pescetelli and Yeung 2021). The Bias-sharing advisor agreed more frequently with a participant when the initial estimate was made with high confidence, and less frequently when the initial estimate was made with low confidence, and the Anti-bias did the opposite. Overall, there was no evidence that participants picked these advisors at different rates. Looking just at the medium confidence trials, where the advisors were exactly equivalent, there was very weak evidence that the Bias-sharing advisor was picked more frequently in the Lab experiment and reasonable evidence against any difference at all in the on-line replication.

The data from these studies were not conclusive against the existence of this effect, especially when integrated with the findings of Pescetelli and Yeung (2021). It is plausible, for instance, that metacognitive moderation does happen in advice evaluation, but that these effects are more subtle than the broader effects of accuracy and agreement, and that the studies here were too underpowered to detect these effects. It seems unlikely that metacognitive moderation effects would exist in the advisor evaluation domain (as shown in Pescetelli and Yeung (2021)) but not in the advisor choice domain, given the rough parity demonstrated between these domains across the dimensions of accuracy and agreement in previous experiments.

3.5 General discussion

Previous work in our lab indicated that people are able to evaluate advice in the absence of feedback using agreement as a proxy (Pescetelli and Yeung 2021). We performed a series of experiments investigating whether the patterns observed for advice-taking are also evident in advisor choice. Our experiments exercised tight control over the advisors' answers in the domains of agreement and accuracy, and allowed us to explore their relative contributions.

Modelling and theoretical work indicates that biased source selection can dramatically reshape communication networks and create echo chamber effects where accurate but unpalatable information is ignored (Sunstein 2002; Madsen, Bailey, and Pilditch 2018). Empirical research on source selection behaviour has found relatively little indication that people behave this way in the real world (Marquart 2016; Sears and Freedman 1967; Nelson and Webster 2017), particularly in terms of avoiding exposure to unpalatable information (Weeks, Ksiazek, and Holbert 2016; Jang 2014). If the effects seen by Pescetelli and Yeung (2021) in the domain of advice-taking also occur in the domain of advisor choice, they may demonstrate in principle a psychological mechanism which could drive biased source selection that are rational and appropriate given the information available.

3.5.0.1 Advisor choice results

The preferences for advisors were broadly consistent with the pattern expected from previous work on advisor influence (Pescetelli and Yeung 2021). Experiments 1B and 3B showed that, where objective feedback could be used to calculate advisor performance, participants showed a systematic preference for picking the advisor who would provide the most accurate advice. Experiment 1A showed that this preference endured even when feedback was removed, although the result was not replicated in Experiment 1B.

Experiments 2A and 2B showed that participants who did not receive feedback systematically preferred to choose to get advice from High agreement advisors over Low agreement advisors. Experiments 3A and 3B indicated that this preference for agreement in the absence of feedback did not extend to a preference for agreement over accuracy. These results are consistent with an account of advisor trust updating which uses agreement as a proxy for advisor accuracy when more reliable information is not available, although some additional mechanism is required to explain how accuracy can dissociate from agreement in the absence of feedback. Pescetelli and Yeung (2021) offer confidence as an additional mechanism, but attempts to replicate their advisor influence experiment in the domain of advisor choice (Experiments 4A and the Lab experiment) did not provide empirical support for this.

3.5.0.2 Accuracy of advice

Our experiments showed that people will prefer to seek advice from an advisor who is more accurate, provided that they can identify that advisor. This coheres well with results from Experiment B.1.0.3 and other literature in the advice-taking domain where greater task accuracy is referred to as 'expertise' (Pescetelli and Yeung 2021; Yaniv and Kleinberger 2000; Gino, Brooks, and Schweitzer 2012; Rakoczy et al. 2015; Sniezek, Schrah, and Dalal 2004; Soll and Larrick 2009; Tost, Gino, and Larrick 2012; Schultze, Mojzisch, and Schulz-Hardt 2017; Wang and Du 2018; Önkal, Gönül, et al. 2017). A review of this advice-taking literature can be found in another chapter§5.2.3.1.

Like Pescetelli and Yeung (2021), we were interested in whether people are sensitive to accuracy for decisions they make where feedback is not provided. The majority of decisions on which we seek advice in everyday life do not come with feedback, or come with only infrequent and often delayed feedback. Although the perceptual decision-making task and the date estimation task used here are highly stylised, they can mimic this feedback regime. When feedback was denied to participants, leaving them unable to use the feedback to determine the accuracy of the advisors, their preferences only favoured the more accurate advisor in the Dots task (Experiment 1A§3.1.1.3), with no systematic preference in the Dates task (Experiment 1B§3.1.2.3). This difference might be explained by the relative difficulty of the tasks.

3.5.0.3 Confidence-weighted agreement as a proxy for accuracy

Where people cannot use objective feedback to determine the quality of advice, they use whether or not the advice agrees with them as a proxy for accuracy. This is shown by Experiments 2A§3.2.1.3 and 2B§3.2.2.3, and is consistent with Experiment 1A§3.1.1.3 (although not with Experiment 1B§3.1.2.3 given Experiment 3B§3.3.2.3). It also reflects, in the advisor choice domain, the findings of Pescetelli and Yeung (2021) in the advice-taking domain.

Pescetelli and Yeung (2021) theorised, on the basis of their experiments, that people assess advice on the basis of confidence-weighted agreement. This means that, where a judge is confident in their own opinion, offering agreement or disagreement results in a large increase or decrease in trust in the advisor, respectively. Where a judge is very unsure of the accuracy of their own opinion, however, neither agreement nor disagreement affect trust in the advisor very much – it is as if the judge has no frame of reference by which to assess the advice they have been given.

Assuming that trust in an advisor translates directly into a preference for hearing advice from that advisor, this theory accounts for some, but not all of our experimental results. It provides a straightforward account of the results of Experiment 2A§3.2.1.3 and 2B§3.2.2.3, where participants who did not receive feedback preferred to hear advice from the advisor more likely to agree with them (while overall advisor accuracy was held constant). It can also explain the results of Experiment 1A§3.1.1.3, because the probability that two independent binary judgements agree is related to the probabilities that they are greater than chance (Soll and Larrick 2009). Given participants were above chance in their perceptual decisions, the more accurate advisor would agree more, as indicated in Table 3.1. Furthermore, if agreement is weighted by confidence, this discrepancy would be greater where the weighting is higher to the extent that participants were well calibrated in their confidence: where they were most confident they were most likely to be correct, and so they were proportionally more likely to be agreed with by the more accurate advisor.

Experiment 1B§3.2.2.3 is harder to explain under the Pescetelli and Yeung (2021) theory. Participants should be sensitive to differential agreement rates as in Experiment 1A§3.2.1.3, and this should be enhanced to the extent that participants' confidence is well calibrated. The results were doubtful that participants with greater accuracy or confidence calibration expressed a greater preference for the more accurate advisor. It could have been the case that individual high-impact trials where a participant was extremely confident of the correct answer happened to systematically coincide with the less accurate advisor agreeing with the participant, but this seems very unlikely indeed. All in all, the results from Experiment 1B do not cohere well with the theory.

Experiments 3A§3.3.1.4 and 3B§3.3.2.3 may also be consistent with the theory. Participants denied feedback did not prefer to see advice from the agreeing advisor over the accurate one. This result would make sense if participants were able to detect the accuracy advantage of the accurate advisor: they might well have noticed that that advisor tended to agree with them where they were confident of the correct answer, and to disagree otherwise. The data did not support that conclusion. In the continuous Dates task used for Experiment 3B, the difference between agreement and accuracy is much clearer to participants because the advisors place markers on a timeline that also shows the participant's initial estimate. Participants in Experiment 3B may have been able to detect the redundancy of the agreeing advisor's advice and preferred the accurate advisor for that reason. We included a debrief questionnaire asking what participants though was the difference between the advisors, and several participants indicated in their responses that they had identified that the agreeing advisor tended to reflect their own answer.

The most clear failure of the theory in accounting for the present results is in the experiments using confidence-contingent advice; the Lab experiment and Experiment 4A§3.4.2.3. In these experiments, the advisors were specifically constructed such that they would be balanced for overall agreement and accuracy, but differentiated at the extremes of participants' subjective confidence. One specific test indicated that there may be an effect – a frequentist test on the moderate-confidence Lab experiment trials where advisors' agreement rates were balanced. This test, and this experiment, are the closest to those reported by Pescetelli and Yeung (2021). The effect, if there is one, does not seem to generalise to advisor choice behaviour in shorter versions of the task performed by participants outside of a laboratory setting.

Overall, the results of this suite of experiments suggest that trust in advisors, as evaluated by advisor choice preference, is determined by a variety of factors, even in highly constrained perceptual and general knowledge estimation tasks. One of these factors may be confidence-contingent agreement, or agreement more generally.

We have offered some speculation as to properties of the task and context that might alter the extent to which confidence-contingent agreement determines trust in advisors. Whether there is any dominant mechanism in play that accounts for a sizeable amount of the variation in trust remains unknown, but it appears unlikely that confidence-contingent agreement is that mechanism.

3.5.0.4 Variability between participants

The experiments produced a range of results, from clear evidence of systematic preferences to clear evidence of the lack of any such systematicity in preferences. Interestingly, participants displayed a range of preference patterns. Most participants in most conditions in most studies demonstrated fairly balanced picking behaviour; perhaps motivated by a sense of fairness or novelty, or perhaps as a result of random selection or the influence of other factors such as the position of the advisor on the screen. Where effects appear in the data, they show up as a fat tail on a skewed distribution: a sizeable minority select a given advisor on most trials, and almost no one selects the other advisor on most of the trials.

The advisor preference distributions differ somewhat between tasks. Although advisor choice distributions in both tasks are roughly normally distributed, those in the Dots task results are sharper. This sharpness may be a result of many participants selecting advisors at approximately equal rates. If participants become bored or fatigued by the experiment they may disengage and select advisors in a random manner.

Where manipulations are effective in the Dots task (e.g. Experiment 3B§3.3.1) they change the direction of preferences: a good many participants continue to pick advisors at approximately equal rates, but preferences in those who do express a preference systematically favour one particular advisor. Systematic differences in the Dates task are signified by distributions in which the modal preference moves to an extreme preference for the relevant advisor while the tails of the distribution continue to cover the whole range. These differences are likely a consequence of the difference in the number of Choice trials in each task. The Dates task only has around 10 Choice trials, and it is relatively common for participants to select a single preferred advisor repeatedly. The Dots task, however, has 30-60 Choice trials, meaning participants may become tired of seeing advice from the same advisor, increasing the novelty value of advice from the non-preferred advisor and thus reducing the apparent strength of preference. In support of this idea, the distribution of preferences in Experiment 3A§3.3.1.4), which had 30 Choice trials, was much more similar to those from the Dates task than the other Dots task experiments, which had 60 Choice trials. The similarity may have reflected the absence of any effect rather than the number of Choice trials, but it also more closely resembles the distributions of the Dates task where effects are present (Experiments 1B§3.1.2.3, 2B§3.2.2.3, and 3B§3.3.2.3) than the other Dots task distribution where they are not (4A§3.4.2.3).

The kind of heterogeneity revealed in the experiments' results is seldom included in population models of influence dynamics. In the following chapter§4, we explore the effects of including this kind of heterogeneity in agent-based models of advice giving.

3.5.0.5 Social and Personality perspectives

Source selection, most robustly selective exposure, is argued in the Social and Personality Psychology domains to be a product of a drive for cognitive consonance (Festinger 1957) or protecting a positive self-concept (Knobloch-Westerwick 2015) rather than an heuristic which is expected to improve the accuracy of decision-making under common real-world conditions. While these experiments do not directly contradict that literature, they do indicate that source selection can occur in the absence of any clear implications of the choice for a person's view of themselves. It may be argued that the effects in the absence of meaningful moral or political contexts reflects a bleed over from processes that perform useful work in vigilantly guarding a person's self-image, and that are always active when choices over information sources are made. Conversely, however, it could be argued that much of the selective exposure evidence represents a drive for accurate information: if

I strongly believe in a liberal world-view I may judge the reporting of a liberal media outlet to be more objectively accurate, and be more willing to view its content on that basis.

3.5.0.6 Limitations

As noted above, there were several issues with the experiments that were only noticed after data had been collected. The most serious of these was that the Dates task did not counterbalance the advisors' positions, and post hoc analysis of the Dots task data suggested that advisor position may have an unexpectedly important role in advisor choice behaviour. Another concern is that the difficulty of the Dates task was high, and also highly variable between participants. In some ways this is a strength, because it could theoretically allow us to detect variations in advisor choice behaviour as a function of participants' ability, but in practice we did not find these. We were left only with the drawbacks, therefore, of higher noise in the data and the possibility of strong effects of choice being driven by difficult-to-identify high-impact individual trials.

A related concern is the frequently large and significant effects of advisor position in the Dots task. Participants frequently picked the advisor at the top of the screen, regardless of that advisor's identity. This suggests disengagement with the task. Interestingly, the advisor at the top of the screen was further away from the answer bar (where participants input their initial estimates and final decisions) than the advisor at the bottom. We would expect lazy participants using a mouse to predominantly select the closer advisor. For participants using mobile phones, it would be equally easy to select either advisor.

Ecological validity is a concern more generally, too. While these studies were appropriate for capturing the behaviour of interest, they are also highly stylised in both their decision-making tasks and in the relationship between the participant and the advisor. The tasks are somewhat unusual in that people are seldom required to compare barely-seen visual scenes or estimate historical dates in their everyday lives.

When they do, they are unlikely to seek advice on them. This is a limitation general to most psychology experiments in one way or another, but no less important for that.

The relationship between the participants and the advisors was also unusual. While some researchers have suggested that social influence phenomena can be explained by reinforcement learning from repeated interactions (Behrens et al. 2008; FeldmanHall and Dunsmoor 2019; Heyes et al. 2020), the kind of repetition presented in these tasks is not a typical feature of human relationships. In some particular cases, for example working with a colleague at a job that requires rapid and repeated joint decision-making, behaviour might approximate that of our tasks, although for our domain of advisor choice to be of particular interest these people would also have to have the power to select their partner.

3.5.0.7 Conclusion

These experiments paint a rather equivocal picture, without being able to offer the kind of strong evidence we would like concerning the validity of the Pescetelli and Yeung (2021) theory of confidence-weighted advice evaluation. Despite the equivocal picture, we do find that people are selective in who they ask for advice, and that selections are dominated by accuracy where there is a reliable cue, and by agreement when there is not. One major observation revealed in the experiments was the extent to which participants exhibited a wide range of preference strengths and directions, particularly where systematic effects of advisor preference were absent. In the following chapter§4 we use agent-based computational modelling to investigate the consequences of agreement as a driver of advisor choice in the formation of echo chambers in information exchange networks. The simulations also explore the role of the kind of heterogeneity revealed in the experiments in this chapter.

4 Network effects of advisor choice

4.1 Introduction

The behavioural experiments in the previous chapter, along with predecessors from others in our lab (Pescetelli and Yeung 2021), indicate that people show preferences for selecting advisors who agree with them and treat advice which agrees with their initial judgements as more valuable. Previous modelling work has shown that either of these properties is sufficient to produce echo chamber-like structures in networks wherein individuals increasingly disregard information from those likely to disagree with them (Pescetelli and Yeung 2021; Madsen, Bailey, and Pilditch 2018). In this chapter, similar models are constructed, with the parameters controlling the agents' trust updating and advice seeking behaviour based on values fitted to the participants in the Dots task behavioural experiments. We aimed to explore the network effects observed by Pescetelli and Yeung (2021) and Madsen, Bailey, and Pilditch (2018) in a slightly different setup, with binary decisions that matched more closely to our behavioural tasks and, beyond this, to begin to explore the impact of individual variation. The key questions were whether the models would reproduce the echo chamber and polarisation effects seen in previous models, and how those effects would be altered by including heterogeneity within the population.

4.1.1 Agent-based models

Agent-based models are models in which numerous discrete decision-making agents interact to produce a model state, typically used to investigate emergent phenomena (Bonabeau 2002; Smith and Conrey 2007). Unlike typical models, which describe the overall behaviour of the system using equations, in an agent-based model each component agent (taken to represent an atom, a neuron, a starling, a person, etc.) has equations describing its behaviour at each time step. Crucially, the equations governing agents' behaviour will have terms relating to other agents' behaviour, and the outcome of the equations will produce changes in those properties. For example, a neuron might have inputs indicating whether or not a neighbouring neuron is in action potential, and the equations will determine whether or not the neuron itself triggers an action potential.

Agent-based modelling is a powerful approach, because it can connect system-level phenomena to individuals' decision rules and incorporate individuals with outlying or different decision rules (Bonabeau 2002). The power of agent-based modelling also comes with drawbacks (Jones 2007). The intuitive nature of its explanations means that it is easy to lose touch with ecological validity (because even non-ecologically valid implementations may seem valid when expressed in agent-based terms) and easy to over-claim on the basis of results. The interactivity and emergent behaviours that it produces can mean that some models are unstable: it is important to demonstrate robustness of emergent behaviours by running repeated tests on families of models. Furthermore, the multifaceted nature of the models means that there are more features to describe; modellers must make sure that the models are precisely specified and that those specifications are shared accurately so that others are able to reproduce and explore the models (Jones 2007; Bruch and Atwell 2015).

In social science research, agent-based models are best used to investigate "the implications for social dynamics of one or more empirical observations or stylised facts" (Bruch and Atwell 2015). The models used here are employed to investigate the effects of introducing the advisor choice phenomena documented in the previous chapter into network models of social influence. Bruch and Atwell (2015) argue strongly that agent-based models would benefit from greater empirical realism, and others, including Bonabeau (2002), have called for greater ecological validity. This chapter provides some evidence for this discussion by demonstrating the effects of implementing agents which copy individual participants versus drawing agents' coefficients from an empirically-defined distribution.

4.1.2 Similar models in the literature

The use of agent-based models has increased substantially in many fields in recent decades. Rather than attempt to present a full picture of agent-based models in which agents advise one another, this section briefly covers models that include a tendency for agents to seek out advice from like-minded others and compare the effects of varying this property. Models such as Song and Boomgaarden (2017) that include inter-agent influence but no variation in this propensity are not discussed.

Pescetelli and Yeung (2021) present a model which is a conceptual predecessor for the current model. In this model, agents make a single binary decision at each step on the basis of a perception of the true value and an idiosyncratic (static) bias. Agents then consult another agent for advice, and reach a final decision by integrating that advice with their initial estimate, weighted by the trust they have in their advisor. Model parameters varied according to whether or not agents received feedback, and whether or not agents sought out other agents in whom they had higher trust (similar to weighted selection in our model). They showed a cluster of interesting results, starting with the demonstration that using agreement as a proxy for accuracy in the absence of feedback is beneficial provided that individual agents' biases are uncorrelated. Secondly, they found that agents' trust weights aggregated into separate camps defined by shared biases where feedback was infrequent. These separate camps are analogous to echo chambers in human social networks. Lastly, they observed that agents' biases tended to polarise in the absence of feedback when biased agents preferentially sought like-minded advice.

Key differences between this model and the current model are that the current model allows for more heterogeneity in agents' properties and draws all agents' biases from a normal rather than uniform or point distribution. We also focus more on an exploration of low-feedback conditions rather than contrasting network behaviour under different feedback regimes.

In a similar vein, Madsen, Bailey, and Pilditch (2018) demonstrated that echo chambers could form within large networks of rational agents. Their models used agents who processed information in a Bayesian way. Crucially, those agents also had a likelihood of accepting advice that was dependent upon the similarity of the advice to their own prior beliefs. Under a range of parameters, at at a range of network sizes from 50 to 1000 agents, they showed that echo chambers formed consistently in most cases, illustrating that the agents' selective behaviour was a robust and dominant force in shaping networks.

van Overwalle and Heylighen (2006) replicated several social psychology results using small networks of connectionist neural networks. These results included a demonstration of majority influence (the authors describe it as 'polarisation'), in which repeated interactions bring about a consensus opinion in a small network of 11 agents. Regarding minority influence, Gao et al. (2015) reported, in a brief paper, that agents preferring to hear from like-minded others allowed stable minority opinion to persist where agents had to subscribe to one or other opinion. The persistence of these minority opinions in the face of widespread majority influence was a consequence of the self-reinforcing nature of the minority echo chamber.

Duggins (2017) offers a model of similar phenomena – social influence of opinions – using a somewhat different approach. The Duggins (2017) model uses spatial placement of agents to govern their ability to interact with others, and is based on a model of social influence driven by the difference between agents' perspectives alone, and does not include the agents' trust in one another as here.

These studies use agent-based models to approach questions concerning how people might decide where to get their information from. We employ the same method to the same ends, with a particular interest in how the heterogeneity seen in our behavioural experiments would alter the observed network dynamics. These studies used agents that were identical or nearly identical, and explored the consequences for the structure on the network on the basis of their position in the network. We aim to replicate these results, and extend them with an exploration of how allowing agents to exhibit the kind of heterogeneity seen in our behavioural experiment participants affects the overall patterns of changes in the network structure.

4.1.3 Aims

There are two parts to the modelling work in this chapter: exploring the characteristics of participants' recovered model parameters; and exploring the effects of those parameters on the dynamics of the agent-based models.

In the first part, models parameters are recovered from the behavioural data for each participant using a gradient descent algorithm. The two free parameters estimated during the parameter recovery process are: how rapidly trust updates (*trust volatility*); and the extent to which more trusted advisors are preferred (*weighted selection*).

The model and parameter recovery process are both tested using simulation. The parameter recovery process is tested by attempting to recover data from simulated data (where the parameters are known), and shows a reasonable level of accuracy in recovering parameters from data, suggesting that the recovered values for participants are likely to capture some aspects of participants' behaviour on the task. The model itself is tested by comparing errors for fitted parameters to errors for those same parameters when fitted to data that have had the advisor agreement information scrambled. Given the models are driven by the advisors' agreement with the judge's initial estimate, scrambling this column should break the temporal relationship between experience and trust updating, leading to worse parameter fits. This is indeed the case.

The parameter recovery results, distributions and correlations of the coefficients, illustrate the heterogeneity of the participants. Some participants' data do not fit well to the model, indicating that these participants may have used strategies which bear little resemblance to the model the data are fit to.

We then explore the behaviour of the agent-based models (within a narrow area of their parameter space), varying the way agents' preferences for more trusted advisors are determined. The Baseline model assigns all agents no preference for more trusted advisors, meaning that they sample from all possible advisors at random. The Homogeneous model applies the mean of the recovered parameter distribution to all agents. The Heterogeneous model draws values from a distribution defined by the mean and standard deviation of the recovered parameter distribution. The Empirical model selects values from those observed in the recovered parameter data (with replacement). For the first three models, trust update values are sampled from a distribution with the mean and standard deviation of the recovered parameters, while in the Empirical model the agents receive the trust update value for the participant whose preference strength coefficient they received. Two versions of each model are run, one which starts with a random trust network and one which starts with a homophilic trust network.

In short, we set up a comparison between the Baseline and Homogeneous model that provides a basic proof-of-concept replication of previous work described above: we expect to see polarisation of opinions and the emergence of echo chamber-like structures. Then we compare the Homogeneous model to the Heterogeneous and Empirical models, asking whether variation in agents' advice-taking and advisor choice behaviour changes the network dynamics. We also compare the Heterogeneous and Empirical models to explore whether the use of idiosyncratic parameter sets results in further alterations to the network dynamics.

4.1.4 Current model

In the current model, the agents face a binary decision in which they integrate evidence with prior expectations in a Bayesian manner to produce an initial estimate, analogous to the perceptual decision-making task in the Dots task experiments§2.1.3.1. The agents then receive advice from another agent, whom they choose on the basis of their trust in that agent compared to all the other agents combined with their preference for receiving advice from more trusted advisors. ‘Trust’ as encapsulated in the model is double-ended: very untrustworthy advisors can give reliable information because they are firmly expected to give the wrong answer (and therefore contain information about the correct answer because the decision is a binary one). The advice is integrated with the initial estimate, again in a Bayesian manner, to produce a final decision.

The final decision is used to update the agent’s perception of the world, and in parallel also to update their trust in the advisor. The agent’s bias drifts towards the final decision answer by weighted averaging. The trust they have in their advisor increases (or decreases) depending upon whether (or not) the advice is in agreement with the final decision. This dual update, of both expectation about the world and trust in the advisor, is a key feature of the model, and reflects ubiquitous situations where relative uncertainties have to be played off against one another (Körding et al. 2007; Behrens et al. 2008).

The estimate-advice-decision-update cycle is repeated thousands of times. The models illustrate the dynamics of advice interactions: how prior expectations change over time and whether and how polarisation or consensuses emerge; and how agents’ trust in one another changes as they gain experience with one another.

4.2 Method

Agent-based modelling is used to simulate the interactive effects of repeated paired decision-making. The agents in the model perform a cycle of making an initial estimate, selecting one of the other agents as an advisor, receiving advice from the chosen advisor, making a final decision, and updating their trust in their advisor and their internal beliefs about the world. The task the agents face is roughly analogous to the task faced by human participants in the Dots task experiments§2.1.3.1, where the participants make a decision about which of two rapidly and simultaneously presented grids contained more dots.

Each model consists of a population of agents implementing the same mathematical model of decision-making, trust-updating, and advisor choice, with different coefficients for parameters of that mathematical model (Table 4.1) drawn from appropriate distributions.

4.2.1 Open code

The agent-based modelling is performed using the custom-written R package `adviseR` (Jaquiery 2021b). The functions which implement each of the steps below are listed in the (non-exported) simulation loop function `simulationStep`. Each of these sub-functions includes unit tests to verify its behaviour. For readers who prefer to follow along with code rather than maths, each section below includes details of the function implementing the equations. These small functions are not exported by the package, so links are to source code rather than package documentation.

The models take a long time to run when creating the thesis from base code. To assist with exploring the models without having to rerun models from scratch, cached data is available from Jaquiery (2021a).

4.2.2 Model details

Each model is defined by a set of model parameters (4.2), which are assigned capital letters in the equations below. These parameters are used to generate further parameters which govern agents' tendencies, and these are assigned lowercase letters and superscripted with the agent to whom they belong. Variables which change over the course of the model have a subscript indicating which step of the model they belong to.

This is best illustrated with an example. Each agent has a sensitivity parameter, s^a which governs the amount of random noise which is attached to their perception of the world. On any given decision, the amount of this noise ϵ_t^a is determined by drawing from a normal distribution with standard deviation defined by the agent's sensitivity ($N(0, \frac{1}{s^a})$). When the agent is created, the parameter defining its

Table 4.1: Agent properties

Property	Description	Updates each step
s^a	Accuracy of agent's perceptions	No
c^a	Agent's subjective confidence scaling	No
τ^a	Size of agent's trust updates	No
λ^a	Size of agent's bias updates	No
w^a	Extent of agent's preference for trusted advisors	No
b^a_t	Agent's prior expectation about the task answer	Yes

sensitivity distribution (s^a) is itself drawn from a normal distribution with mean and standard deviation defined by the model settings ($N(S^\mu, S^\sigma)$).

4.2.2.1 Creating agents

Each agent has the properties listed in Table 4.1. An agent's values for each of these properties are created by drawing values from normal distributions defined by parameters in the model settings (Table 4.2).

Additionally, each agent has a 'trust' in each other agent: an estimate of that agent's reliability ranging from being convinced that the other agent is always correct to being convinced that the other agent is always wrong. These values are updated at each step and initialised by drawing from a uniform random distribution with limits [0.5, 1].

There is an ongoing debate about whether these kinds of advice models should support expectations of lying. In some models, an agent's lowest trust value for another agent indicates that they expect that agent's advice will be useless (no better than random chance), while others use a floor which indicates the agent believes the advisor to be deliberately misleading (providing information in the strict sense that advice is predictive of correct answers because they reliably point to the incorrect answer). Collins et al. (2018) provide a discussion of this issue along with initial evidence that advice can be considered misleading in some cases. The current model treads a hybrid path between these positions, by initialising the trust weights such that the lowest level of starting trust is equivalent to considering the advice as random (containing no information about the truth), but allowing trust weights to decrease following interaction so that advice is considered as pointing away from the truth.

Agents are created in the `make_agents` function.

Connecting agents Agents were joined together using a fully-connected network. It is common for models of social interactions to use connectivity graphs that reflect structural features common to social networks, including small-world (Watts and Strogatz 1998) and small-world scale-free (Humphries and Gurney 2008) networks, in which each agent is connected to several others such that the network is constructed of small clusters with a low average path length (the minimum number of steps required to connect two nodes, for all pairs of nodes in the network). Small-world scale-free networks also implement a scale-free power law structure, in which a few nodes are highly connected while most are less well connected.

The networks used here do not have these properties, and are instead fully-connected networks, wherein each agent is connected to each other agent, and the likelihood of two agents interacting is governed by their trust weights for one another. This is because we follow Pescetelli and Yeung (2021), who used fully-connected networks, and because non-fully-connected networks introduce unnecessary complexity, such as the dependency of behaviour on an agent's position in a network (McClain 2016) and the mechanisms used for creating and breaking connections.

4.2.2.2 Model step

Establishing the stimulus The same stimulus is presented to each agent, in the form of a value (v_t) drawn from a normal distribution with mean (V^μ) and standard deviation (V^σ) defined in the model settings:

$$v_t \sim N(V^\mu, V^\sigma) \tag{4.1}$$

The agents' task is to determine whether v_t is greater than or less than zero.

The establishing of true values is achieved in the code within the `simulationStep` function, and executes a function supplied by the user. The default function, used in all models described below, is drawing from the normal distribution. This default can be seen in the `truth_fun` parameter for the `runSimulation` function.

Initial estimate-making Agents' perception of the stimulus is imperfect, simulated by combining the true stimulus value with random noise (ϵ^a_t), to produce a percept q^a_t.

$$q^a_t = v_t + \epsilon^a_t \tag{4.2}$$

Where: $\epsilon^a_t \sim N(0, \frac{1}{s^a})$

This sensory percept is then converted into a subjective probability that the stimulus was greater than zero (p^a_t) using a sigmoid function with a slope defined by the agent's subjective confidence scaling parameter.

$$p^a_t = \varsigma(q^a_t, c^a) \tag{4.3}$$

Where: $\varsigma(x, y) = \frac{1}{1+e^{-xy}}$

The conversion of the percept (q^a_t) to the subjective probability (p^a_t) changes the representation from a theoretically-unbounded normal distribution centred around 0 to a probability in the interval [0,1] centred around 0.5. The subjective probability expresses a judgement about whether or not v_t is greater than or less than zero, but does not contain an estimate of v_t itself.

In the code, the percept is calculated using `getPercept`. The inclusion of the agent's confidence scaling is done as part of the second step in the determination of the initial estimate.

This subjective probability is then integrated with the agent's prior expectation about whether stimuli are generally less than or greater than 0 (b^a_t), or 'bias', to produce an initial estimate, i^a_t. This integration is performed using Bayes' rule.

$$i_t^a = \frac{b_t^a \cdot P(p_t^a|v_t > 0)}{b_t^a \cdot P(p_t^a|v_t > 0) + (1 - b_t^a) \cdot P(p_t^a|v_t \leq 0)} \tag{4.4}$$

Where: $P(p_t^a|v_t > 0) = z(p_t^a, 1, V^\sigma)$;

and: $P(p_t^a|v_t \leq 0) = z(p_t^a, 0, V^\sigma)$;

with: $z(x, \mu, \sigma)$ giving the density of $N(\mu, \sigma)$ at x.

The prior, b_t^a, is equal the prior probability that the value is greater than 0: $b_t^a = P(v_t > 0)$

In the code, the calculation of initial estimate from the percept (including the scaling of the percept according to the agent's confidence scaling parameter) is performed in `getConfidence`.

Note that in these initial estimate equations the agents have direct access to a model property, V^σ, the standard deviation of the true values of the stimuli. Ideally, in an agent-based model, the agents would not have such direct access to non-observable properties, and would instead build up specific expectations about the variability of stimuli from observation, perhaps seeded with a loosely-informative prior. The agents are allowed to know the value here as a shorthand for such exploration. This makes the agents' task analogous to well-practised real-world tasks such as perceptual decision-making. It is unlikely to seriously affect any conclusions or illustrations drawn from the models.

The initial estimate contains both a discrete decision (whether the stimulus value was more likely to be less than zero, $i_t^a < 0.5$, or greater than zero $i_t^a \geq 0.5$), and how much so ($|i_t^a - 0.5|$). It thus represents both the agent's decision and the confidence in that decision.

Advisor choice Having made an initial estimate, each agent selects a single advisor from whom to receive advice. This choice is made based on the trust in each potential advisor scaled by the strength of the agent's preference for receiving more-trusted advice.

The identity of an agent's advisor on a given trial is designated by a', and the weight assigned to advisor a' by agent a at step t by $\omega_t^{a,a'}$. Each agent's trust value is adjusted to be relative to the most (or least) trusted advisor, depending upon whether the agent prefers trusted ($\mathrm{w}^a > 0$) or untrusted ($\mathrm{w}^a \leq 0$) advisors.

$$\omega'^{\ a,a'} = \begin{cases} \mathrm{w}^a > 0, \omega_t^{a,a'} - \max(\Omega_t^a) \\ \mathrm{w}^a \leq 0, \omega_t^{a,a'} - \min(\Omega_t^a) \end{cases} \tag{4.5}$$

Where Ω_t^a is the set of trust weights in all potential advisors for agent a at step t.

These relative trust values are then assigned probability weightings for selection based on a sigmoid function. Because of the relative scaling, each probability is in fact drawn from a half sigmoid, where the probability weight assigned to the most likely candidate is 1 and other candidates' probability weights between 1 and 0. The w^a parameter is a crucial one in the model; it governs the steepness of this sigmoid function (Figure 4.1). When its value is large (either negative or positive), an agent is much more consistent in their selections – smaller differences in trust values in different potential advisors translate into a much higher likelihood of selecting advisors. When its value is low, the agent's trust in potential advisors exerts much less influence on which potential advisor is selected.

The identity of advisor ${}_t^{a,a'}$ is determined by sampling at random from the advisors, weighted by the probability weights assigned. In the R code the advisor choice procedure occurs in `selectAdvisorSimple`.

Differences from parameter estimation The approach used for fitting participant data from the behavioural experiments is subtly different from that used for advisor choice in the agent-based models. In the behavioural experiments, rather than being presented with a choice of many different advisors whose appeals all had to be considered, participants were presented with a choice of two advisors only. The parameter recovery process estimated a trust update rate (τ^a in the agent-based model) which tracked the trust in each advisor, and the difference between these trust values was fed into a sigmoid governing selection. The slope of

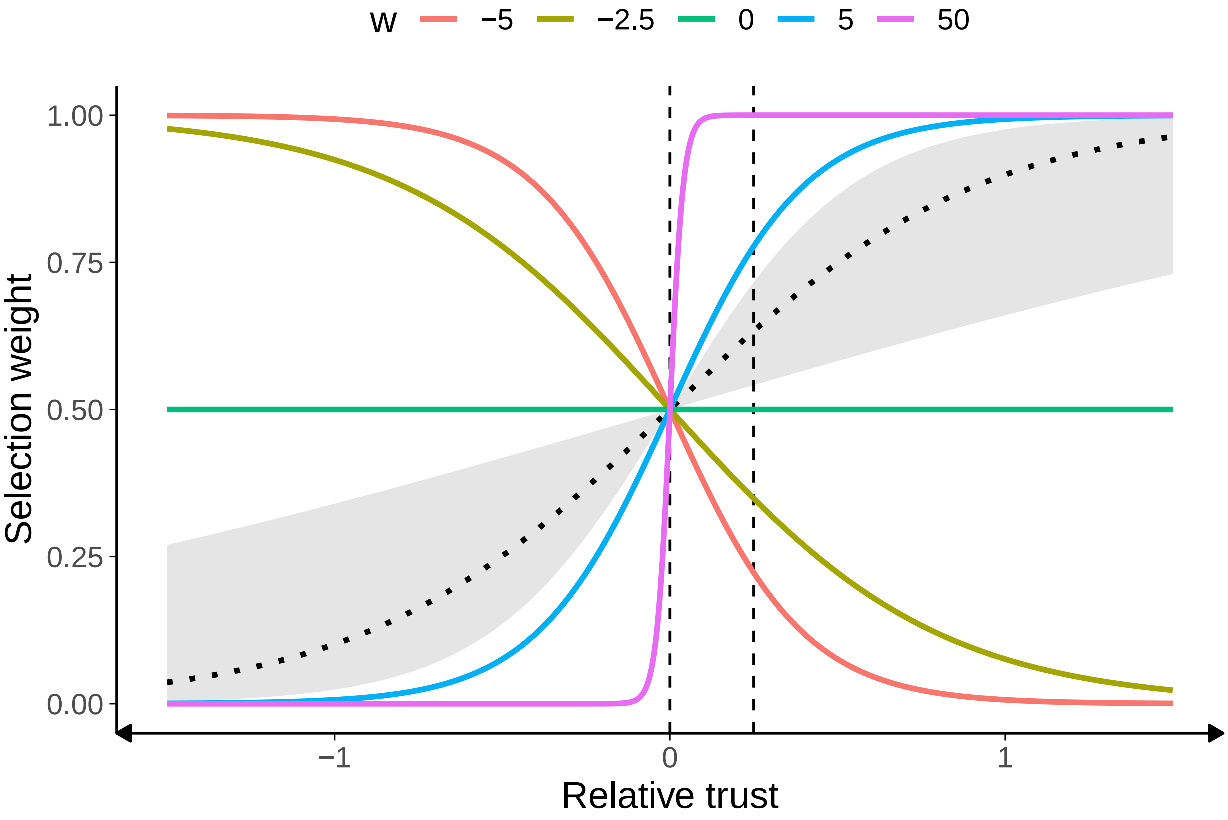


Figure 4.1: Weighted selection parameter.
Example values of weighted selection are shown. For postive values of w, as relative trust increases, the selection weight also increases. Correspondingly, for negative values of w the selection weight decreases as trust increases. Compare the intercepts with the dashed lines: all seleciton weights are 0.5 where relative trust is zero, equating to random picking between equivalently-weighted advisors. Where w is 0 (green line), relative trust differences make no difference to selection weighting. Where it is large (purple line), on the other hand, relatively small differences in relative trust translate into much greater selection weighting.
The dotted line and shaded area represent the sigmoids calculated using the mean empirical value and its 95% confidence limits.

that best-fitting sigmoid function was taken as equivalent to w^a in the agent-based models. This approach is reasonable given the differences in the advisor choice task facing the human participants and model agents, but should be noted as a caveat for drawing interpretations concerning the role of weighted selection values based on human participants' performance.

Code for the implementation of advisor choice fitting can be found in the function `advisor_pick_probability`.

Final decision-making Final decisions are made by Bayesian integration of the initial estimate and advice. Advice takes the form of a binary recommendation, and is weighted by the trust the agent has in their advisor.

First, initial estimates and advice are reoriented to the direction of the initial estimate, such that initial estimates represent confidence in the initial estimate and advice represents agreement with that decision:

$$i'\,_t^a = \begin{cases} i < 0.5, 1 - i_t^a \\ i \geq 0.5, i_t^a \end{cases} \tag{4.6}$$

$$i'\,_t^{a,a'} = \begin{cases} i < 0.5, \text{round}(1 - i_t^{a,a'}) \\ i \geq 0.5, \text{round}(i_t^{a,a'}) \end{cases} \tag{4.7}$$

The trust weight is slightly truncated to avoid very extreme values, and oriented based on whether the advice agrees, giving the expectedness of the advice provided the initial answer was correct:

$$\omega'\,_t^{a,a'} = \begin{cases} i'\,_t^{a,a'} = 0, 1 - \min(0.95, \max(0.05, \omega_t^{a,a'})) \\ i'\,_t^{a,a'} = 1, \min(0.95, \max(0.05, \omega_t^{a,a'})) \end{cases} \tag{4.8}$$

The final decision is obtained by performing Bayesian integration. In Bayesian terms, agents are trying to discover the probability of their final answer being correct given the agreement (or disagreement) observed from their advisor. Thus they multiply the initial probability of being correct (subjective confidence, $i'\,_t^a$) by the probability of the advisor agreeing if they are correct ($w'\,_t^{a,a'}$). This is divided by all of the options that could have led to the observed advice: the probability that they are correct multiplied by the probability of the advice if they are (the numerator), plus the probability that they are incorrect multiplied by the probability of the advice if they are incorrect. Because correctness and advice agreement probability are both mutually exclusive binaries, the probability of being incorrect is 1 - the probability of being correct, and the probability of the advice if they are incorrect is 1 - the probability of the advice if they are correct:

$$f'\,{}_t^a = \frac{i'\;a_t \cdot \omega'\,{}_t^{a,a'}}{i'\;a_t \cdot \omega'\,{}_t^{a,a'} + (1 - i'\;a_t)(1 - \omega'\,{}_t^{a,a'})} \tag{4.9}$$

The final decision (with confidence) is acquired by reversing the transformation applied earlier:

$$f_t^a = \begin{cases} i < 0.5, 1 - f'\,{}_t^a \\ i \geq 0.5, f'\,{}_t^a \end{cases} \tag{4.10}$$

Final decisions are calculated in the code in the `bayes` function.

Feedback We are primarily interested in exploring how trust updates in situations where feedback is rare or absent, as is often the case in real-world tasks. Correspondingly, the extent to which feedback occurs is a parameter in our model. On a proportion of trials a proportion of the agents are randomly selected to receive feedback. These proportions can range from 0 – no trials have feedback or no agents get feedback – to 1 – all trials have feedback or all agents get feedback. The extent to which feedback saturates the model is the product of these parameters. The feedback is always accurate, and implicitly trusted by all the agents. When they receive feedback, the agents use that feedback rather than their own final decisions to update their trust and bias. Where they do not receive feedback, agents have to rely on their own subjective estimates to determine the validity of the advice§4.2.2.2. The R implementation is part of the `simulationStep` function.

Bias updating The agents update their bias – i.e., their experience-based estimate of the relative likelihoods of the two outcomes of the binary decision – after each final decision, taking an average of their current bias and the feedback or their final decision, weighted by the agent's bias volatility (λ^a). They therefore adjust their prior expectations on the basis of their experience (and advice). Participants in perceptual decision experiments can update their prior expectations in response to base rates, even when feedback is not provided (Zylberberg, Wolpert, and Shadlen

2017). Indeed, (Haddara and Rahnev 2020) suggest part of the effect of feedback is to reduce bias. Bias is specifically updated according to the following equation:

$$b'^{\,a}_{t+1} = b^a \cdot (1 - \lambda^a) + f^a_t \cdot \lambda^a \tag{4.11}$$

The bias is clamped to within 0.05 and 0.95 to keep agents at least a little open-minded:

$$b^a_{t+1} = \min(0.95, \max(0.05, b'^{\,a}_{t+1})) \tag{4.12}$$

This is implemented using the `getUpdatedBias` function.

Trust updating The agents update their trust in an advisor by taking an average of their current trust in their advisor and that advisor's agreement, weighted by their trust volatility (τ^a). This is the second of our key parameters in the model (alongside weighted selection§4.2.2.2), and the way they interact is of critical importance. High values of trust volatility lead to an agent rapidly deciding an advisor is excellent or worthless, and will dramatically alter the likelihood of selecting that advisor again, depending on the agent's weighted selection parameter. Trust update volatility is given by:

$$x = w^{a,a'}_t \cdot (1 - \tau^a) + i'^{\,a,a'}_{t} \cdot \tau^a \tag{4.13}$$

A tiny cap is used for truncation to prevent agents wholly disregarding or blindly trusting others:

$$w^{a,a'}_{t+1} = \min(0.9999, \max(0.0001, x)) \tag{4.14}$$

Trust updates are accomplished using the `newWeightsByDrift` function.

In the current model, trust weights for an advisor do not change unless advice is received from that advisor. This is not the only approach; alternatives include decay of trust in unselected advisors towards a default value and normalisation of trust weights such that the sum of an agent's trust weights for all their advisors remains constant. Whether or not trust decays over time and whether or not increased trust in one advisor diminishes trust in other advisors are empirical questions (and may change based on context), but we do not expect the results of the model would be substantially different if trust weights for unselected advisors were updated in these ways.

4.2.3 Model parameters

Beyond the critical parameters of interest – weighted selection and trust volatility – the models have a large number of parameters. These can be broadly divided into parameters which govern the distributions from which the agents' parameters are drawn, and parameters which define the operation of the model.

These parameters can be seen in full by inspecting the parameters for the `runSimulation` function.

4.2.3.1 Parameters varied between model runs

There are three parameters which are varied between runs:

- `decision_flags`, whether or not there is time for the random network weights to update before agents' biases begin to update
- W^{μ}, W^{σ}, the agents' weighted selection values

The exact specification of how these parameters are varied can be seen in the source code for this document.

4.2.3.2 Constant parameters

Many parameters are held constant across runs. Within reasonable tolerances, these parameters do not have substantial effects upon the model dynamics. A detailed exploration of the model space is beyond the scope of this work, but all parameters were varied in some way during model development. The main parameters which could in principle vary but which are held constant are:

Table 4.2: Model properties

Property	Description	Type
`n_agents`	Number of agents to simulate	model
`n_steps`	Number of decisions to simulate	model
`decision_flags`	Binary flags indicating whether trust and/or bias update at each step	model
`feedback_probability`	Probability feedback is provided each step	model
`feedback_proportion`	Proportion of the population receiving feedback when provided	model
`random_seed`	Seed used for the pseudorandom number generator to allow repetition of runs	model
`truth_sd`	Standard deviation for true world values	model
`confidence_weighted`	Whether agents use their own confidence to modify their trust updates	model
B^μ, B^σ	Mean, standard deviation of distribution of agents' biases	agents
S^μ, S^σ	Mean, standard deviation of distribution of agents' sensitivities	agents
T^μ, T^σ	Mean, standard deviation of distribution of agents' trust volatilities	agents
$\Lambda^\mu, \Lambda^\sigma$	Mean, standard deviation of distribution of agents' bias volatilities	agents
C^μ, C^σ	Mean, standard deviation of distribution of agents' confidence scaling	agents
W^μ, W^σ	Mean, standard deviation of distribution of agents' trusted advisor preference	agents

- `feedback_probability` and `feedback_proportion`, which force the agents' biases to remain closer to the optimal value of 0.5. The bias reduction effect is as expected from Haddara and Rahnev (2020).
- B^μ, which increases polarisation and homophily by separating the mean of the agent groups
- B^σ, which has little effect independent of B^μ, but can increase the frequency of extreme and moderate biases

- S^{μ} and S^{σ}, which interact with `truth_sd` and C^{μ}, C^{σ} to determine agreement rates given a constant bias, increasing or decreasing the power of weighted selection to shape network dynamics
- Λ^{μ} and Λ^{σ}, which increase the speed of the network dynamics (so that similar trajectories occur over fewer generations)
- and `confidence_weighted`, which decreases the speed of the network dynamics.

It is plausible that there are interactions between these and other model parameters that were not detected during model development. The model code is made available to allow others who may be curious about aspects of the model beyond the scope of this thesis to investigate behaviour in these regions of the parameter space. Formal runs illustrating the effects described above are not provided here because they would require a huge amount of computational time to generate for the full models described in this chapter.

4.2.3.3 Empirically estimated coefficients

The models which have their weighted_sampling value marked as 'emp' are constructed by drawing parameter values for trust volatility (τ^{a}) and weighted selection (w^{a}) directly from the parameter coefficients estimated from the empirical Dots task data. The weighted selection values in other models are taken from distributions defined by the estimated values for the population, which allows for the generation of an unlimited number of unique participants but means that the agents produced will be more homogeneous in their overall strategies. For example, if a minority strategy exists within the population, the parameter estimate distributions reflective of the dominant strategy would be slightly modified through averaging towards the minority strategy, which may in turn produce agents whose behaviour is not reflective of any plausible real individuals.

Using parameters estimated from actual individuals instead of drawing from a distribution derived form those values has both strengths and weaknesses. The strengths are that the values represent genuine best-estimates of both trust volatility and weighted selection of real participants who completed one of our Dots task behavioural experiments. Given that these two parameters are related to one another, with weighted selection being dependent on trust volatility, it may be important to use observations of both simultaneously to appropriately model individuals' behaviour. The weaknesses are that the task given to human participants differed in potentially important ways from the task modelled in the agent-based model. The most potentially important difference in this respect is the choice of advisor: in the behavioural experiments the human participants were familiarised with two advisors and then given the choice between them, whereas in the model the agents are picking from a large number of potential advisors. The parameter is estimated on the basis of a sigmoid function applied to the trust difference between advisors, whereas it is used in the agent-based models in a half-sigmoid applied to the difference between each advisor and the most trusted advisor.

These considerations mean that the model dynamics arising from using estimated coefficients may be informative, but only in an illustrative capacity. Too much differs between the behavioural and simulated situations to draw strong conclusions.

Behavioural data source The behavioural data were taken from participants who completed one of the Dots tasks experiments. This included participants in experiments that were unsuitable for analysis because of bugs in the experiment implementation. Participants in experiments that did not offer a choice of advisor were not included, because it is not possible to estimate the weighted selection parameter w^a for participants who never make choices between advisors. Ultimately, this meant that data were drawn from 593 participants over 12 different experiments. Only trials that had advice were included, whether the advice came from a chosen advisor or an assigned advisor. Overall, there were 170590 trials in the data, of which 56940 trials (33.38%) had a choice of advisor.

The Dots task data were used for this because the Dots task presents a more consistent experience between participants. Whereas in the Dates task some participants perform far better than others, in the Dots task overall performance is clamped to around 71% using a staircase procedure. Another advantage to the Dots task data is that there are many more trials, meaning that the recovered parameters are much more precise.

4.2.3.4 Model results analysis

The model behaviours of interest are polarisation – the extent to which the agents' biases tend towards extreme values – and echo chamber formation – the extent to which agents form separate trust bubbles wherein they have high trust towards others inside the bubble but low trust towards those outside. The true probability of the binary answer being greater than 0.5 is 0.5; i.e. either answer is equally likely.

Polarisation is operationalised as the mean absolute deviation of agents' biases from 0.5. Echo chamber formation is quantified in two ways: shared bias-weight correlation and group ratio. Shared bias-weight correlation is the correlation between the amount of shared bias agents have (1 - the absolute difference between their biases) and the trust weight of one agent for the other. This is calculated for each directional tie in the network, meaning that each shared bias is accompanied by two trust weights. The overall correlation between these values gives the extent to which agents' trust in other agents is associated with their shared expectation about the result. Group ratio is calculated by splitting the population of agents into two groups based on whether or not their bias is above 0.5. The mean of the trust weights for agents within their group is then divided by the mean of the trust weights for agents outside their group. The higher this ratio, the more extreme agents' tendencies to trust those with similar biases compared to those with different biases.

4.3 Results

4.3.1 Model checks

Verification of this model fitting approach was conducted on the Advanced Research Cluster (Richards 2015) in three ways.

Firstly, simulated participant datasets were generated by using the recovery model. This was achieved by running the recovery model in reverse with known coefficients for w^a and λ^a parameters. This produced simulated data generated from known coefficients. The recovery model was then run on these simulated data, and the recovered estimates for the w^a and λ^a parameters were compared with the known coefficients used to generate the data. This illustrated that known coefficients could be recovered.

Secondly, error scores were compared for coefficients fitted to each participants' advice-taking and advisor choice data and coefficients fitted to versions of that participant's data where the advisor agreement column values had been shuffled. Overwhelmingly, the fitting error on the shuffled data tended to be higher than the fit to the original data, indicating that the models were sensitive to participant patterns related to advisor agreement (Figure 4.2).

Finally, the error for coefficients fitted to the participants' data were compared to errors for the same coefficients fitted to shuffled data. Error values for the shuffled data tended to be higher than values for the original data, suggesting that participants' behaviour shared some features with the model's expectations.

4.3.2 Participants' coefficients

The model specified above§4.2.2.2 was fit to each participant's data using a gradient descent algorithm. The model was run with random values for w^a and λ^a, and its fitness measured. The fitness function for this algorithm assessed the combined error of the model in predicting the participant's advisor choice and advice-taking behavioural data. The algorithm then updated the w^a and λ^a parameters individually and reran the model. If the model run with the new parameter had better fitness, that new value was kept for that parameter. The algorithm then began again, and halted when it detected that no more progress with increasing fitness was being made. The model assumed that at the first encounter the advisors were equally trusted.

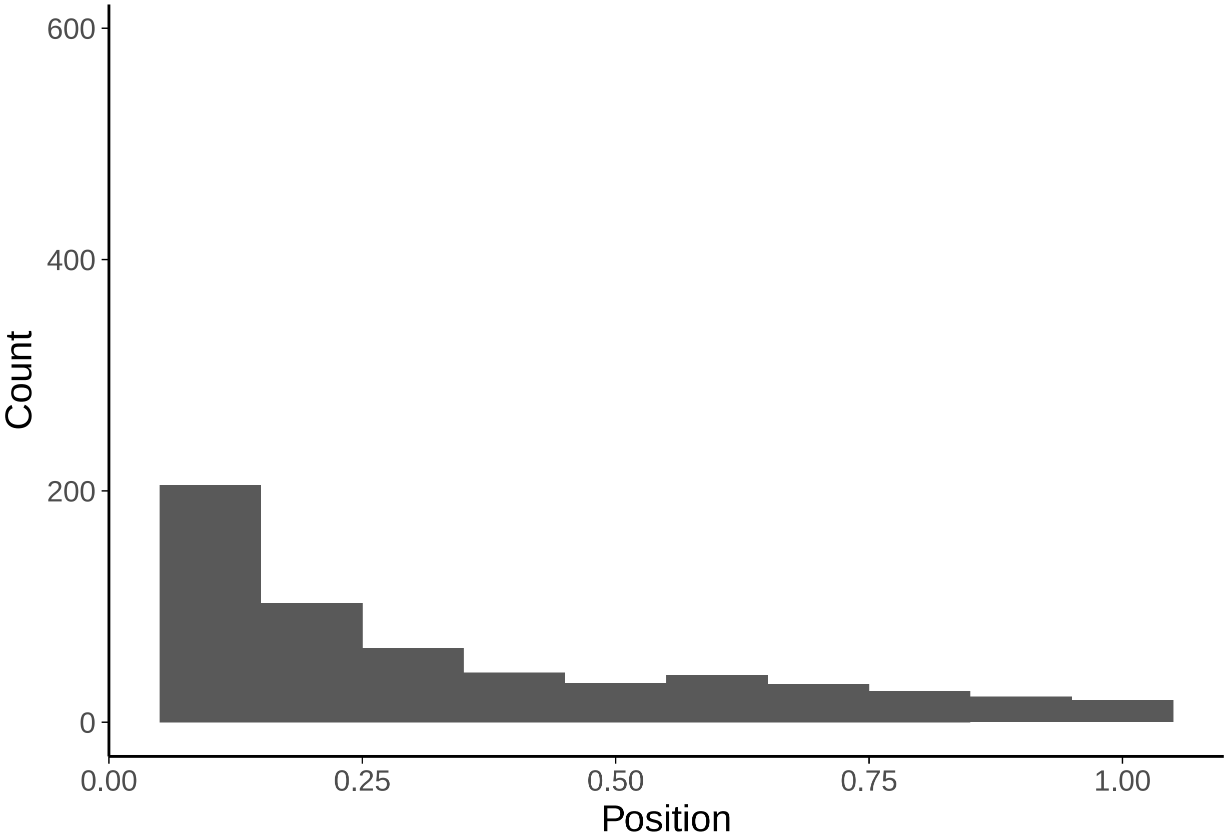


Figure 4.2: Real versus shuffled data fitting error.
Each participant's data were shuffled such that the advisor agreement value for each trial was no longer systematically tied to the trial on which it actually occurred. This process was repeated 9 times, and then joined to the real data. A joint error score was created by summing standardized scores for trust volatility error and weighted selection error. The plot shows the distribution of position of the real data within the combined dataset, with lower numbers indicating lower combined error. For many participants, especially those encountering advisors with particularly high or low agreement rates, this shuffling effect may be quite small, meaning that the difference between shuffled and unshuffled data is less pronounced and that therefore the chance of a better fit to shuffled data is higher.

The resulting coefficients were clustered close to zero for most participants, with a roughly Gaussian distribution, as shown in Figure 4.3. The values are generally greater than zero for both w^a ($t(592) = 2.82$, $p = .005$, $d = 0.12$, $\mathrm{BF}_{\mathrm{H1:H0}} = 2.35$; $\mathrm{M}_{\mathrm{WeightedSelection}} = 2.18$ [0.66, 3.71], $\mu = 0$) and λ^a ($t(592) = 5.10$, $p < .001$, $d = 0.21$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1.4\mathrm{e}4$; $\mathrm{M}_{\mathrm{TrustVolatility}} = 0.01$ [0.01, 0.02], $\mu = 0$). This makes sense because it indicates that participants tended to trust advisors more to the extent that their advice agreed with participants' own initial judgements (λ^a), and that people tended to choose advisors they trusted more over less trusted advisors (w^a).

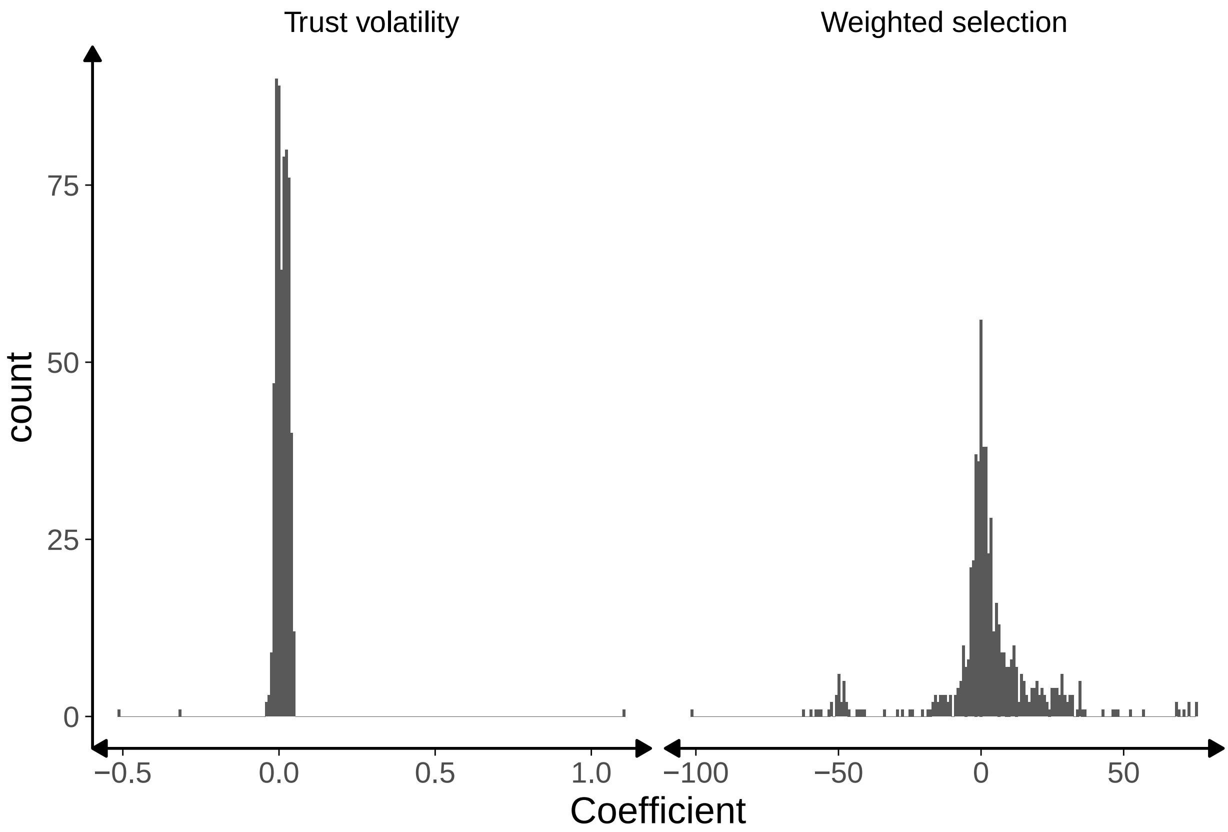


Figure 4.3: Histograms of recovered parameters.

Nevertheless there was considerable variability in these estimates. There was no significant correlation between trust volatility and weighted selection (Figure 4.4).

4.3.2.1 Stability of coefficients

Several participants performed experiments that had two core sections. In each of these sections, participants were familiarised with and then allowed to choose between pairs of advisors who had the same advice profiles but different identities. For example, a participant in a High versus Low advisor accuracy experiment might see two pairs of advisors, with each pair having one High accuracy and one Low accuracy advisor, and all the advisors having different names and visual presentations. This repetition meant that we could perform parameter recovery on the data for each participant separately for each of the repetitions. If the coefficients estimated for each of these repetitions were more similar within a participant than

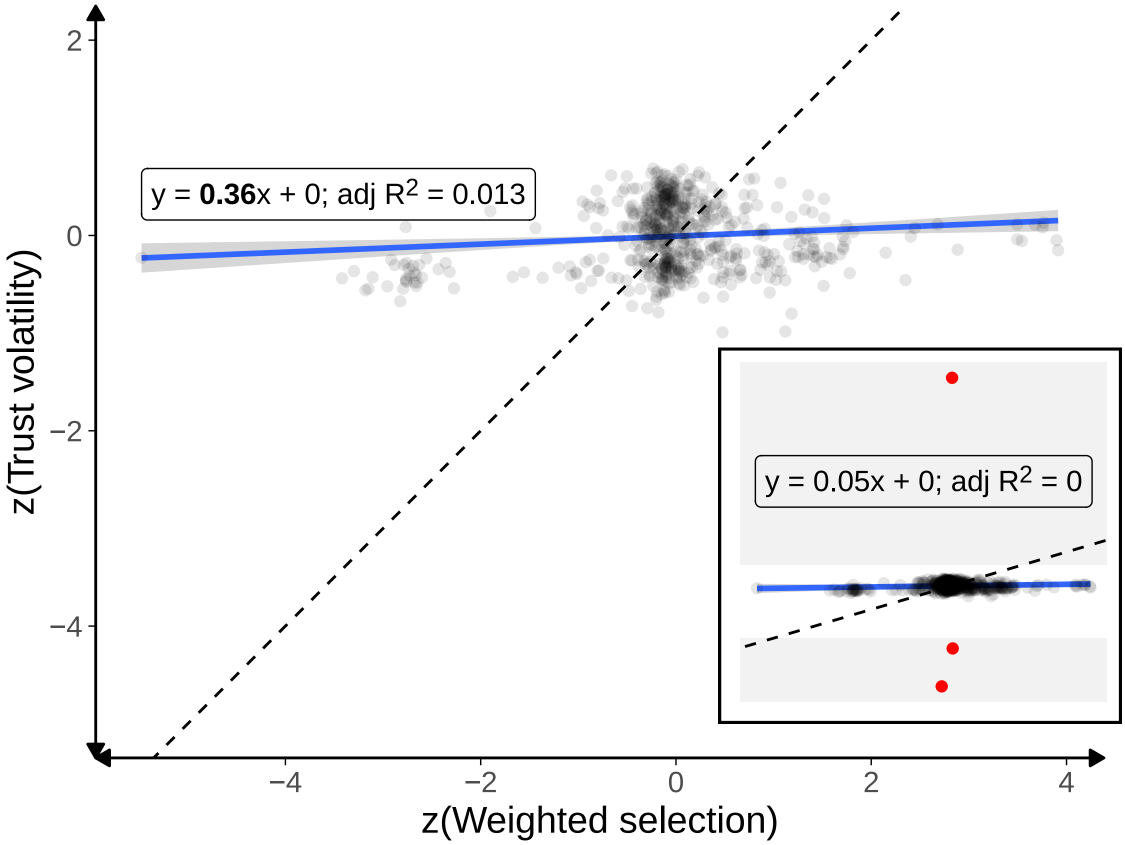


Figure 4.4: Correlation of recovered parameters.
Outliers with extreme trust volatility values ($|z| > 5$) have been dropped, and are shown on the inset plot for context. Each point is a participant, and the solid blue line shows the overall relationship (with 95% confidence intervals shaded in grey). The formula for the line is given in the top-left. Dashed line indicates a 1:1 correlation.

between participants in the same experiment, then that would provide evidence that the parameters we recover are stable properties of a participant.

We performed this analysis and ran a t-test to compare the absolute differences between a participant's first and second estimate versus the absolute differences between a participant's first estimate and the second estimate of another participant in that experiment. The results provided strong evidence that the differences were indistinguishable, indicating that either the coefficients recovered did not represent a stable property of a participant, or that we were not able to estimate them precisely enough to demonstrate this (Trust volatility: $t(434)$ = -0.29, p = .774, d = 0.01, $\text{BF}_{\text{H1:H0}}$ = 1/17.8; $\text{M}_{\text{Difference|real}}$ = 0.04 [0.02, 0.05], $\text{M}_{\text{Difference|shuffled}}$ = 0.04 [0.02, 0.05]; Weighted selection: $t(434)$ = -1.77, p = .077, d = 0.08, $\text{BF}_{\text{H1:H0}}$ =

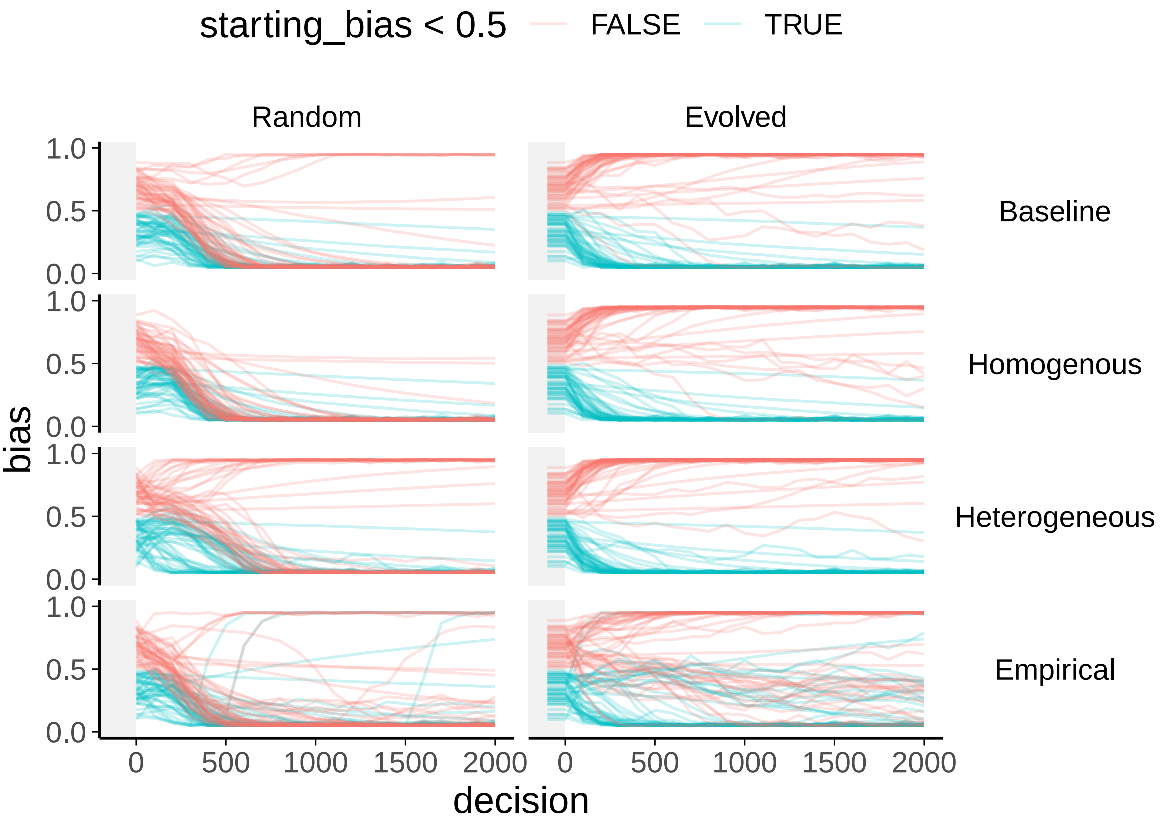


Figure 4.5: Bias evolution within each model.
Each plot shows a model run with a different combination of starting trust weights and weighted selection values. Each line shows the bias of a single agent, coloured according to whether the agent's starting bias was more or less than 0.5. In the Evolved weights graphs the evolution period prior to biases being allowed to shift is truncated (and marked with grey shading).

1/3.93; $\text{M}_{\text{Difference|real}} = 18.73$ [16.28, 21.19], $\text{M}_{\text{Difference|shuffled}} = 20.76$ [18.38, 23.14]). The variation within agents is not implemented in our model; our interest is in heterogeneity between agents. Nevertheless, this result suggests that heterogeneity *within* agents may be an important feature to model in future work.

4.3.3 Validity of the model

A first set of simulations aimed to evaluate basic network behaviours in terms of trust formation. We ran 40 simulations comprising 144000 interactions among agents as defined above. Of interest was the evolution of agents' trust and beliefs (biases) over time, in particular as a function of their starting bias. Agents were

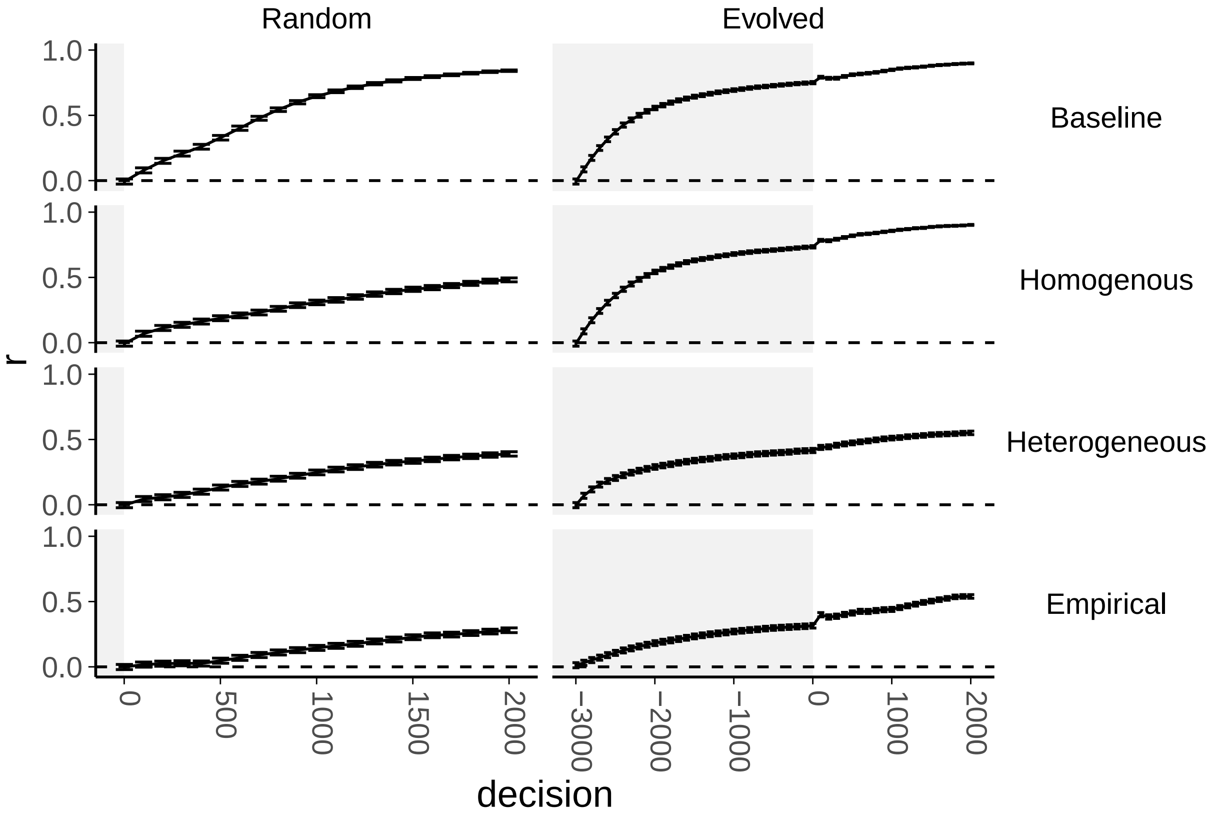


Figure 4.6: Shared bias-weight correlation evolution for each model.
Each plot shows a model run with a different combination of starting trust weights and weighted selection values. Each graph shows the mean correlation between shared weight between agents and trust weights at each decision.

initially defined as belonging to one of two populations with differing (mirrored) biases in their prior expectations about the decisions. Figure 4.5 shows the evolution of agents' beliefs in a single simulation run. The left-hand panels show simulations where agents' trust weights are randomised, while the right-hand panels show simulations where the agents' trust weights are allowed to evolve on the basis of advice exchanges for 3000 interactions prior to biases updating. The simulations are differentiated by the manner in which the agents' key parameters, w^a and λ^a are determined. In the Baseline simulation the value of w^a is zero, while λ^a is drawn from the empirically observed distribution. In the Homogeneous and Heterogeneous simulations, λ^a is again taken from the empirically observed distribution and w^a is set to the mean of the empirically observed distribution (Homogeneous) or drawn

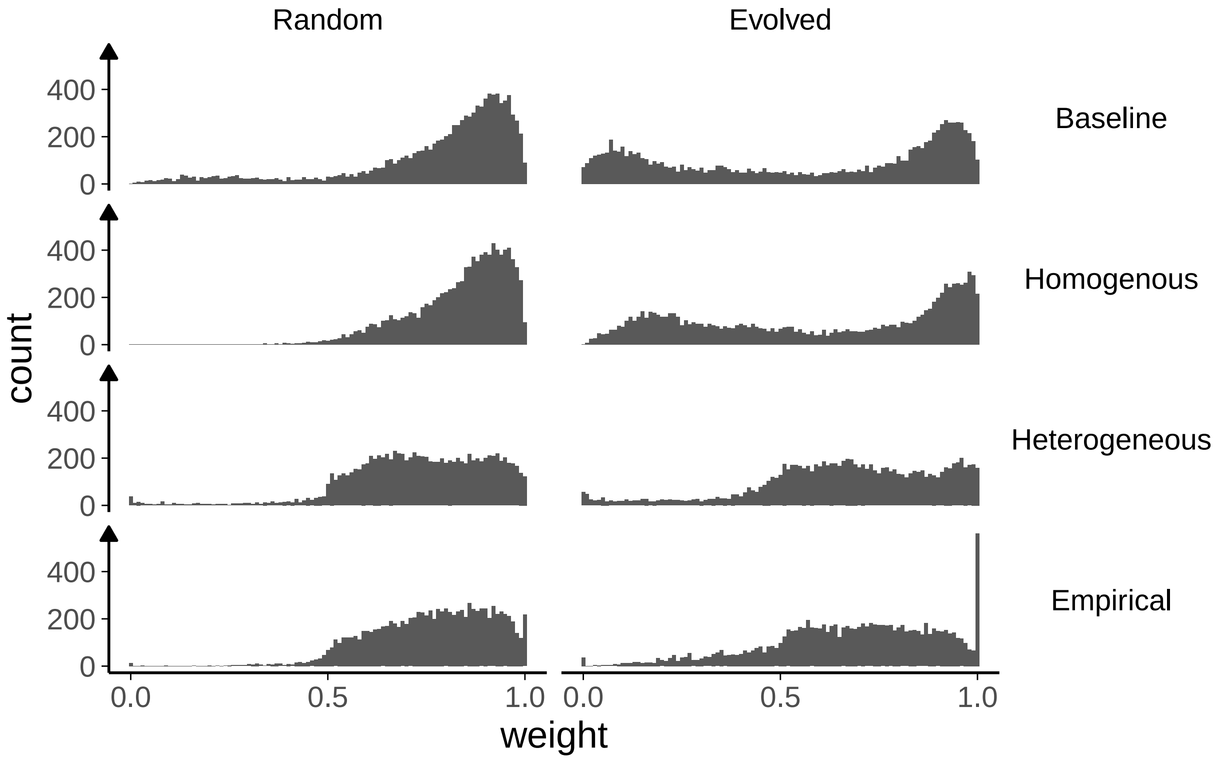


Figure 4.7: Distribution of final weights for each model.
Each plot shows a model run with a different combination of starting trust weights and weighted selection values.

from that distribution (Heterogeneous). In the Empirical simulations, each agent selects a participant at random from the empirical data and uses that participant's estimated values of w^a and λ^a.

The model reproduces key effects of interest. The key patterns of interest are visible even in the Baseline simulations, where agents update trust depending on agreement but do not preferentially select their advisors based on that trust. Firstly, agents reinforce one another's biases, leading to the emergence of extreme biases and echoing the effects in the literature. This can be seen in Figure 4.5 (top row) where the biases of all but a handful of agents reach and remain at extreme values. Whether or not the entire population tends collapse into a single bubble, adopting the same extreme value, or to polarise into two distinct bubbles depends upon whether or not the trust network is already network-chamber-like (left versus right

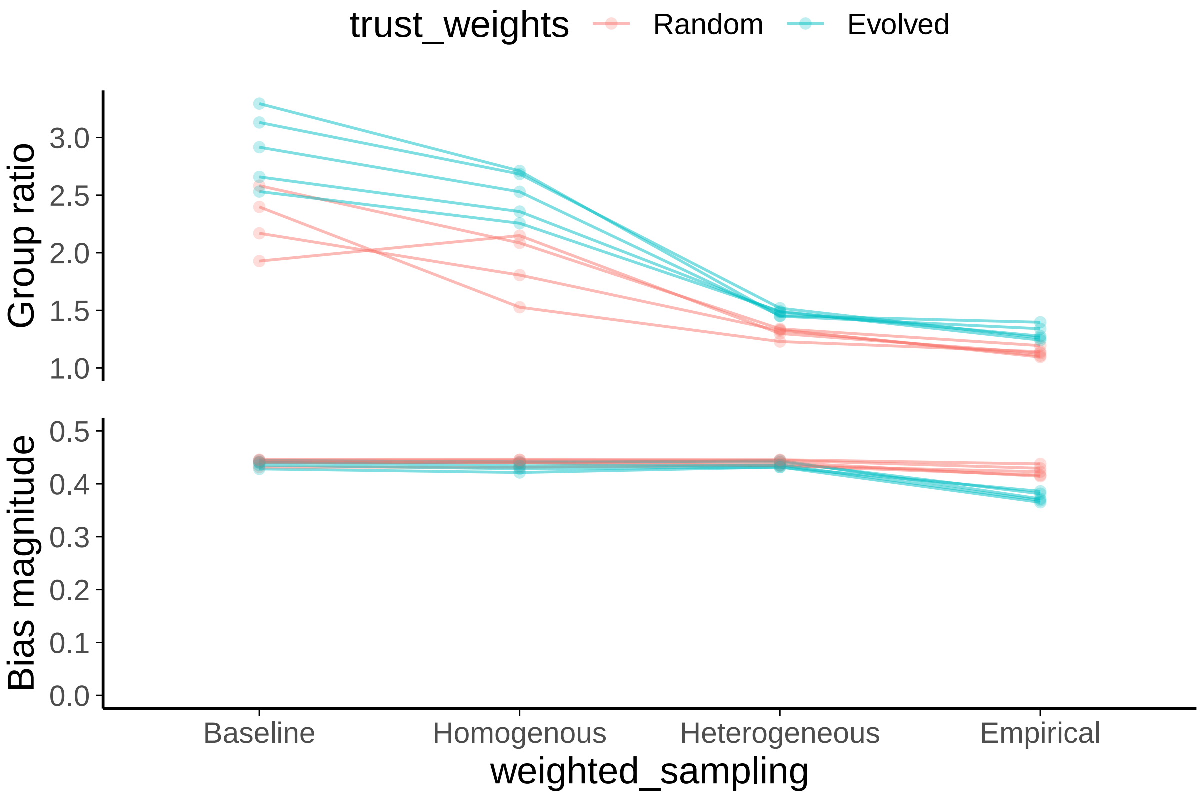


Figure 4.8: Model final decision group ratio and mean final bias magnitude. Each of the trust weights and weighted selection sampling combinations in the figures above was run multiple times with different random seeds. Each of these runs is represented by a single line. Some models have no group ratio because there is only one group at the final decision (because there is a consensus wherein all agents have the same direction of bias), which means that the ratio of weights is undefined and consequently the line is broken. Group ratio is not bounded to the vertical axis values shown.

columns). Secondly, agents develop greater trust in agents who share their bias, as shown by Pescetelli and Yeung (2021). This can be seen in Figure 4.6 (top) where the correlations between agents' shared bias with an advisor and their trust in that advisor rises over time. Agents who have more similar biases to one another ought to arrive at the same conclusion as one another more frequently, meaning that they will agree more frequently. This, in turn, leads to greater trust over the course of repeated interactions, producing the correlation in Figure 4.6 that offers a rough measure of the extent to which agents' trust in one another is dictated by the similarity in their biases. Below, we explore how these features vary as a function of our different sampling strategies for w^a and λ^a.

4.3.3.1 Random versus evolved trust networks

The model shows a substantial effect of whether or not the agents' trust weights have a chance to evolve prior to biases being allowed to shift. The models in Pescetelli and Yeung (2021) implemented this approach, and, like the current model, demonstrated group polarisation where agents selectively sampled from their own groups and thereby increased their bias.

Where the models are started with random trust weights, agents also tend towards extremes, but there is a single consensus for all agents in the model. When this happens, the population is no more accurate than in other models, but *is* more unified. This happens because of differences in the initial correlation between agents' biases and levels of trust. In the random case, the network starting state has no correlation between agents' biases and levels of trust (Figure 4.6, left). As a consequence, early on, agents are equally strongly influenced by others with different biases as they are by others with similar biases. In the case with Evolved trust there is an initial correlation between agents' biases and their levels of trust (figure 4.6, right, unshaded area), meaning that the network begins with correlated biases and trust. This results in segregation into two stable, self-reinforcing sub-networks. This is further reinforced by agents' acquiring negative trust in agents with dissimilar biases (Figure 4.7, top-right). Trust values between 0 and 0.5 indicate that agents believe other agents are predictive of the truth, but that their advice points *away* from the truth. This means that interaction with others serves to *strengthen* rather than weaken biases.

Figure 4.8 shows an overview of this pattern for multiple runs with different random seeds: the evolved trust weight runs tend to show higher group ratios, indicating greater polarisation.

4.3.3.2 Effect of weighted selection

Figure 4.5 (second row) plots the evolution of biases when we add weighted selection of advisors to the simulations. Specifically, in this Homogeneous case, all agents used the mean value of w^a that we observed in the parameters estimated from our participants' data. The agents used this value to weight their selection of advisors. The figure shows that this form of weighted selection has very modest effects of network dynamics. This is likely because the mean value estimated from the Dots task participants was small. There is virtually no effect on bias trajectories, and only a slight one on the distribution of trust weights (Figure 4.7), with fewer very low weights in the weighted selection runs. The most notable effect was in the correlation between shared bias and trust weight where trust weights are random. In this case, the correlation between shared bias and trust weight does not increase as fast or get as high when weighted selection is introduced. The group ratio is nearly always lower when weighted selection is introduced.

These effects are due to agents becoming increasingly unlikely to sample advice from advisors as the trust in those advisors decreases, which in turn means that only the higher trust weights are likely to be adjusted. In effect, when weighted selection occurs the lower tail is ignored, meaning that trust remains unchanged whether the shared bias is moderate or low, and thus reducing the correlation between them. In this sense, the effects of weighted selection itself are more a feature of the operationalisation than the underlying model. The models offer agents a choice of advisors at each step, and leave the trust weights of unsampled advisors static. Other plausible modelling choices might not show this pattern, for example if agents' trust in unsampled advisors gradually decayed to a neutral value; or if the overall level of trust in potential advisors were normalised such that an increase (or decrease) in trust in one agent meant that trust in all other agents underwent a slight decrease (or increase), perhaps proportional to their current trust values.

4.3.3.3 Heterogeneity

Allowing heterogeneity in weighted selection coefficients has a more pronounced effect. Whether the heterogeneity comes from sampling from the distribution of estimated participant coefficients (Heterogeneous) or sampling directly from the empirically-estimated values themselves (Empirical), varied weighted selection values create subsets of agents who polarise away from the consensus value formed when starting trust weights are random. In Figure 4.5 this is visible in the left-hand column: in the bottom two panels there is a small cluster of lines that trend towards 1 rather than 0, and that remain there throughout the simulation. Additionally, there is more heterogeneity in the agents' biases, with several agents retaining moderate biases (lines that do to tend to extremes in Figure 4.5).

There are differences, too, between the Heterogeneous and Empirical sampling approaches. Compared to the Heterogeneous sampling, the model with Empirical sampling shows more moderate biases and more extreme weights (Figure 4.8). This apparently contradictory finding is explicable because agents in the Empirical runs include both positive and negative weighted selection coefficients. This leads to the peaks of weights at both the 0 and 1 end of the histogram for evolved weights. The negative weighted selection values can also drive agents to sample advice that is likely to disagree with them, moderating their biases as seen by a larger minority of agents with moderate biases in the Empirical run than in other runs. Madsen, Bailey, and Pilditch (2018) termed agents with this advice-seeking strategy 'Socratic' because they aimed to engage with others of a different opinion. This sampling approach may be behind the fact that many more agents switch the direction of their bias in the Empirical simulations than in the other simulations.

Finally, Figure 4.7 illustrates that a substantial minority of weights reach the maximum value in the Empirical simulations. This suggests that many agents adopt a favourite advisor and stick with that advisor because they tend to agree more often than not, meaning they retain their relative advantage over other advisors. This likely happens more in the Empirical simulations because more of the agents have extreme values of w^a.

4.3.4 Friends in strange places

Combining the results above, the models raise the interesting observation that consensuses form much more readily where issues arise on which people are likely to encounter different opinions within their circle of trusted informants. This can be seen in Figure 4.5: in the Random case, where agents are likely to have trusted potential advisors who have very different biases, the population tends to drift together towards a single majority consensus. It is important to note that, at least in these models, the extreme that is selected by consensus is not more accurate than the extreme that is neglected by the consensus (and is less accurate than the starting positions), but it is nevertheless intriguing to see dynamics that resemble broad versus insular consultation.

Heterogeneous weighted selection can prevent a full consensus forming where individuals have access to trusted advisors of a variety of different opinions. This was evident in the Random starting trust weights simulations in Figure 4.5: minority groups emerge and reinforce their own biases to adopt the opposite extreme to the consensus favoured by the majority; and several agents continue to occupy the middle-ground long after the majority and minority groups have adopted their extreme positions. This pattern in the Heterogeneous and Empirical simulations emerges because those who have a strong preference for agreement rapidly disregard the advice of their previously-trusted advisors who offer a different opinion. These agents form their own polarised minority, consulting and advising one another consistent with their bias (as the majority are also doing, just with the other bias). This pattern differed from the Baseline and Homogeneous simulations, illustrating the impact of inter-individual variability on network level behaviour.

4.3.5 Consistency of the models

Each model was run 5 times with different random seeds to check which features were consistent across runs, indicating the level of stochastic uncertainty (Bruch and Atwell 2015). The features described above were all consistent across runs, as can be seen by observing the variability in Figure 4.8 in both of the outcome variables§4.2.3.4.

For parameter sets in which all agents' biases tend towards the same extreme, which extreme was favoured varied according to the random seed used. This stochastic is akin to placing a ball on a gabled roof: tiny variations in the initial conditions affect which way it will roll, but it will always roll down one pitch or the other.

Similarly, in some of these models, adding in weighted selection switches which bias extreme is adopted. This effect is not consistent between runs with different random seeds, and is an outsized effect of minor differences to early states, analogous to a breath of wind nudging the ball in the previous example one way or another.

Output in the style of Figures 4.5, 4.6, and 4.7 for different random seeds can be visualised by tweaking the source code for this chapter.

4.3.6 Summary

The results shown in Figure 4.8 capture the key observations of this work. Weighted selection ($\mathrm{w}^a \neq 0$) is not necessary for the formation of echo chambers. Indeed, perhaps counter-intuitively, we found some tendency that weighted selection can reduce these effects, because unsampled potential advisors are not learnt about (and therefore are not distrusted over other unsampled potential advisors). Heterogeneity complicates and enriches this picture in interesting ways, supporting the various minority strategies visible in Figure 4.5.

4.4 Discussion

As with previous simulation-based approaches§4.1.2, this work showed that echo chambers and polarisation can occur in networks of agents using heuristics that appear from the agent's perspective to be rational. In the absence of an objective source of information, using your own opinion as a proxy for the truth is a viable strategy under a range of broadly-true assumptions (Pescetelli and Yeung 2021). Likewise, choosing to hear advice from more trusted sources, and choosing to hear advice you are more likely to take (often the same thing in our models and in real life), is a reasonable strategy. The emergence of collective irrationality from the interaction of rational individuals is similar to the well-known phenomenon of the

Tragedy of the Commons (*Tragedy of the Commons* 2021), in which individual rewards purchased with distributed costs results in everyone becoming poorer.

4.4.1 Weighted selection

The models indicate that weighted selection (w^a), the extent to which agents prefer to receive advice from those whom they consider most trustworthy, has relatively little effect on model dynamics, at least at the level estimated from participant behaviour. The results from more heterogeneous runs discussed below indicate that agents with high weighted selection values do behave in notably different ways. During development we observed that higher weighted selection values (above 10) tended to accelerate the polarisation of the model.

We must note, however, that the weighted selection parameter is used somewhat differently in the model than its analogue was in the behavioural experiments from which it was derived. In those experiments, participants were familiarised with two advisors and then offered the choice between them. In our model, there are numerous potential advisors available and the weighted selection parameter weights the relative probability that each of those advisors is selected. This means that the parameter values derived from our empirical observations may not correspond closely to the values we would have obtained had the behavioural task been a closer match to the simulated one. While our models illustrate that variation in weighted selection plays a part, therefore, we do not make strong claims that the specific values used are meaningful, even in the Empirical case.

4.4.2 Heterogeneity

Most work using agent-based models, in particular the work on the topic of advice-taking (Pescetelli and Yeung 2021; Madsen, Bailey, and Pilditch 2018), has used agents whose key properties either do nor vary or vary in a predictable way (e.g. sampling from a standard distribution). Here, we explored the effect of sampling agents' properties directly from a collection of values estimated from participants in our behavioural experiments. This variation potentially allows us to see the effects of minority strategies that may be hidden where extreme values, or particular combinations of values, are smoothed away by the use of a mean or normal distribution to describe the entire population.

Beyond the basic effects of including weighted selection, some models used weighted selection coefficients that were heterogeneous across agents. In both the Heterogeneous and Empirical models, heterogeneous weighted selection coefficients led to the emergence of hold-out groups – small subsets of the population who formed their own minority consensus at the opposite extreme to the majority consensus.

Where the models had evolved trust networks prior to beginning their bias updates, polarisation was common, and heterogeneous weighted selection coefficients sped up this process. They also appeared to make it harder for agents to maintain low biases, although many of the agents in the Empirical model managed to maintain these moderate biases for the duration.

These effects, the presence of hold-out groups and the speeding up of polarisation, suggest that other models of advice dynamics may underestimate the speed with which polarisation occurs while overestimating the degree to which it takes over.

The hold-out groups captured in the heterogeneity models is an observable feature of opinion networks in the real world. For each consensus opinion there is likely to be an outspoken minority who cling resolutely to the opposite opinion. The models presented here cannot tell us how such heterogeneity comes about, but they do underscore its importance in the wider ecosystem of opinion interactions, and invite questions concerning its origins. One suggestion we raise here is that being contrarian may be a stable sub-strategy within an environment where consensus dominates.

We acknowledge that the specific values of parameters estimated here may not be a good description of people's tendencies in the kind of task simulated: because of differences between the binary advisor choice situation faced by the participants and the multiple advisor choice situation faced by the agents; or because of difficulties in accurately estimating the coefficients for some participants; or because the the participant population may deviate somewhat from humanity in general. Despite this, we are confident that individual variation in these advice-taking and source-selection propensities exists, and that it matters in interesting ways for network dynamics. Idiosyncrasies certainly occur in confidence ratings (Ais et al. 2016; Navajas et al. 2017), and likely in advice-seeking and advice-taking behaviour, too (Soll and Mannes 2011; Pescetelli, Hauperich, and Yeung 2021). Additionally, there may be variation in confidence across personality and psychiatric dimensions, and computational models are increasingly used to explore the implications of these kind of variations in psychiatric disorders (Ais et al. 2016; Navajas et al. 2017; Rouault, Seow, et al. 2018; Moses-Payne et al. 2019; Hauser et al. 2017).

There are numerous potential causes for inter-individual variation in advice-taking and advisor choice behaviour. In behavioural experiments in our lab, we often see a small minority of participants will take no advice at all – never seeking it or changing their answer if they receive it – and these participants typically report very high subjective confidence. It is plausible, therefore, that whatever causes the heterogeneity seen here, it has some relationship to confidence. Advice-seeking has been argued to serve several purposes, including increasing accuracy (Soll and Larrick 2009), diffusing responsibility (Rader, Larrick, and Soll 2017), and social inclusion (Mahmoodi et al. 2015). Correspondingly, individual variation in the importance of any of these components may alter the tendency to seek and use advice.

4.4.3 Considerations

There are several considerations which should be borne in mind while interpreting the results of the models. These considerations fall into several different categories, from technical constraints to limitations on generalisability.

4.4.3.1 Stability of coefficients

We were unable to demonstrate that independent estimates of participants' coefficients were more similar than would be expected by chance. This means that we are unable to rule out the suggestion that there is as much variation in advice-taking and -seeking behaviour within people as between them. In our models, agents have set individual propensities governing how they take and use advice. The literature suggests that people have idiosyncrasies governing confidence and its relationship with advice-seeking (Ais et al. 2016; Navajas et al. 2017; Pescetelli, Hauperich, and Yeung 2021), but it is nevertheless plausible that these idiosyncrasies vary dramatically across time and context. Further simulation work might investigate the effects of heterogeneity within rather than between agents.

4.4.3.2 Context

Polarisation and homophily are not inherently bad features. In some discussions (e.g. debates about whether slave ownership is morally permissible) we can be glad that near-universal opinion reigns on the issue, despite its once being a contested issue. In others (e.g. discussion concerning the extent to which the state should be involved in individuals' lives) it is helpful to have a range of opinion to maintain the diversity of approaches (Diamond 1998, p482).

This means that the context must be considered when evaluating the implications of these models. The mathematics and computational results will remain the same, but the real-life implications would change dramatically depending upon the context which the models were chosen to represent[1]. The model presented here could be considered a model of many different kinds of social interactions, from those where consensus is desirable, such as norm formation, to those where consensus is subordinate to accuracy, such as group and individual decision-making. The implications of polarisation and consensus formation appear beneficial in some cases and harmful in others. The model is sufficiently specified that it is unlikely to fit many interpretations outside of the domain of social influence, but there are plausibly areas in other disciplines which may have a different gloss on this or a very similar model implementing a combination of Bayesian and weighted average integrations and updates.

[1] In one exchange in the literature, Jones (2007) view a model as showing drinking behaviour over a day, while the authors, Mezic et al. (2007), see each tick in the model as representing a day!

4.4.3.3 Generalisabilty

There are two major question marks concerning the generalisability of the model results. The first is to do with the assumptions of the model, specifically whether the recovered parameters are features of individuals that remain stable across contexts; and the second to do with the results, specifically whether people polarise to the extent indicated by the model.

The first of these generalisability concerns is part of what Bruch and Atwell (2015) term 'input uncertainty'. This is uncertainty over the inputs of a model – in the present case the coefficients recovered from participant data. For each of our participants, we fitted our underlying mathematical model to their behavioural data so that we obtained coefficients that minimised the fitting error across both parameters simultaneously. The existence of such minima, though not unique for every participant, is an observed fact. What is less clear, however, is whether these coefficients represent a stable and enduring feature of the participants or whether they are largely arbitrary – an ephemeral product of time, psychological context, and task. It is a basic assertion of the model that the coefficients observed for an individual do not change during the model: the size of the trust update is the same at decision 1000 as it is at decision 1, and likewise the strength of preference for hearing from an advisor who is more likely to agree. The results of the model can inform us about what might happen if these properties are stable, although even then it is unlikely that the values themselves are exactly accurate to some posited underlying propensity. The constant model parameters were tuned during development to give sensible results given the mean value of weighted selection and the mean and standard deviations of trust volatility, so the performance of the model should not be taken as strong evidence for the existence of the posited psychological features.

The second generalisability concern is in the conflict between the model behaviour and the results of research into the level of polarisation and consensus in society. Research demonstrates that most people occupy central positions on political issues, and that while political polarisation varies across operationalisations and across time, it is not becoming increasingly pronounced in recent years (Park 2018; Perrett

2021).[2] The interpretation of the simulations in terms of real-world politics is dependent on what we think models are showing. We may consider them as showing the drift leftwards of political opinion over time on moral issues (Park 2018), i.e. as models of consensus formation. Alternatively, we may view them as suggesting that moderate areas are uninhabitable, which is clearly not the case on most issues. The model does allow for the emergence of a middle-ground (accurate) consensus where feedback rates are higher, however, although what constitutes feedback in real-world political scenarios is debatable; we pursue this work in part because we regard real-world decisions as frequently lacking in timely actionable feedback. The precise nature of political polarisation, whether and to what extent it occurs and what modifies it, is an open question and the subject of an entire literature in its own right.

4.4.3.4 Known psychological aspects which are left out

The modelling work in this chapter follows the recommendations of Bruch and Atwell (2015), who suggest that modelling is best used to characterise the effects on system dynamics of a particular effect derived from empirical observation, and Jackson et al. (2017), who advocate for simplicity where possible. Consequently, many decision-making and advice-taking phenomena reported in the literature are left out. East et al. (2016) argue that an agent-based model only achieves sufficiency by including "all relevant processes and conditions". To allow the reader to determine whether the omissions are important, this section briefly covers those aspects which are left out of the models.

Advice is the key feature of the model rather than the task, and so the agents' performance on the task does not vary over time (they do not get better or worse, and they do not get bored). Advice in the model, as in the experiments, is a one-way transaction. The agents giving advice are unaffected by having given the advice, except by the roundabout route of potentially causing other agents to give advice in the future which is likely to make them appear trustworthy to the agent. This means that there is no deliberate reciprocity to advice-seeking

[2] The paper by Park (2018) covers politics of the USA prior to the controversial tenure of President Donald Trump.

behaviours (Mahmoodi et al. 2015), and agents do not alter their advice to appear more palatable to others or so that they are selected as advisors more frequently (Hertz, Palminteri, et al. 2017; Hertz and Bahrami 2018).

The agents use advice rationally, given their estimation of the trustworthiness of their advisor. A plethora of research has shown that people do not do this in practice. Most research has documented processes which result in the under-weighting of advice (e.g. Yaniv and Kleinberger 2000, and many others), while some has indicated conditions under which people are over-reliant on advice (Schultze, Mojzisch, and Schulz-Hardt 2017; Dietvorst and Simonsohn 2019). This literature is reviewed in the chapter on the Context of Advice§5. Capping the extent to which agents take advice to within a relatively narrow range between a minimum and maximum might alter the simulations quite extensively: the simulations where agents consider advisors as anti-predictive would no longer occur; and the time taken to differentiate groups and form consensuses would likely increase because the range of trusts would be lower. The effect of weighted selection, and the effects of heterogeneity in trust volatility and weighted selection values, would likely also diminish because the range in which they can operate is narrower.

Lastly, and in keeping with the simplicity aim advocated by Jackson et al. (2017), the reader is encouraged to consider the differences that including empirically estimated coefficients in the model makes to the overall dynamics, and draw their own conclusions as to the necessity of including these in similar models. We feel the effects are interesting, although, depending on the purpose of the model in question, they may not exert sufficient influence on the dominant network dynamics to be a necessary inclusion.

4.4.3.5 Differences between experiments and model

For the most part, the task in the model was analogous to that facing the participants in the Dots task experiments. The key difference in task terms was in the choice of advisor. Participants in the experiments were offered a choice of two advisors (with whom they had recently had blocked practice), whereas agents in the model chose freely from all other agents in the model and begin with a knowledge of how trustworthy they consider each of those potential advisors.

4.4.3.6 Technical constraints

Agent-based models are computationally expensive to run because each step in the model depends upon the step before. This poses some challenges for sources of uncertainty Bruch and Atwell (2015) term ‘stochastic uncertainty’ and ‘model uncertainty’. Stochastic uncertainty refers to the consistency of a model’s dynamics when different random numbers are selected in places where they are required. Model uncertainty refers to the consistency of a model’s dynamics to variation in parameters other than those selected for exploration. Stochastic uncertainty has been addressed to the extent it is possible given temporal and computational constraints by running each model 5 times with different random seeds.

Model uncertainty is more difficult to address. There are certainly values for model parameters which have substantial effects on dynamics. A trivial example is that when agents are given feedback on every decision they tend towards very low bias magnitudes. Feedback and other parameters were explored to some extent during the development of the model, as documented in Constant parameters§4.2.3.2. Despite this, readers should be aware that the parameter space is very large, and there are likely to be substantial interaction effects which were not found during development.

Many models in the literature use very large populations. While the models discussed above§4.1.2 had smaller sizes, the largest being 1000 agents, it is not uncommon to see models with 2-3000 agents (Nowak, Szamrej, and Latané 1990; Duggins 2017; Song and Boomgaarden 2017). The size of the current model runs seems sufficient to illustrate the effects of interest, but it is possible that different effects may emerge in larger populations. Smaller populations examined during testing showed less consistent effects. For very large models the underlying package code would have to be adjusted and the data extracted from model runs sampled much more sparsely.

4.4.4 Conclusion

We used agent-based models to explore the implications of selective advice sampling. The experiments in Chapter 5 showed that, people use their own opinion as a proxy for the truth when evaluating advice, although the extent to which they do this varies dramatically from person to person. The limitations of the simulations notwithstanding, they show that simple models of interaction on the basis of homophily are enriched by including the kind of variety observed in human participants.

Advisor choice on the basis of similarity of opinion does seem to occur, although it depends on the absence of feedback and does not appear to be universal. There are enough people for whom it does occur, however, for it to disrupt the uniform dynamics of simulated networks of information exchange, allowing stable bubbles of minority opinion and individuals who preserve more moderate opinions.

In the chapters that follow, we turn our attention from exploring how individuals seek and take advice to the reasons why people take less advice than normative models predict. I argue that this phenomenon, known as egocentric discounting, may be due to prior expectations beyond the scope of the experimental tasks used to study advice-taking behaviours, and that these expectations are warranted given typical advice-taking conditions.

5
Context of advice

5.1 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 that 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 other participants. In a typical design, the judge offers an initial estimate for some decision, e.g. the total value of coins in a jar of change, 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 by a 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 (Soll and Larrick

2009; Bahrami et al. 2010). This is analogous to multi-sensory integration (Ernst and Banks 2002; Körding et al. 2007), and many other cognitive processes argued to be modelled on Bayesian integration, such as those listed in Section 2 of Colombo and Hartmann (2015). 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 (Equation 1.7). 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 accurately combine estimates that do not include their own opinion (Yaniv and Choshen-Hillel 2012), even adjusting for differences in ability between advisors (Soll and Mannes 2011). Similarly, Trouche, Johansson, et al. (2018) showed that when advice and initial estimates were surreptitiously switched, participants discounted their own initial estimates in favour of advice, 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 that 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.

5.2 Manipulations affecting egocentric discounting

A sizeable body of research has been conducted into egocentric discounting, using 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.

Many of the experiments detailed below have not looked at egocentric discounting itself, but have instead looked at changes to the weight given to advice in integrated decisions. While egocentric discounting and advice weighting are inversely related, establishing the degree to which advice has been discounted requires a reference value from which its weighting has deviated. Many experiments do not calculate this normative value, nor even give sufficient details to establish a reference value, either an objective normative value or an expected value from the perspective of the participant.

One obvious reference value is the case where equal weight is given to one's own opinion and the advice received. This is often the value implicitly or explicitly stated. However, equal weighting is only normatively prescribed if advice is exactly as reliable as one's own initial estimate.

It should be borne in mind, therefore, that depending on the circumstances, very high or low levels of advice weighting might not correspond to very low or high levels of discounting. When Schultze, Mojzisch, and Schulz-Hardt (2017) provide participants with advisory estimates from a random number generator, ascribing any weight at all to the advice is suboptimal. Conversely there are many experiments in which expert advice is provided, and in these experiments weighting advice evenly corresponds to egocentric discounting because the advisor's estimates are more accurate on average.

As a consequence of the uncertainty about which normative weighting strategy is required by each experiment (and sometimes this strategy is different from the perspectives of the researcher and the participant), egocentric discounting is primarily examined here in terms of changes in advice weighting. Where advice weighting diminishes, egocentric discounting is said to increase, without specific comment being possible on the exact level of discounting on display. In studies where normative strategies can be determined (e.g. Yaniv and Kleinberger 2000), advice weighting is below that predicted by the normative strategy, indicating egocentric discounting.

5.2.1 Task properties

The properties of the task chosen can affect the levels of egocentric discounting. Task difficulty is a major factor, perhaps mediated by the judge's confidence, but broader features also play a role, including how advice is provided and whether unified estimates are required at any point.

5.2.1.1 Task difficulty

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 and Du (2018) 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. Yonah and Kessler (2021) included a condition where there was no objectively correct answer in a random dot motion display, and saw increased willingness to take advice compared to conditions where participants were on average 67% correct. They also found that participants were less likely to seek advice where they rated their performance on the task as more competent, i.e. where they experienced the task as easier.

5.2.1.2 Judge's confidence

Wang and Du (2018) saw full mediation of their difficulty manipulation by participants' confidence on the task, while Gino and Moore (2007) 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, and Schweitzer (2012) used anxiety manipulations to decrease judges' confidence. In both cases, partial or full mediation through confidence occurred such that participants' higher confidence in themselves and their decisions was associated with greater egocentric discounting. In many other experiments, including other experiments in Wang and Du (2018), confidence is not manipulated but is still associated with greater egocentric discounting. Using a similar methodology, but a different analytical approach, Moussaïd et al. (2013) observed that highly confident participants rarely updated their views following advice.

5.2.1.3 Judge-Advisor System structure

More complex task designs, in which reflection and discussion are encouraged, can reduce discounting. Minson, Liberman, and Ross (2011) and 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 reduced, but still evident in this process, as it was in van Swol (2011), which used a traditional Judge-Advisor System paradigm where advice was delivered face-to-face. Liberman et al. (2012) 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 more than by justifying estimates or exchanging bids. Schultze, Mojzisch, and 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.

5.2.2 Advice properties

Several features of advice itself have been explored: the confidence of advice, its similarity to the initial estimate, whether it is solicited, and the amount of it provided.

5.2.2.1 Confidence of the advice

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 and Larrick (2009) measured the confidence of advice and saw that higher advice confidence was associated with higher influence of advice. Moussaïd et al. (2013) 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. More generally, the assumption that more confident advice will be more accurate is known as the 'confidence heuristic', and has been investigated as a phenomenon in its own right (Pulford et al. 2018; Price and Stone 2004; Bang et al. 2014).

5.2.2.2 Similarity of advice to the initial estimate

A 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.[1] The evidence concerning the effects of advice distance on advice influence is equivocal. Some studies show that advice is less influential the further it is from the initial estimate, while other studies show a greater influence of more distant advice. Other studies have indicated that the relationship is quadratic: low weight is assigned to advice which is too near or far from the initial estimate, and a greater weight assigned to advice which is in the moderately distant.

Several studies have provided evidence for a simple agreement effect whereby advice that is nearer to the initial estimate is more influential. 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, and Ross (2011) found that more distant advice

[1] Note that the influence of advice that is perfectly in accord with an initial estimate is undefined when using Weight on Advice§2.2.1.2 measures.

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.

The opposite effect was demonstrated by Hütter and Ache (2016). They found consistently higher influence for advice that was further from the initial estimate, both for single pieces of advice and for integrating multiple pieces of advice.

A non-linear, U-shaped relationship between advice distance and advice influence has been shown in other studies. Moussaïd et al. (2013) asked participants to estimate answers to a variety of questions and give confidence ratings, both before and after receiving another person's initial estimate as advice. They identified a three-zone structure to the influence of advice according to the distance between the initial estimate and the advice. Similar advice fell into the 'confirmation zone', where opinion was unchanged but confidence increased; moderately distant advice fell into an 'influence zone' where opinion changed to accommodate the advice; and distant advice was generally ignored. Likewise, Schultze, Rakotoarisoa, and Schulz-Hardt (2015) showed in an elegant series of Judge-Advisor System experiments that relationships between egocentric discounting and advice distance were U-shaped. Schultze, Rakotoarisoa, and Schulz-Hardt (2015) also showed that confidence in final decisions was dramatically boosted by near advice, and that confidence gains decreased sharply with distance, consistent with the Moussaïd et al. (2013) account.

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, and 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, and 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.

The level of agreement may act as a cue to the plausibility of both the advice and the initial estimate. Somewhat counter-intuitively, this can lead to agreeing advice being assigned less weight than we might expect because of its bolstering of the confidence in the initial estimate.

5.2.2.3 Solicitation of advice

The extent to which advice is discounted may also be related to whether advice is wanted. This is hard to disentangle from the effects of task difficulty, because people are more likely to seek advice when they find the task more difficult to do (and hence their confidence in their response is lower).

Gino and Moore (2007) compared advice-taking behaviour across two experiments using a task in which participants had to estimate people's weight from photographs. In one of these experiments participants received advice automatically and in the other participants had the option of clicking a button to receive advice. Participants opted to receive advice on almost all trials, including in an easy condition, and no differences were found in the extent to which advice was used between the compulsory and optional advice experiments. The very high rates of advice seeking in the optional experiment suggest that participants in the compulsory advice experiment may have been very welcoming of the advice due to the difficulty of the task. Gino (2008) showed that more expensive advice was sought less frequently but used more heavily, that expensive advice was used more heavily than free advice even when both are compulsory, and that paid-for advice was used more heavily than when the same advice was given for free as the result of a coin-flip. This study packaged questions together in blocks, and participants purchased advice for a whole block at once. This procedure means that the solicitation of advice is decoupled from the difficulty of the question on a trial-by-trial basis, although it is still likely that those participants who found all the questions in a block more difficult were more likely to solicit advice. In the real world, price is often an indicator of quality, albeit an imperfect one, and solicitation is likely to act as a proxy for confidence, which is again related to task difficulty: participants in the experiments may have taken more advice because they were less sure on the blocks in which they sought advice. Hütter and Ache (2016) found that participants opted to see a large number of advisory estimates in a calorific content estimation task when allowed to sample ad lib, although the influence of the advice was low.

5.2.2.4 Number of advisory estimates

Yaniv and Milyavsky (2007) presented participants in a Judge-Advisor System the advice of 2, 4, or 8 advisors and saw that discounting behaviour did not lessen as the number of advisors rose. Hütter and Ache (2016) saw levels of advice usage for multiple pieces of advice which were relatively similar to levels of advice usage for a single piece of advice. In other words, people integrating their own opinion with two advisory estimates tend to integrate the advisory estimates and then treat them as a single piece of advice for integrating with their own opinion.

Minson and Mueller (2012) assigned participants to be members of a dyad or to act alone, and crossed their design such that some dyads received another dyad's estimate as their advice and some received an individual participant's estimate as their advice, while some individual participants received advice from dyads and some from other individuals. The advice was labelled as having come from an individual or a dyad. Despite the optimal policy being to weight initial estimates and advice according to the number of judges and advisors, weights were almost identical for individual advisors and dyad advisors, meaning that advice that represented the average of two advisors was treated as a single estimate.

This may be an artefact of presentation because these studies presented multiple estimates simultaneously as a list, or as a single combined estimate. If this is a real phenomenon, however, it is a critical bias: while sensible motivations for favouring one's own opinion over another's will be posited below, it is far harder to argue that the weight assigned to one's own opinion should remain the same whether being integrated with one other estimate or ten other estimates.

5.2.3 Advisor properties

A common strategy in advice-taking experiments is to manipulate the properties of the advisors (either within- or between-participants). Expertise is often manipulated, but some research has investigated factors such as the pre-existing relationship between judge and advisor, whether advisors are human or algorithmic, and whether advisors have a clear conflict of interest.

5.2.3.1 Expertise of advisors

By far the most frequently manipulated property of advisors is their ability to perform the task in question, known as their expertise, and identifiable with the Mayer, Davis, and Schoorman (1995) dimension of ability§1.1.2.1. Advisor expertise can be communicated to participants in various ways, which can be broadly categorised into 'showing' approaches in which participants build up a picture of advisors' performance over a series of instances, and 'telling' approaches where participants are presented with a summary of an advisor's performance. As described by Equation 1.7, the ability of the advisor to perform the task (relative to the judge) alters the optimal level of advice-taking according to the normative model. This means that discounting may still occur even when the advisor's estimate is weighted more highly than the judge's.

Yaniv and Kleinberger (2000) showed in a series of historical date estimation experiments that better-performing advisors were more influential than worse-performing advisors, especially where feedback was provided on judge's final decisions but also where it was not. Gino, Brooks, and Schweitzer (2012) used a coin-jar estimation task where participants were shown the advisor's past performance to demonstrate that people give the same advice more weight if it comes from an advisor with a history of good performance, although this effect was not visible in an anxiety arm of the experiment. Rakoczy et al. (2015) saw that 3-6 year-old children took advice from advisors who had named animals correctly more seriously than advice from advisors who acknowledged their own ignorance in a version of the Judge-Advisor System adapted for young children.

Sniezek, Schrah, and Dalal (2004) saw greater dependence on advice provided by specially-trained 'expert' peers, but only where the proportion of reward money for accurate judgement paid to the advisor had been decided *before* advice was provided. Soll and Larrick (2009) observed greater influence of more expert advisors across three experiments in which expertise was signalled by familiarity ratings with a university for which the graduate salary was being estimated, country of origin in a geography knowledge task, and confidence in a set of trivia questions. Likewise, Soll and Mannes (2011) saw that participants paired with advisors who were more accurate than they were at estimating basketball teams' point-per-game from other team statistics were more influenced by advice than those paired with advisors who were less accurate than they were. Tost, Gino, and Larrick (2012) used a weight-estimation task and observed that greater weight was placed on advice from advisors labelled as experts compared to the same advice from advisors labelled as novices. Schultze, Mojzisch, and Schulz-Hardt (2017) labelled advisors using a ranking system and saw that participants placed a higher Weight on Advice from highly-ranked advisors compared to low-ranked advisors, although the actual advice was (unbeknownst to the participants) the same. Wang and Du (2018) showed that participants placed greater Weight on Advice from expert as opposed to novice advice in a coin-jar estimation task using advice that was genuinely expert or novice. Önkal, Gönül, et al. (2017) reported a series of experiments using a stock price estimation task in which advisor expertise was manipulated using both labels and experience. Where experience alone was provided there was no clear effect of expertise, while labelling had a substantial effect. Follow-up experiments did demonstrate effects of experience.

5.2.3.2 Familiarity of advisors

So far as I can ascertain, no one has reported on a Judge-Advisor System in which comparable advice is received from friends and non-friends. Sniezek and van Swol (2001) and van Swol and Sniezek (2005) found a correlation between the amount classmates had interacted and ratings of trust in those classmates as advisors, but they did not use a measure of advice-taking which allows calculation of advice weighting. Minson, Liberman, and Ross (2011) had long-term dance partners make estimates about their own dance performances in relation to professional assessment, and saw normal levels of egocentric discounting.[2]

It seems intuitive that advice will be weighted more highly when coming from people we know, but this does not appear to have been tested. With regard to the three-factor model of trust put forward in Mayer, Davis, and Schoorman (1995), knowing and trusting an advisor would be expected to increase the extent to which the advisor's advice is taken.

5.2.3.3 Humanity of the advisor

Some experimenters have examined advice weighting for non-human advisors. In a stock-market forecasting task, Önkal, Goodwin, et al. (2009) found that judges placed less weight on (identical) advice when it was labelled as coming from a statistical model versus a human expert forecaster. In their study on over-weighting, Schultze, Mojzisch, and Schulz-Hardt (2017) provided participants with advice labelled as coming from a random number generator and noted that its estimates were still assigned some weight by judges. The influence of this randomly-generated advice was roughly equivalent to that of human advisors not labelled as having high expertise.

5.2.3.4 Advisor conflict of interest

Where advisors have a conflict of interest, following the advice may benefit the advisor at the expense of the judge. Gino, Brooks, and Schweitzer (2012) observed that non-anxious judges assigned less weight to advice from advisors with a conflict of interest, while anxious judges assigned similar weight regardless of conflict of interest. Bonner and Cadman (2014) similarly saw less influence from advice from advisor with a conflict of interest in a CEO-remuneration task.

[2] The partners' estimates in this study were highly correlated, meaning that they were not independent and thus did not necessarily bracket the correct answer. This means that both partners assigning positive weight to the other's advice is not likely to approach the correct answer.

5.2.4 Wider social factors

The Judge-Advisor System in experiments is usually quite abstracted away from advice-taking in the real world, in which broader social factors are likely to offer a highly influential context dictating or suggesting norms for advice-taking. Despite being somewhat shielded from these broader factors, they are nevertheless apparent in some Judge-Advisor System experiments.

Fairness may be a universal value (de Waal 2014), and it is certainly a lauded value in the developed Western societies where the majority of Judge-Advisor System research takes place. Consistent with this, judges in the Sniezek, Schrah, and Dalal (2004) study tended to offer both novice and expert advisors equal shares in their reward money. This sense of fairness is seen to extend to advice-weighting, too. Harvey and Fischer (1997) saw expert judges consistently placed some weight on novice advice, and attributed this advice-taking behaviour to a social requirement for fairness. Likewise, Mahmoodi et al. (2015) showed that dyads making perceptual decisions would consistently over-weight the estimate of the less accurate dyad member despite evidence that this impaired performance and thus decreased the dyad's reward money.

Feeling powerful was observed to lead to decreases in advice-taking by See et al. (2011) and Tost, Gino, and Larrick (2012), partially mediated by the judge's confidence. Somewhat similarly, Gino, Brooks, and Schweitzer (2012) saw angry judges took less advice through being more self-confident, while anxious judges sought and took more advice through being less self-confident.

Finally, developmental maturity seems to be an important factor. Rakoczy et al. (2015) found that while 3-6 year old children did differentiate between knowledgeable and ignorant advisors, they nonetheless placed exceptionally high Weight on Advice from both advisors.

5.3 Purported explanations for egocentric discounting

Almost as numerous as the studies into egocentric discounting are the explanations offered to account for it. Despite this, no explanation has managed to withstand critical empirical scrutiny. Below, I offer a brief review of the explanations which have been put forward to date.

5.3.1 Egocentric bias

Among the earliest explanations for egocentric discounting was egocentric bias: the belief that one's judgement is superior to that of others. Harvey and Fischer (1997) attributed egocentric discounting to such self-serving estimates of ability, such as when car drivers report on average being better than average (Svenson 1981). The explanation also suggested a mediating role of overconfidence, tying it neatly into later similar explanations from See et al. (2011), Gino, Brooks, and Schweitzer (2012), and Tost, Gino, and Larrick (2012). For all these authors, a judge's confidence in the initial estimate is the principal driving force behind the weighting of advice. This view is appreciable from a Bayesian perspective on advice integration (Bahrami et al. 2010): as with multi-sensory integration (Fetsch et al. 2012), informational cues coalesce optimally onto the correct answer where they are weighted according to their precision. In such a framework, there is an optimal (Bayesian) integration process which is being fed faulty inputs (because the judge's own estimate is believed to have erroneously high precision).

Overconfidence certainly seems to have a role, and accounts well for many of the nuances in advice-taking experiments including the findings that advice is taken more readily for difficult tasks (where judge confidence is lower) and that judges made to feel more powerful (and hence more confident) take less advice. It is unlikely to be the whole story, however: Trouche, Johansson, et al. (2018) note that Soll and Mannes (2011) were able to disentangle ratings of ability and advice-taking behaviour and saw that the egocentric discounting occurred even if the judge's assessment of relative ability were taken as true.

5.3.2 Access to reasons

One of the most influential explanations of egocentric discounting has been Yaniv's argument that judges have greater access to the reasons justifying their own decisions than to those justifying the decisions of others, due to the opaqueness of other minds (Yaniv and Kleinberger 2000). This differential access to reasons suggests, from the perspective of the judge, that there is greater evidence favouring the judge's own opinion than favouring the estimate of their advisor, and that, analogous to the confidence case above, the more well-supported opinion should be given a greater weight during integration.

Trouche, Johansson, et al. (2018) presented judges with their own estimates labelled as advice, and with advice labelled as their own estimates, and observed that judges persisted in placing greater weight on what they *believed* was their own initial estimate rather than what was *actually* their own initial estimate. This result is a serious problem for the access-to-reasons explanation, because there is no good reason why simply relabelling estimates should change the judge's internal census of evidence supporting the estimates.

5.3.3 Anchoring

Some researchers have argued that egocentric discounting can be explained by anchoring. Anchoring is a well-established phenomenon whereby numbers clearly unrelated to a numerical estimation task can nevertheless bias estimates, as when participants asked to estimate the height of Mount Everest in feet but first asked to say whether '2,000' is higher or lower than the correct value (12,000) give far lower estimates than participants asked to say whether '45,500' is higher or lower than the correct value (Jacowitz and Kahneman 1995). Bonner and Cadman (2014) suggested that anchoring was responsible for judges' over-use of outlandishly extravagant suggestions for CEO remuneration. In a more thorough set of studies, Schultze, Mojzisch, and Schulz-Hardt (2017) observed a consistently greater-than-zero weight on transparently useless advice, which they ascribed to anchoring to the advice. While these cases for anchoring may be admitted, they are to do with over-weighting advice rather than the under-weighting advice which characterises egocentric discounting. This is because the putative anchor is the advisory estimate.

Historically, it has been suggested that the judge's initial estimate acts as an anchor (Harvey and Fischer 1997), but this was later ruled out when Harvey and Harries (2004) demonstrated egocentric discounting persevered when the labels for the judge's initial estimate and the advisor's advice were switched. If anchoring to the initial estimate were responsible for egocentric discounting the relative weighting of initial estimate and advice would have followed the actual values rather than the labelled values, whereas the data showed that discounting occurred towards the *labelled* rather than *actual* initial estimate.

5.3.4 Sunk costs

Another general cognitive bias recruited as an explanation for egocentric discounting is the sunk costs fallacy, in which one perseveres with a poor strategy in order to justify the cost or effort that has already gone into pursuing it. Interestingly, sunk costs have been recruited to explain both egocentric discounting *and* following advice.

Gino (2008) had participants receive advice for free or for a fee depending upon the outcome of a coin flip. They found that the same advice was more influential where payment had been taken, and that the more expensive the more influential it was whether paid for or free. Assuming that participants viewed the greater cost as a marker of quality, the remaining effect contingent on whether or not payment had actually been taken can be explained by sunk costs. Similarly, Sniezek, Schrah, and Dalal (2004) found that, at least for expert advisors, judges who allocated a portion of their prospective reward money to the advisor *before* receiving the advice placed more weight on that advice.

Ronayne and Sgroi (2018) also invoked the sunk costs fallacy, but suggested that it could account for using less advice, i.e. discounting. These authors presented participants with the opportunity to use another participant's results rather than their own in a reward lottery, which they somewhat oddly labelled 'advice'. Despite this 'advice' being transparently better, participants frequently chose to keep their own results rather than switch, and this finding is explained on the basis that those participants were loath to forfeit the work they had done to obtain their own results by adopting another's. Extended to the Judge-Advisor System, this explanation would predict that more effortful tasks would lead to greater egocentric discounting. This is an intriguing prediction, but perhaps because effortfulness tends to covary with difficulty (which in turn decreases egocentric discounting), it does not appear to have been studied.

5.3.5 Naïve realism

A third cognitive bias invoked to explain egocentric discounting is naïve realism. Naïve realism occurs when people treat their own perceptions as reflective of shared underlying reality and others' perceptions as misguided or biased to the extent that they do not agree. Minson, Liberman, and Ross (2011) argue strongly for a naïve realism explanation of egocentric discounting, although it is never fully explained why, on a naïve realism view, *any* adjustment to advice is warranted (because ex hypothesi the initial estimate reflects the true answer). The most creative element of the experiments involves funnelling integrated decisions (corresponding to judges' final decisions in the typical Judge-Advisor System) through a collaborative joint decision process before extracting a final individual decision. Naïve realism once again fails to provide a compelling explanation why the final individual decisions closely reflected the collaborative joint decisions rather than the initial estimates: rather than continuing to endorse their own view of 'reality', participants appeared to be willing to accept the joint view once it had been established.

5.3.6 Responsibility / feeling of deserving outcomes

Some advice-taking may be explicable on the basis of responsibility sharing. Harvey and Fischer (1997), Mahmoodi et al. (2015), and Ronayne and Sgroi (2018) have all suggested that taking advice can transfer some of the responsibility for the outcome of a decision onto the advisor. Where uncertainty is high, or where rewards are shared, this can be particularly useful.

While distribution of responsibility is more a reason for *reduced* rather than *increased* egocentric discounting, when combined with an account that predicts ascribing very low or no weight to advice by default (e.g. naïve realism), it can explain why advice weighting is higher than the zero that would be expected. It seems more plausible, however, that factors which promote advice-taking such as fairness, advisor expertise, and distribution of responsibility serve to place a limit on egocentric discounting, rather than that complete discounting is a default strategy from which these factors move judges.

5.3.7 Wariness

Trouche, Johansson, et al. (2018) designate the above explanations 'proximal explanations', because they offer a mechanistic account of how discounting occurs. Rather than partaking in this discussion, they instead provide an 'ultimate explanation', which may explain why discounting occurs. They suggest that discounting occurs because advisors' interests do not always align with judges', and thus some level of discounting offers protection from relying too heavily on advice that may be deliberately harmful. They suggest thus that discounting is an evolved response to misaligned incentives between judges and advisors. The account I offer below is in the mode of this explanation.

5.4 A wider view of egocentric discounting

The explanations outlined above 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 might be averaging one's own opinion with advice.

As an anecdotal 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. The approach taken here works to codify this intuition as a set of hyper-priors: expectations about the relative utility of advice as compared to one's own opinion. I argue that the normative model of advice-taking, and the pared-down experimental design with which is it entwined, seek to take the situation out of the evolutionary history of advice, but cannot take the evolutionary history of advice out of the situation.

According to this hypothesis, the hyper-priors on advice-taking are a consequence of the many opportunities for deception and misunderstanding which apply to advice but not to one's own opinion. The most obvious of these is the opportunity for deception. As Trouche, Johansson, et al. (2018) point out, advisors do not always have the judge's best interests at heart when supplying advice. Consider, for example, a situation where one co-worker, Sally, asks another, Hanan, whether it is a good idea to apply for a promotion. Hanan may think it would be good for Sally to apply, because Sally is well-qualified and hard-working, but nevertheless discourage Sally from applying because Hanan herself is going to apply and wishes to reduce the competition.

It is not necessary, however, for there to be misaligned incentives of this kind. Advice may be less informative than one's own opinion where the advisor is less able to perform the task, or does not perform the task as effectively. Consider, for instance, Sally asking Hanan for advice on where to go on holiday. Hanan may wish to maximise the probability Sally has a wonderful holiday by offering the best advice possible, but, because Hanan does not have a perfect knowledge of Sally's preferences, nevertheless advise Sally to select a non-optimal destination. Likewise, Sally may well have spent considerable time researching and thinking about the question, and it is not reasonable to believe that Hanan would do likewise because it is, after all, Sally's holiday.

There is room for misunderstanding even where an advisor's interests are aligned with the judge's and the advisor's ability equals the judge's. Advice must be communicated to the judge, and communication in the real world is inherently noisy. Communication of advice requires something in the mind of the advisor to be encoded into a set of signals, transmitted to the judge, and then re-encoded into something in the mind of the judge, at which point it can be integrated with what the judge already believes. Information can be degraded at any of these steps, resulting in advice that is less informative than the judge's own opinion. When Sally asks Hanan what to do about a work problem, and Hanan rapidly and confidently rattles off a suggestion, Sally may be forgiven for thinking the rapidity and confidence are a property of Hanan's confidence in the suggestion rather than an underlying characteristic of Hanan. If Sally does not adjust for the fact that Hanan is always more confident about things than Sally is, then Hanan's suggestion will be overly dominant relative to its informational value.

5.4.1 Compatability with existing explanations

As noted by Trouche, Johansson, et al. (2018), ultimate explanations of the kind offered here do not invalidate proximal explanations of the kind offered in the middle of this chapter§5.3. My view is most consistent with a Bayesian integration view in which advice is weighted by a range of features of advice, advisor, and context, with the further proviso that there are hyper-priors which govern the default level of discounting. I am thus content to observe the contest to provide proximal explanations for changes in the level of egocentric discounting, and only baulk at claims that egocentric discounting is 'irrational'. I suggest that, once egocentric discounting as a default is accepted, the adjustments in the level of discounting are almost all transparently rational.

5.4.2 Evidence

In the chapters that follow, I present evidence from computational agent-based evolutionary simulations and on-line 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. I note that only the first of these is necessary for illustrating the plausibility of the theory: the behavioural adjustments serve more to support the extension of the argument, that changes in the level of discounting are rational adjustments.

6 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. Demonstrating the existence of prior expectations is difficult, because they are beyond the bounds of any task, but we aim instead to demonstrate that they are adaptive in terms of fitness as revealed by evolutionary pressure. Using computational agent-based evolutionary simulations, we demonstrate in this chapter that conservatism can emerge within a population even where detrimental advice is rarely experienced, and can even emerge where individuals always interact in good faith.

As discussed in the previous chapter§5, 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. We explored this tendency as a function of three plausible scenarios. The first scenario§6.2 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§6.3, 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§6.4, 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, we hypothesised that some level of egocentric discounting will emerge as the dominant strategy, i.e., the mean population weighting for an agent's initial estimates versus advice they receive will be greater than .50.

6.1 General method

Agent-based computational models of an evolutionary process were programmed in R (R Core Team 2018) and run variously on a desktop computer and the Oxford Advanced Research Computing cluster (Richards 2015). The code is available at `https://github.com/oxacclab/EvoEgoBias` (Jaquiery 2021d), and a cached version of the specific data presented below are available to allow inspection without rerunning the models completely (Jaquiery 2021a).

The models reported here use 100 generations of 1000 agents which each make 60 decisions in each generation on which they receive the advice of another agent. Each scenario is run with 7 different strengths of the manipulation value, and each of these instances is run 50 times. Decisions are either point estimation (Scenarios 1§6.2 and 2§6.3) or categorical decision with confidence (Scenario 3§6.4). 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 60 decisions of its lifetime.

6.1.1 Initial estimates

The agents perform a value estimation (category estimation in Scenario 3) task. Agent a'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 (in)sensitivity parameter (s^a, which is itself drawn from a normal distribution with mean and standard deviation 1 when the agent is created, and clamped to a minimum value of 0.1).

An agent's initial estimate (i^a_t) is thus:

$$i^a_t = v_t + N(0, s^a) \tag{6.1}$$

This structure is the same as in the normative model of advice-taking§1.1.5 presented in Equation 1.2. The normally distributed error terms cancel out on average over multiple estimates. Where s^a and $s^{a'}$ (an advisor's sensitivity) are equal or unknown (as in these simulations) the normative model indicates that the best strategy is to average these values.

6.1.2 Advice

Each agent receives advice from another agent which it combines with its initial estimate to reach a final decision. Each agent is reciprocally connected to 10 other agents at random when they are spawned, and on each decision they receive advice from one of their connections at random.[1] The advice has a probability of being mutated in some fashion. The mutation depends upon the scenario and is described separately for each.

[1] A small number of connections allows agents to sample multiple pieces of advice from the same advisors over their lifetime. The agents do not update during their lifetime, so more realistic connectivity structures are not needed in these simulations.

6.1.3 Final decisions

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

An agent's final decision (f_t^a) is thus:

$$f_t^a = i_t^a \cdot w^a + i_t^{a,a'} \cdot (1 - w^a) \tag{6.2}$$

The final decisions in Scenario 3§6.4 are more complex, but follow a similar structure.

6.1.4 Reproduction

After their 60 decisions are up, agents reproduce and then die, creating a new generation that will perform their 60 decisions and reproduce, and so on. Reproduction favours those agents who performed better on their 60 decisions, and each newly created agent has a small chance to mutate slightly.

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§6.2 and 2§6.3, which use continuous decisions, this fitness is obtained by subtracting the absolute difference between the final decision and the true value for each decision (mean absolute deviance):

$$u^a = -\sum_{t=1}^{60} |v_t - f_t^a| \tag{6.3}$$

Scenario 3§6.4 uses categorical decisions. These are processed using a fitness function that awards agents -1 point for each incorrect categorisation of the true world value.

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. Each agent is then given a probability to reproduce equal to their share of the sum of all fitness scores:

$$p^a = \frac{u^a}{\sum_{j=1}^{n} u^j} \tag{6.4}$$

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. When spawned this way, each agent has a 1% chance to mutate slightly, taking instead a value sampled from a normal distribution around its parent's egocentric bias with a standard deviation of .1. All other agent features, decision-making accuracy, neighbours, and confidence scaling (in Scenario 3§6.4) 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 (higher w^a) as the dominant adaptive strategy.

6.2 Scenario 1: misleading advice

In scenario 1, advisors sometimes offer misleading advice.

6.2.1 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 ($i_t^{a'}$), or an extreme answer in the opposite direction to the advising agent's initial estimate

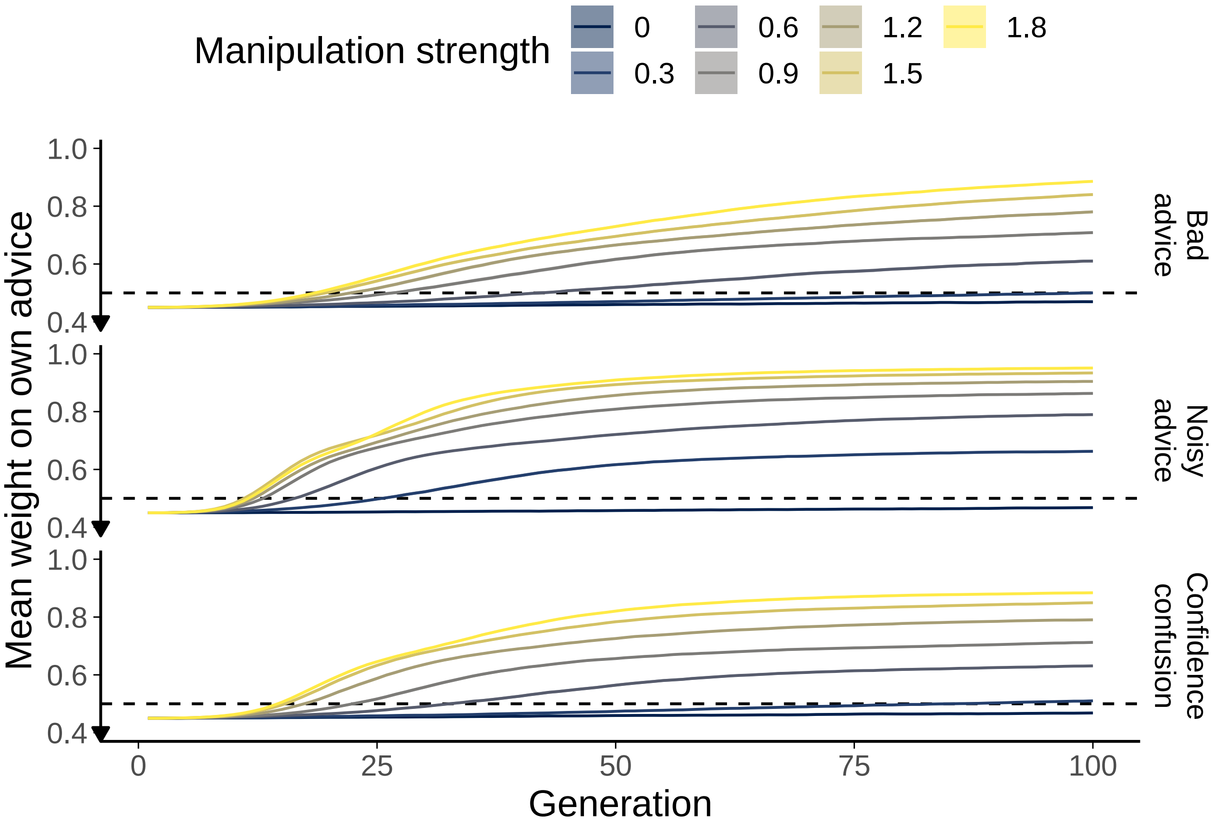


Figure 6.1: Egocentric discounting simulation results.
The mean weight on agents' own advice is shown averaged across 50 runs of each scenario at each manipulation strength. The width of the ribbons shows the mean 95% confidence intervals for the population self weights over the 50 runs. The dashed line shows equal weighting (.5), the mathematically optimal value for integrating a single estimate from an advisor of equivalent ability to oneself.

(i.e. lower than 50 if $i_t^{a'}$ was above 50, and vice-versa). The probability of extreme advice is x%, where x is the manipulation value for the simulation. If an agent gives extreme advice they add 3 standard deviations of their own sensitivity to the answer they would otherwise give.

6.2.2 Results

Figure 6.1 (top) shows the results of 350 simulations (50 at each level of bad advice frequency). The starting value of self-weight in all simulations is set to .45 so that evolution should still occur in the absence of a manipulation towards the optimum level of self-weight in that case (.5). The evolutionary pressure in the absence of a manipulation is quite low, so over the 100 generations modelled the mean self-weight in simulations without a manipulation barely moves towards .5. Longer simulations show that the population does evolve to and remains at the optimum value of .5 in the manipulation-free cases.

Where advice is sometimes misleading, egocentric discounting emerges rapidly. In longer simulations, this value remains stable throughout the rest of the stimulation. The greater the probability of misleading advice, the more rapid the rise towards higher mean self-weights, and the higher the stable self-weight value eventually attained.

The actual values of self-weight evolved in the population are dependent upon how frequently bad advice occurs and how bad the advice is when it does occur. For the particularly bad advice offered by the agents in this scenario, probabilities of bad advice around 0.6-0.9% produce average self-weights in the 70-80% region typically observed experimentally (Yaniv and Kleinberger 2000).

6.2.3 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.

Note that, because of the rarity of bad advice (it occurs 1.8% of the time in the most pronounced case), most individuals never experience bad advice. This means that any mutant that is more conservative than their parent is likely to lose fitness by ignoring good advice much of the time, but in the rare cases where bad advice is given they save themselves from taking such a heavy penalty. Perhaps the advice in these simulations is too bad to go undetected in more realistic interactions (the agents here do not consider the plausibility of the advice they receive); more plausible misleading advice would have to be more common to produce the same magnitude of effects in self-weight evolution, but the results would be qualitatively similar. The reproductive function also matters: if we selected a single agent to spawn the entire subsequent generation, we would likely see a wholly trusting agent be selected each time, just one that happened to never be given bad advice. Of course, in many models of this kind of behaviour there would be a fitness benefit to deceiving other agents, leading to an increase in the frequency of bad advice (Alexander 2021). Overall, there is a balance between the frequency and extremity of bad advice and the specific nature of the reproductive function, but the simulations demonstrate that, in circumstances where bad advice is a possibility, it is adaptive to guard against it to some extent.

6.3 Scenario 2: noisy advice

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

6.3.1 Method

As with Scenario 1§6.2, the true value (v_t) is fixed at 50.

Advice in this scenario has additional noise added by increasing the standard deviation of the advisors' error term by the manipulation value:

$$i_t^{a,a'} = v_t + N(0, s^{a'} + x) \tag{6.5}$$

Where x is the manipulation strength for the condition.

Models were also run where the advisors' initial estimates were used as the basis for advice, and noise added onto that (perhaps to represent noisy communication rather than less effort being made in the assessment):

$$i_t^{a,a'} = i_t^{a'} + N(0, x) \tag{6.6}$$

The results were the same and are not reported here.

6.3.2 Results

As before, egocentric discounting emerges rapidly in the active condition and remains stable throughout the rest of the simulation (Figure 6.1, middle). To attain mean self-weight values in the 70-80% region usually observed experimentally, the manipulation strength needs to be around 0.3. The population s^a values are drawn from a distribution with mean 1, so this equates to roughly 30% extra noise on the advice compared to the initial estimates.

6.3.3 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 estimates: difficulties in communicating the advice or different levels of conscientiousness in decision-making may also produce this effect.

6.4 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, however, there are large and consistent individual differences in people's expressions of confidence, both when expressed numerically (Ais et al. 2016; Song, Kanai, et al. 2011) and when expressed verbally (MacLeod and Pietravalle 2017; Wallsten et al. 1986). In this scenario, we therefore investigated emergent strategies when 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 understanding of how expressed confidence maps onto internal confidence.

6.4.1 Method

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

$$v_t = N(50, 1) \tag{6.7}$$

This allowed categorical answers which identified whether or not v_t was greater than 50. Agents were equiprobably assigned a personal confidence factor (c^a) of 0 or 0.5. This was used to scale the difference between the advising agent's initial estimate and the category boundary to produce the advice:

$$i_t^{a,a'} = (i_t^{a'} - 50)(c^{a'} \cdot x + 1) \tag{6.8}$$

This meant that around half the agents in any given simulation expressed their internal confidence directly (because $c^a = 0$), while other agents magnified their confidence by a factor between 0 and 1.9. Advice in this scenario, unlike the previous, is on a scale in the range [-50, 50], where the magnitude of the advice represents its confidence.

Each agent then used the reciprocal of this process to translate advice back into its own confidence scale:

$$r_t^a = i_t^{a,a'} / (c^a \cdot x + 1) \tag{6.9}$$

Agents then integrated advice with their initial estimate according to their egocentric discounting tendency to arrive at a final decision:

$$f_t^a = i_t^a \cdot w^a + r_t^a \cdot (1 - w^a) \tag{6.10}$$

Crucially, because the confidence coefficient can differ between advisor ($c^{a'}$ in Equation 6.8) and judge (c^{a} in Equation 6.9), some agents will under- or over-estimate their advisor's actual confidence.

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.

6.4.2 Results

Egocentric discounting again emerges in the condition where agents can have different confidence factors (Figure 6.1, bottom). The adaptiveness of egocentric discounting in this scenario arises because advice from a mismatched partner (e.g. one inflating their confidence and another not) requires a different interpretation than advice from a matched partner. The application of an inappropriate interpretation results in misleading advice, so there is a pressure for a middle-of-the-road strategy whereby advice is counted, but not too much.

6.4.3 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. This scenario is all the more plausible given the empirical observations that, although people's own mappings of verbal confidence terms to numerical confidence ratings are fairly consistent, and rank-order of terms between individuals is fairly consistent, mappings of verbal confidence terms to numerical confidence ratings varies substantially between individuals (MacLeod and Pietravalle 2017; Wallsten et al. 1986).

6.5 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, and Schweitzer 2012; Sniezek, Schrah, and Dalal 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 hyper-priors 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§5 can be seen as eliciting exactly this behaviour. All advice-taking experiments occur in a context that differs from a hypothetical average advice-taking context, but some experiments specifically manipulate the context as part of their investigation. In the clearest example of this, Soll and Mannes (2011) generated advice so that participants would either be better, the same as, or worse than advisors at a task where they had to estimate basketball teams' points-per-game from other team statistics. The participants paired with advisors who were better than they were had lower average self-weights in their final decisions, indicating that they were more influenced by the advice. Those participants paired with advisors more accurate than themselves still placed more weight on their own initial estimates, but not so much as the participants who were paired with advisors whose competence matched their own. The participants paired with advisors less accurate than themselves took the least advice. This response to a context that mimics the manipulation in scenario 2§6.3 (the relative accuracy of advisor and judge) indicates that, unlike the fixed agents in our model, people can flexibly adapt to different contexts. In the next chapter I present new behavioural experiments which suggest that people can indeed modulate their advice-taking to the specifics of the context in which they are in, and show that this happens rapidly.

7 Behavioural responses to advice contexts

I suggest that the long-standing view that egocentric discounting reflects sub-optimal information processing is a consequence of taking a narrow view of the problem being solved. While it is indeed demonstrable that people in Judge-Advisor System experiments would perform better and earn more reward if they took more advice (Soll and Larrick 2009), the people who participate in those experiments also have to function in the real world where being too trusting can produce very negative outcomes. Furthermore, advice-taking and decision-making are, like the behaviour of other organisms, ultimately mechanisms for the more efficient propagation of genes. While participants in a Judge-Advisor System are working out how to best weigh their own experience with another opinion, the challenge faced by the genes in their cells is to somehow navigate a complex, often-unreliable, and frequently-changing world in order to produce more copies of themselves. I suggest that egocentric discounting may result from genetic and cultural evolution favouring the default assumption that an individual's own information is a more reliable basis for decision-making for that individual than another individual's information.

The evolutionary models offered a proof-of-concept illustration that hyper-priors may evolve under a variety of contextual features that are almost always at least partially true of advice-taking situations. Hyper-priors are expectations that are not changed as a function of experience within the scope of a given scenario, and help form the context within which a scenario occurs, similar to a frame of reference in physics. In advice-taking contexts, a judge's view of their advisor's benevolence may fluctuate over the course of a few back-to-back exchanges, whereas their view of the general benevolence of people as a whole is unlikely to change in a meaningful way in that time. The latter, therefore, is a hyper-prior because it contributes to the advice-taking behaviour without being altered by the situation.

Hyper-priors are so termed to distinguish them from priors: expectations that are updated following evidence. In a chaotic world dominated by complex phenomena emerging from the interaction of agents with sophisticated mental processes, sometimes the best genetic strategy is to hedge your bets by building a phenotype that can respond flexibly to different contexts. It may well be, therefore, that people not only have hyper-priors concerning the likely value of advice, but also priors that can respond to different contexts and to changes in context. In this chapter, we explore the flexibility of egocentric discounting in the contexts presented in the previous chapter§6. We present three experiments that examine how advice-taking changes according to the benevolence and the identifiability of advisors. We also include a brief discussion of the literature on advisor expertise.

7.1 Benevolence of advisors

The evolutionary models discussed in the previous chapter§6 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, is one of the three pillars of the Mayer, Davis, and Schoorman (1995) model of advice-taking. Despite this, relatively little investigation has been made into the role of benevolence in trust, as discussed previously§5.2.3.2.

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 it comes from an advisor who is highly unlikely to be 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.

We expected that advisor influence would be higher for advice that participants rated as 'honest' versus advice rated as 'deceptive'. Likewise, we expected that influence would be higher for advisors who were described as 'always honest', even for advice rated as 'honest'. In other words, we expect that participants' advice-taking depends upon both the plausibility of the advice and the benevolence of the advisor.

Early versions of the experiments we conducted used a minimal groups paradigm (Rabbie and Horwitz 1969; Pinter and Greenwald 2011) in an attempt to induce an in-group/out-group distinction in participants. We were not able to get this manipulation to produce difference perceptions of the advisors, as measured by participants' questionnaire responses, and so we resorted to directly cuing participants about the benevolence of advisors. The experiments presented below, Experiments 5§7.1.1 and 6§7.1.2, are the result of previous experiments exploring how we could represent the manipulation in a way that participants paid attention to and remembered.

7.1.1 Experiment 5: benevolence of advisors

In this experiment, participants were cued as to the benevolence of their advisors. The advice on each trial came from one of two advisors the participants became familiar with over the course of the experiment. Participants were asked to rate the advice prior to submitting their final decisions. We expected that participants would rate advice from an advisor who was more benevolent as more honest, and that they would weigh that advisor's advice more heavily, even where the advice itself was rated the same.

Open scholarship practices This experiment was preregistered at https://osf.io/tu3ev. This is a replication of a study of identical design that produced the same results. The data for both this and the original study can be obtained from the esmData R package (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/b4289fea196f71ccf0ba0b2ae8fde12139a16301/ACv2/db.html.

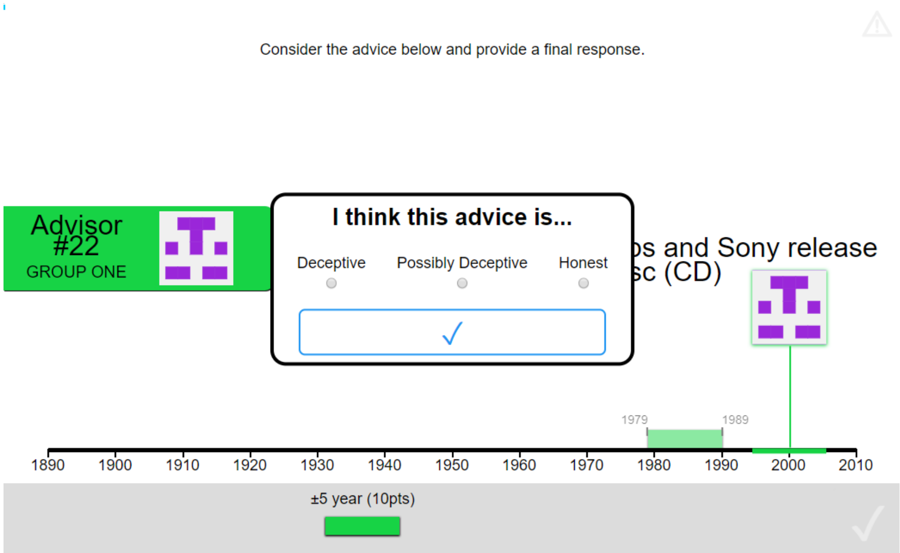
There were two deviations from the preregistered analysis in this experiment. Several participants used translation software to translate the experiment website. We could not guarantee that the questions were accurately translated and so these participants were excluded. Some participants never rated the Sometimes deceptive advisor's advice as 'Honest', and so they were excluded in the t-test comparing advisors.

7.1.1.1 Method

Procedure 20 participants each completed 28 trials over 3 blocks of the continuous version of the Dates task§2.1.3.2. Participants' markers covered 11 years, meaning that it would cover an entire decade inclusively, e.g. 1965-1975. When participants received advice, but before they submitted their final decision, they rated the honesty of the advice on a three-point scale (Figure 7.1).

Participants started with 1 block of 7 trials that contained no advice to allow them to familiarise themselves with the task. All trials in this section included feedback for all participants indicating whether or not the participant's response was correct.

Participants then did 7 trials with a practice advisor to get used to receiving advice. They also received feedback on these trials. They were informed that they would "get advice on the answers you give" and that the feedback they received would "tell you about how well the advisor does, as well as how well you do". Before starting the main experiment they were told that they would receive advice from multiple advisors and that "advisors might behave in different ways, and it's up to you to decide how useful you think each advisor is, and to use their advice accordingly".



Consider the advice below and provide a final response.

I think this advice is...

○ Deceptive ○ Possibly Deceptive ○ Honest

Figure 7.1: Advice honesty rating.
Participants rated advice on a three-point scale according to whether they thought the advice was deceptive or honest.

Participants then performed 2 blocks of trials that constituted the main experiment. In each of these blocks participants had a single advisor for 6 trials, plus 1 attention check. No feedback was given on answers in the main experiment blocks.

The two advisors were identical in how they generated advice, but they were labelled differently. The advisors had different coloured backgrounds (e.g., purple and green), and participants were told that the advisor whose background colour matched the participant's colour "will give you the **best advice that they can** [original emphasis]", while the advisor who did not match the participant's colour "might sometimes try to direct you **away from the correct answer** [original emphasis]". The advisor whose background matched the participant's colour was labelled as being in 'group one', while the other advisor was labelled as being in 'group two'. This visual presentation arose from earlier experiments that implemented (unsuccessfully) a minimal groups paradigm. Colours and the order in which the advisors were encountered were counterbalanced.

Advice profiles Despite differences in labelling, the advisors were identical in terms of how they actually produced advice. The advisors offered advice by placing an 11-year wide marker on the timeline. The marker was placed with its centre

Table 7.1: Participant exclusions for Experiment 5

Reason	Participants excluded
Attention check	2
Multiple attempts	0
Missing advice rating	2
Odd advice rating labels	1
Not enough changes	2
Too many outlying trials	0
Total excluded	**3**
Total remaining	**17**

on a point sampled from a normal distribution around the correct answer with a standard deviation of 11 years in the manner described earlier§2.1.3.2.

7.1.1.2 Results

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 60s to complete, 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. The latter exclusion was added after data were collected because it was not anticipated that participants would use translation software in the task. The numbers of participants who failed the various checks are detailed in Table 7.1.

The final participant list consists of 17 participants who completed an average of 11.88 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 (F(1,16) = 31.85, p < .001; $\mathrm{M}_{\mathrm{Initial}}$ = 19.07 [14.98, 23.16], $\mathrm{M}_{\mathrm{Final}}$ = 11.12 [8.27, 13.97]), which suggests that they incorporated the advice, which was indicative of the correct answer (Figure 7.2). The participants had less error on decisions made with the Always honest advisor than the Sometimes deceptive advisor (F(1,16) = 9.38, p = .007; $\mathrm{M}_{\mathrm{AlwaysHonest}}$ = 13.24 [10.16, 16.32], $\mathrm{M}_{\mathrm{SometimesDeceptive}}$ = 16.95 [13.18, 20.72]), although surprisingly there was no statistically significant

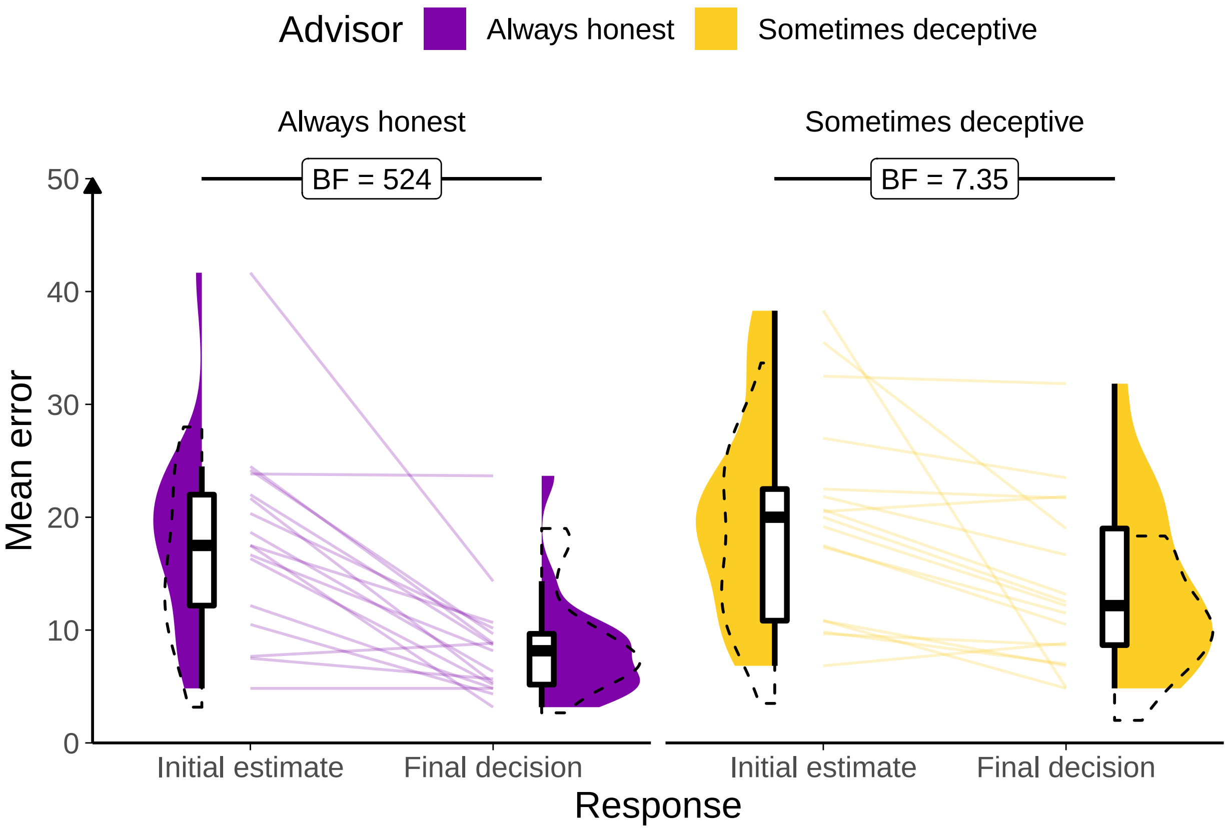


Figure 7.2: Task performance for Experiment 5.
A: Response error. Faint lines show individual participant mean error (the absolute difference between the participant's response and the correct answer), for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of participant means on the original study which this is a replication. The dependent variable here is error, the distance between the correct answer and the participant's answer, and consequently lower values represent better performance. The theoretical limit for error is around 100.

interaction to indicate that this was due to greater reduction in error scores over time for that advisor (F(1,16) = 1.94, p = .183; $\mathrm{M}_{\mathrm{Reduction|AlwaysHonest}}$ = 9.70 [5.95, 13.44], $\mathrm{M}_{\mathrm{Reduction|SometimesDeceptive}}$ = 6.20 [1.96, 10.43]).

Participants only had one marker they could place, and separate confidence judgements were not asked for, so we cannot directly assess confidence in these data.

Advisor performance The advice given by the advisors was generated stochastically from the same distribution. Any differences will be random. This was demonstrably the case in the domain of advice error (absolute distance between the

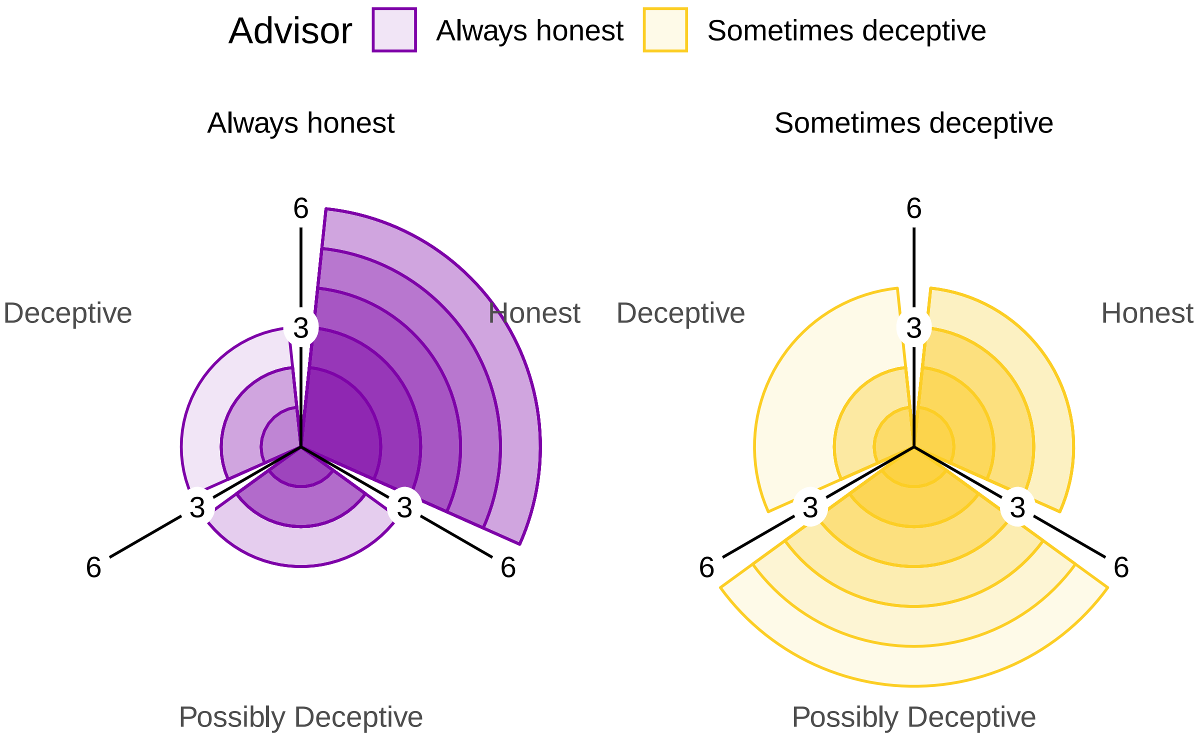


Figure 7.3: Advice rating in Experiment 5.
The polar plots show the number of times each participant gave the given rating to the advice of each advisor. The colour density illustrates the number of participants who gave at least that many ratings to the advice of an advisor.

centre of the advice marker and the correct year; $\text{BF}_{\text{H1:H0}} = 1/3.92$; $\text{M}_{\text{AlwaysHonest}} = 8.05$ [6.81, 9.29], $\text{M}_{\text{SometimesDeceptive}} = 7.90$ [6.84, 8.96]). The Bayes' Factor for the domain of agreement – the absolute distance between the centre of the advice marker and the centre of the participant's initial estimate marker – was essentially on the threshold ($\text{BF}_{\text{H1:H0}} = 1/2.96$; $\text{M}_{\text{AlwaysHonest}} = 20.14$ [16.09, 24.19], $\text{M}_{\text{SometimesDeceptive}} = 21.85$ [17.00, 26.70]).

Advice ratings The advisors' advice was rated differently by the participants, as shown in Figure 7.3. The difference was statistically significant, indicating that the patterns of ratings differed depending on the advisor giving the advice ($\chi^2(2) = 18.14$, $p < .001$, $\text{BF}_{\text{H1:H0}} = 335$; AlwaysHonest:SometimesDeceptive ratio: Deceptive 0.65, Possibly Deceptive 0.49, Honest 1.82). This indicates that participants understood the task and that the manipulation worked as intended, with more suspicion applied to the advice from the Sometimes deceptive advisor.

Effect of advice Participants chose their own ratings for the advice, and it was common for participants not to use all ratings for all advisors (e.g. many participants never rated advice from the Always honest advisor as Deceptive). This meant that the statistical tests preregistered for this hypothesis were broken down into separate contrasts of advice and advisor. Using a more complete test, such as 2x3 ANOVA, would have suffered greatly from missing values.

To explore the effect of advice, a 1x3 ANOVA was run on Weight on Advice across ratings. In all, 11/17 (64.71%) participants had at least one trial rated with each of the three ratings. The Weight on Advice differed according to the rating assigned the advice (F(2,20) = 9.37, p = .001; $\mathrm{M}_{\mathrm{Deceptive}}$ = 0.18 [-0.04, 0.41], $\mathrm{M}_{\mathrm{PossiblyDeceptive}}$ = 0.32 [0.09, 0.55], $\mathrm{M}_{\mathrm{Honest}}$ = 0.60 [0.46, 0.74]; Mauchly's test for Sphericity W = .935, p = .739). As expected, participants were more influenced by advice they rated as Honest compared to advice they rated as Deceptive.

Effect of advisor To distinguish the effects of the advisor from the effects of the advice, we compared Weight on Advice for only those trials where the participant rated the advice as Honest. This approach has the limitation that 2 (11.76%) participants had to be dropped due to never rating advice from the Sometimes deceptive advisor as Honest.

Comparing the two (Figure 7.4) showed that participants placed more weight on the Honest advice from the Always honest advisor ($t(14)$ = 2.67, p = .018, d = 0.78, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 3.38; $\mathrm{M}_{\mathrm{AlwaysHonest}}$ = 0.67 [0.54, 0.80], $\mathrm{M}_{\mathrm{SometimesDeceptive}}$ = 0.48 [0.34, 0.62]).

Exploratory analyses Figure 7.5 shows the overall pattern of advice-taking in the experiment. The difference in the number of dots of each colour reflects the difference in participants' ratings of advice between advisors. The lines are best-fits

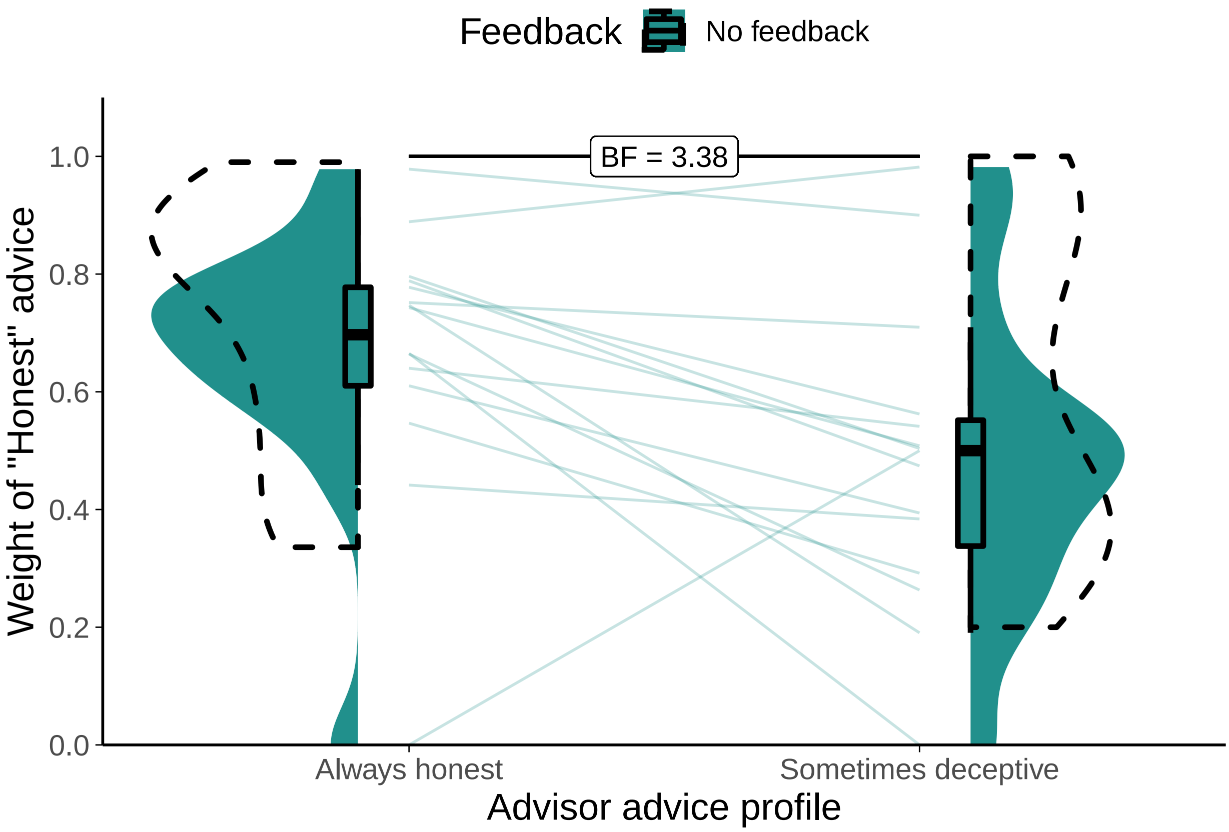


Figure 7.4: Weight on Advice in Experiment 5.
Shows the weight of the advice of the advisors. Only trials where the participant rated the advice as Honest are included. The shaded area and boxplots indicate the distribution of the individual participants' mean Weight on Advice. Individual means for each participant are shown with lines in the centre of the graph. The dashed outline shows the distribution of participant means in the original study of which this is a replication.

to the dots, and illustrate the relationship between the distance of the advice from the participant's initial estimate and the amount the participant moved the marker towards the advisor's advice for their final decision.

A linear mixed model was used to better understand the significance of the relationships in the figure, predicting influence on a trial from the advisor, the distance of the advice, and rating (dummy-coded). Random intercepts were included for participant. A similar Bayesian model was included to produce Bayes factors.

Overall, the distance between the initial estimate and the advice was hugely influential, with an increase in distance leading to an approximately equal increase in influence (β = 0.94 [0.83, 1.05], df = 184.30, p < .001, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 4.2e27).

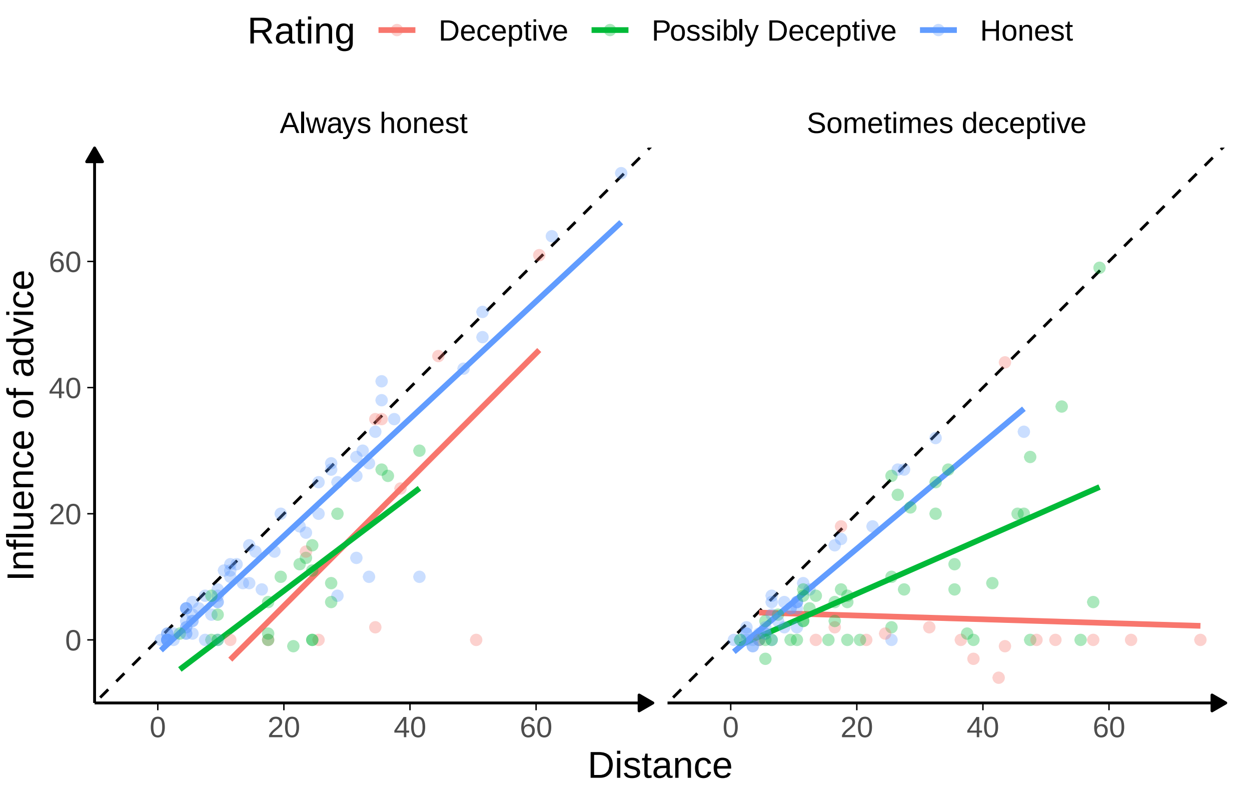


Figure 7.5: Advice response in Experiment 5.
Each point is a single trial, coloured according to the rating given to the advice by the participant. Coloured lines give the best fit linear regression for dots of that colour. The horizontal axis gives the distance between the centre of the participant's initial estimate marker and the centre of the advisor's advice marker, and the vertical axis gives the distance the participant's final decision had moved in the direction of the advisor's advice. Points that lie on the dashed line indicate that the participant moved their marker all the way to the advisor's marker, thus wholly adopting the advisory estimate as their final decision.

This indicates that the advice rated Honest from the Always honest advisor was generally followed, as is shown clearly in Figure 7.5. This pattern indicates that the participants see the difference between their own estimates and the advisor's as indicating that they themselves are mistaken. Conversely, the relationship between distance and influence was decreased when the advice came from the Sometimes deceptive advisor, even for advice rated as Honest (β = -0.16 [-0.43, 0.11], df = 180.78, p = .267, $\text{BF}_{\text{H1:H0}}$ = 79.3).

Advice rated as Deceptive from the Sometimes deceptive advisor was substantially less influential (β = 16.24 [1.51, 31.07], df = 175.94, p = .037, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/6.84), and this was complemented by the relationship between distance and influence being substantially reduced in this case (β = -0.78 [-1.25, -0.33], df = 177.51, p = .001, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 22.4).

There are other patterns in Figure 7.5 that do not come out in the statistics, possibly due to the very low trial and participant numbers, but are suggestive of a pattern we might expect. The relationship of advice rated Honest is similar between the advisors: as the advice gets further away it retains its level of influence and the participants are willing to go further to match it. This is not tested directly in the statistics, but there is no significant effect of advisor alone (β = 0.87 [-3.57, 5.28], df = 180.44, p = .705, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/2.57), the presence of which would indicate that Honest-rated advice differed between advisors. The Bayesian analogue of the test indicated that there was not enough evidence to decide whether Advisor alone was an important factor. The preregistered test of this relationship above§7.1.1.2, however, did indicate that there was a difference when tested directly.

Similarly, there may be a difference in the way the Possibly Deceptive advice was received from the Sometimes deceptive advisor as compared to the Always honest advisor (7.57 [-2.34, 17.43], df = 181.63, p = .144, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/5.80), or in how the influence of this advice related to the distance between the advice and initial estimate markers (-0.29 [-0.74, 0.17], df = 182.50, p = .225, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/3.11). The statistics did not support this, and the Bayesian statistics suggested that there was sufficient evidence against either effect being present.

7.1.1.3 Discussion

This experiment provided a first test of the sensitivity of participants' egocentric discounting behaviour to the context of advice, specifically the likely benevolence (honesty) of the advisor and specific pieces of advice. Both the source and the plausibility of the advice matter, and in this task they interacted such that discounting primarily occurred where advice was offered that differed substantially from the participant's initial estimate and came from the advisor the participants were told might mislead them. Participants adapted to the context of the advice, both in terms of how readily they were to categorise the advice as dishonest and in terms of how much they were influenced by the advice. As advice got more distant from the initial estimate, i.e. decreased on our measure of plausibility, participants had to decide whether the discrepancy was due to their own error or their advisor's. Where they believed the advisor was trustworthy, they were more likely to ascribe the error to themselves, rating the advice as honest and adopting it for their final decision. Where they believed the advisor was not trustworthy, they were more likely to label the advice as misleading and to retain their initial estimate as their final decision. Going beyond the obvious, we also saw that participants were more sceptical of advice from the less trustworthy advisor *even when they believed the advice was well-meaning.*

These observations complement the evolutionary simulations and the benevolence component of the three-factor model of trust described by Mayer, Davis, and Schoorman (1995). Nevertheless, we have demonstrated sensitivity to context, but not universal discounting due to hyper-priors. It remains a matter of speculation that people exercise epistemic hygiene by ensuring that information comes from trusted sources before integrating it, and that no source is as trusted as one's own mind. This result is in keeping with evidence from experiments where initial estimates are labelled as advice (and vice-versa). Soll and Mannes (2011) collected initial estimates from participants and then presented those estimates back to participants along with advice so participants could provide final decisions in a classic Judge-Advisor System. Unbeknownst to the participants, for some of the questions, the participant's initial estimate was labelled as the advice, and the advice was labelled as the participant's initial estimate. For those questions that were switched in this manner, the participants appeared to treat the advice as if it were their own initial decision – placing more weight on the advice than their actual initial decision. If egocentric discounting of advice were due to judges having better access to the reasons for their own estimates, for example, their own initial estimates ought to appear most reasonable and therefore be more influential, regardless of whether they were labelled as "initial estimate" or "advice". On the other hand, if people use a heuristic that their own opinion is more trustworthy because it is *their own* opinion, they will rely most on whichever figure is presented as their own initial estimate, as indeed they did.

Limitations Even compared to the other experiments in this thesis, these experiments had a low participant count. As with other Dates task studies, data collection stopped when the Bayes Factor for the main experimental hypothesis reached one of the two thresholds ($1/3 > \text{BF} > 3$). This, combined with the low trial count for each participant, meant that there were interesting follow-up questions that the data were unable to address. In keeping with the open science approach, we suggest that future investigations exploring those questions take them as preregistered hypotheses.

The Dates task is one that participants find challenging, leading to generally high levels of advice-taking in the absence of other effects (Gino and Moore 2007; Yonah and Kessler 2021). This means that the Honest advisor seemed to be entirely trusted. As discussed in the introduction to this section§5, however, even greatly trusted advisors' advice is usually subject to some egocentric discounting.

7.1.2 Experiment 6: benevolence of the advisor population

Experiment 5§7.1.1 showed that participants responded appropriately to benevolent versus less benevolent advisors. In this experiment, participants' advisors are no longer the same individuals throughout the experiment but are members of two different groups, a benevolent group and a less benevolent group. From the participant's perspective, one group's members are all benevolent, while some of the other group's members are less benevolent. Encountering members of a group of advisors, as opposed to learning about a single advisor, is a closer representation of the situation faced by the agents in scenario 1§6.2 of the evolutionary simulations.

Once again, participants rate the advice before entering their final decision. We expect that advice from the less benevolent group will be less likely to be rated as honest and it will be weighted less in final decision-making.

Open scholarship practices This experiment was preregistered at https://osf.io/qjey5. The experiment data are available in the esmData package for R (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from https://github.com/oxacclab/ExploringSocialMetacognition/blob/1ba333b91366c63a8ab1aed889dd87ea9295a01d/ACv2/dbc.html. In addition to the exclusion criteria listed in the preregistration, we excluded participants who used translation software when visiting the experiment webpage.

7.1.2.1 Method

Procedure The procedure was the same as the procedure for Experiment 5§7.1.1.1. Once again, the two advisor groups were identical in how they generated advice, but they were labelled differently. Advisors had different names and avatar images on every trial. The advisors' background colours indicated which group they were in, with the benevolent advisors' background matching the participant's own. Participants were told at the start of each block which context they were in. Before the block with benevolent advisors they were told that the advisors "will all try their best to help you". Before the block with less benevolent advisors they were told that "some of the advisors may sometimes try to mislead you". Colours and the order in which the advisor groups were encountered were counterbalanced.

Advice profiles Despite differences in labelling, the advisors were identical in terms of how they actually produced advice. The advisors offered advice by placing an 11-year wide marker on the timeline. The marker was placed with its centre on a point sampled from a roughly normal distribution around the correct answer with a standard deviation of 11 years.

Table 7.2: Participant exclusions for Experiment 6

Reason	Participants excluded
Attention check	11
Multiple attempts	3
Missing advice rating	10
Odd advice rating labels	9
Not enough changes	11
Too many outlying trials	3
Total excluded	**25**
Total remaining	**46**

7.1.2.2 Results

Exclusions Participants (total n = 71) could be excluded for a number of reasons: failing attention checks, having fewer than 11 trials which took less than 60s to complete, 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. The latter exclusion was added after data were collected because it was not anticipated that participants would use translation software in the task. The numbers of participants who failed the various checks are detailed in Table 7.2.

The number of participants excluded was quite high. In part, this was due to an unexpectedly high number of participants completing the experiment using translation software. It may also have been due to the study being run on a weekend, whereas most of the other studies were run during the working week, and it is possible that participants using the recruitment platform at the weekend are less well practised at taking experiments than those using it during the week. The final participant list consists of 46 participants who completed an average of 11.93 trials each.

Task performance Participants decreased the error between the midpoint of their answer and the true answer from the initial estimate to the final decision (F(1,45) = 76.09, p < .001; $\text{M}_{\text{Initial}}$ = 15.03 [13.36, 16.71], M_{Final} = 10.89 [9.66, 12.12]), which suggests that they incorporated the (generally accurate) advice (Figure 7.6). The participants had less error on decisions made with the Honest

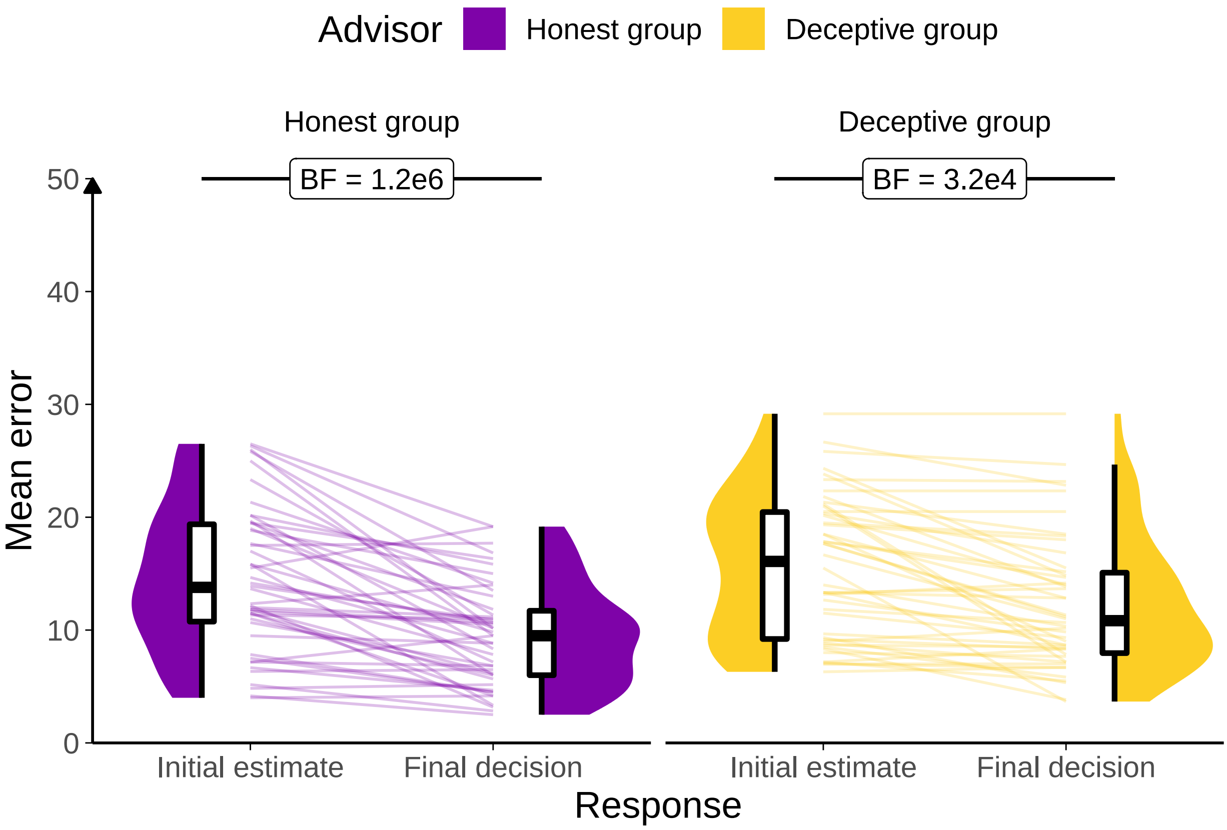


Figure 7.6: Task performance for Experiment 6.
A: Response error. Faint lines show individual participant mean error (the absolute difference between the participant's response and the correct answer), for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of participant means on the original study which this is a replication. The dependent variable here is error, the distance between the correct answer and the participant's answer, and consequently lower values represent better performance. The theoretical limit for error is around 100.

group of advisors than the Deceptive group of advisors (F(1,45) = 6.12, p = .017; $\text{M}_{\text{HonestGroup}}$ = 11.97 [10.49, 13.45], $\text{M}_{\text{DeceptiveGroup}}$ = 13.95 [12.23, 15.68]), although once again the interaction was not significant, so the ANOVA did not demonstrate that this was due to greater reduction in error scores over time for those advisors (F(1,45) = 3.71, p = .060; $\text{M}_{\text{Reduction|HonestGroup}}$ = 4.98 [3.55, 6.41], $\text{M}_{\text{Reduction|DeceptiveGroup}}$ = 3.32 [2.18, 4.46]).

Participants only had one marker they could place, and separate confidence judgements were not asked for, so we cannot directly assess confidence in these data.

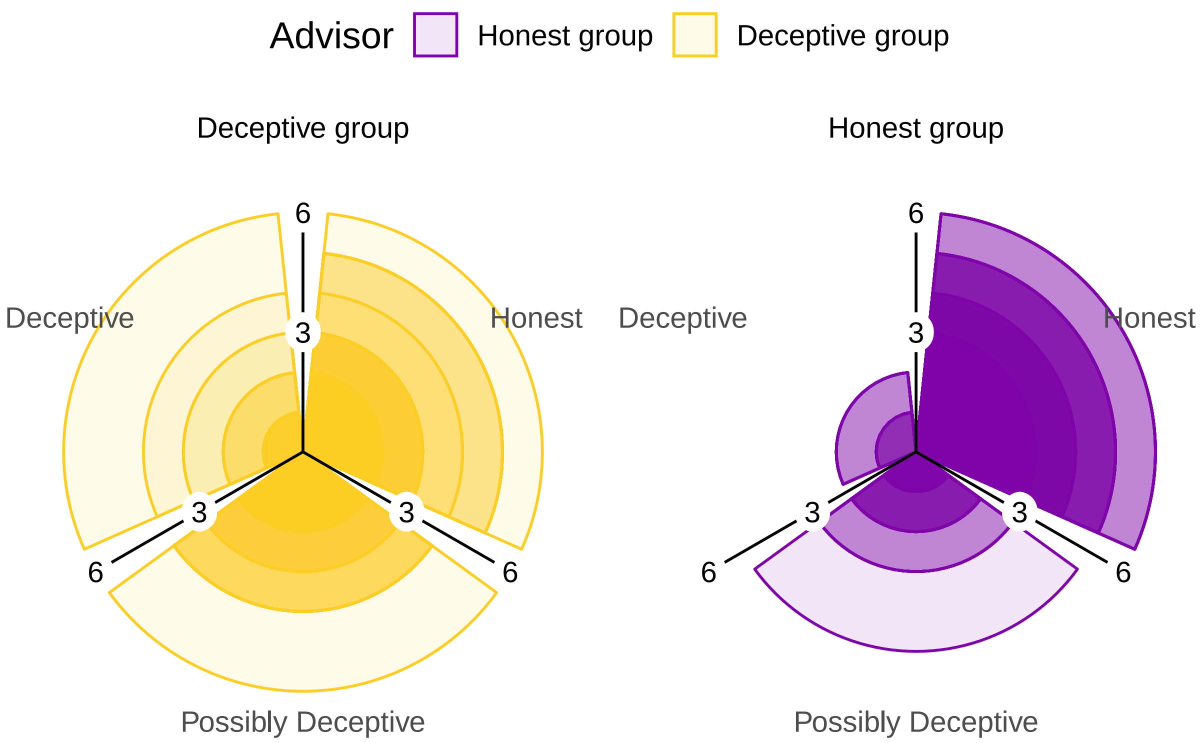


Figure 7.7: Advice rating in Experiment 6.
The polar plots show the number of times each participant gave the given rating to the advice of each advisor. The colour density illustrates the number of participants who gave at least that many ratings to the advice of an advisor.

Advisor performance The advice given by the advisors was generated stochastically from the same distribution. Any differences will be random. This was demonstrably the case for both the domain of advice error (absolute distance between the centre of the advice marker and the correct year; $\mathrm{BF}_{\mathrm{H1:H0}} = 1/4.85$; $\mathrm{M}_{\mathrm{HonestGroup}} = 9.12$ [8.18, 10.06], $\mathrm{M}_{\mathrm{DeceptiveGroup}} = 8.64$ [7.90, 9.38]) and the domain of agreement (absolute distance between the centre of the advice marker and the centre of the participant's initial estimate marker; $\mathrm{BF}_{\mathrm{H1:H0}} = 1/3.94$; $\mathrm{M}_{\mathrm{HonestGroup}} = 17.56$ [15.82, 19.29], $\mathrm{M}_{\mathrm{DeceptiveGroup}} = 18.58$ [16.80, 20.36]).

Advice ratings The advice was rated differently by the participants according to context they were in, as shown in Figure 7.7 ($\chi^2(2) = 40.29$, $p < .001$, $\mathrm{BF}_{\mathrm{H1:H0}} = 8.7\mathrm{e}6$; HonestGroup:DeceptiveGroup ratio: Deceptive 0.47, Possibly Deceptive

0.56, Honest 1.62). This indicates that participants understood the task and that the manipulation worked as intended, with more suspicion applied to the advice from the advisors in the Deceptive group.

Effect of advice Participants chose their own ratings for the advice, and it was common for participants not to use all ratings for all advisors (e.g. many participants never rated advice from the Honest group of advisors as Deceptive). This meant that the statistical tests preregistered for this hypothesis were broken down into separate contrasts of advice and advisor. Using a more complete test, such as 2x3 ANOVA, would have suffered greatly from missing values.

To explore the effect of advice, a 1x3 ANOVA was run on Weight on Advice across ratings. In all, 33/46 (71.74%) participants had at least one trial rated with each of the three ratings. The Weight on Advice differed according to the rating assigned the advice ($\text{F}(2{,}64) = 41.43$, $p < .001$; $\text{M}_{\text{Deceptive}} = 0.06$ [0.03, 0.10], $\text{M}_{\text{PossiblyDeceptive}} = 0.29$ [0.21, 0.36], $\text{M}_{\text{Honest}} = 0.51$ [0.41, 0.61]; Mauchly's test for Sphericity $\text{W} = .940$, $p = .381$). As expected, participants were more influenced by advice they rated as Honest compared to advice they rated as Deceptive.

Effect of advisor To distinguish the effects of the advisor from the effects of the advice, we compared Weight on Advice for only those trials where the participant rated the advice as Honest. This approach has the limitation that 3 (6.52%) participants had to be dropped due to never rating advice from the Deceptive group of advisors as Honest.

Comparing the two (Figure 7.8) showed that participants placed more weight on the Honest advice from advisors in the Honest group ($t(42) = 2.55$, $p = .014$, $d = 0.39$, $\text{BF}_{\text{H1:H0}} = 2.91$; $\text{M}_{\text{HonestGroup}} = 0.55$ [0.47, 0.64], $\text{M}_{\text{DeceptiveGroup}} = 0.44$ [0.34, 0.54]), although the Bayes factor was not above our threshold of 3.

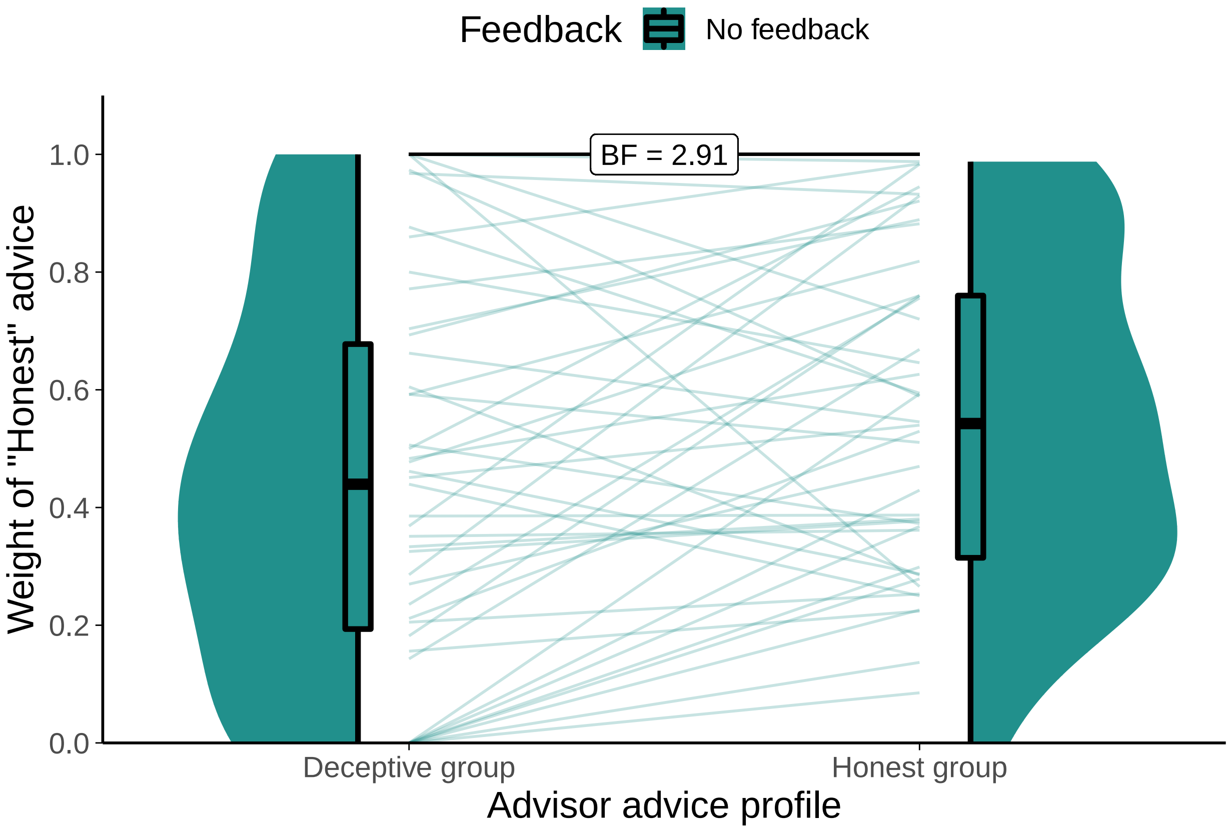


Figure 7.8: Weight on Advice in Experiment 6.
Shows the weight of the advice of the advisors. Only trials where the participant rated the advice as Honest are included. The shaded area and boxplots indicate the distribution of the individual participants' mean Weight on Advice. Individual means for each participant are shown with lines in the centre of the graph.

Exploratory analyses Figure 7.9 shows the overall pattern of advice-taking in the experiment. A linear mixed model was used to better understand the significance of the relationships in the figure, predicting influence on a trial from the advisor, the distance of the advice, and rating (dummy-coded). Random intercepts were included for participant. A similar Bayesian model was included to produce Bayes factors.

The overall pattern showed both advisors were responded to similarly to the Sometimes deceptive advisor in Experiment 5 (Figure 7.5). For advice rated as Honest from the Honest group of advisors, the distance between the initial estimate and the advice was hugely influential, with an increase in distance leading to an approximately equal increase in influence (β = 1.01 [0.94, 1.08], df = 521.77, p < .001, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 4e100). This pattern was neither clearly different nor clearly

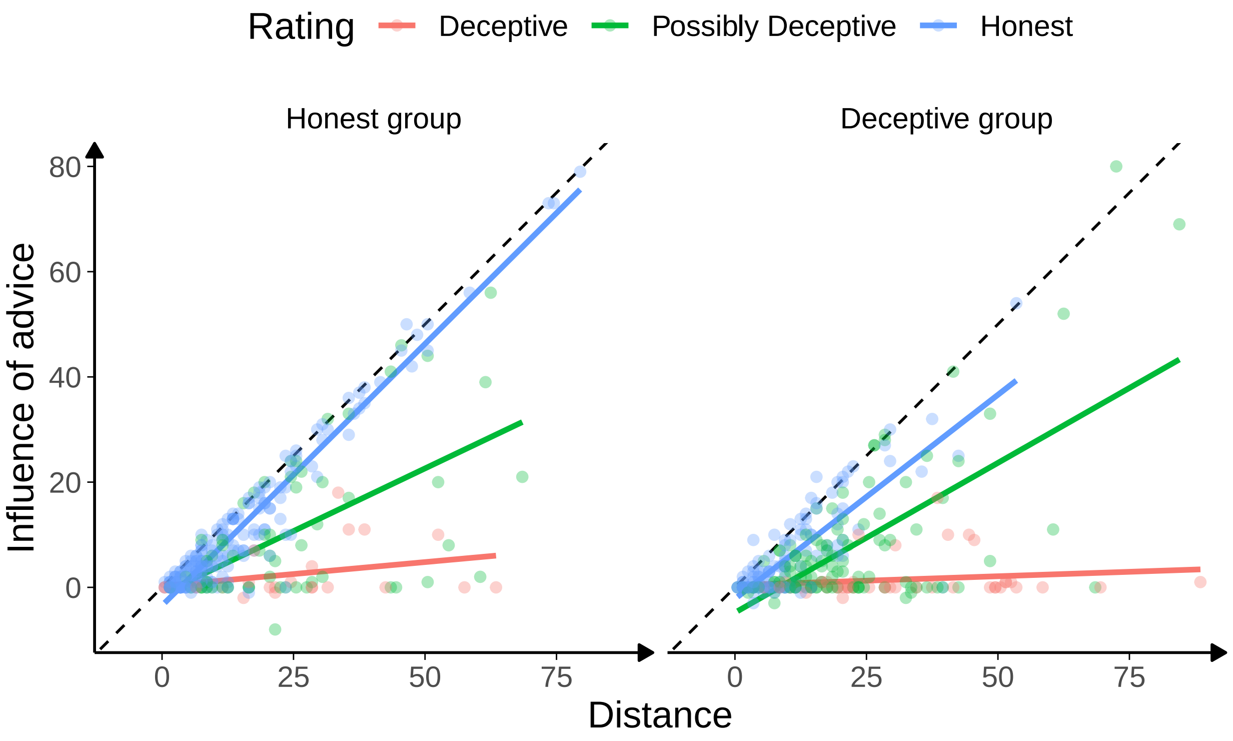


Figure 7.9: Advice response in Experiment 6.
Each point is a single trial, coloured according to the rating given to the advice by the participant. Coloured lines give the best fit linear regression for dots of that colour. The horizontal axis gives the distance between the centre of the participant's initial estimate marker and the centre of the advisor's advice marker, and the vertical axis gives the distance the participant's final decision had moved in the direction of the advisor's advice. Points that lie on the dashed line indicate that the participant moved their marker all the way to the advisor's marker, thus wholly adopting the advisory estimate as their final decision.

the same for advice rated as Honest from the Deceptive group of advisors (β = -0.22 [-0.37, -0.08], df = 522.52, p = .003, $\text{BF}_{\text{H1:H0}}$ = 1/1.18).

Compared to advice rated as Honest, advice rated as Deceptive had a much flatter relationship with distance (when coming from the Honest group of advisors: β = -0.91 [-1.09, -0.73], df = 521.92, p < .001, $\text{BF}_{\text{H1:H0}}$ = 5.7e30; when coming from the Deceptive group of advisors: β = 0.17 [-0.08, 0.42], df = 524.15, p = .177, $\text{BF}_{\text{H1:H0}}$ = 1/4.54). Advice rated as Possibly Deceptive did not show this nearly flat relationship, but neither was it clearly the same as advice rated Honest in its relationship with distance (β = -0.52 [-0.64, -0.39], df = 522.22, p < .001, $\text{BF}_{\text{H1:H0}}$ = 2.2e9), and it is likely more evidence would reveal a subtler pattern in the direction of that found for advice rated as Deceptive. The relationship of advice rated as Possibly Deceptive to distance was, however, different between advisor groups, with the Possibly Deceptive advice being considered more trustworthy as it got more distant when it came from the Deceptive group of advisors (β = 0.32 [0.12, 0.51], df = 520.54, p = .002, $\text{BF}_{\text{H1:H0}}$ = 12.7). This surprising result may be due to participants applying different criteria to judging the deceptiveness of advice according to the advisors giving it.

More clearly visible in Figure 7.9 than Figure 7.5, but present in both, is the tendency for participants to rate advice as Deceptive more frequently as it is further away from their initial estimates. Notably, this appears more common for the Sometimes deceptive advisor and the Deceptive group of advisors. For the Always honest advisor and the Honest group of advisors, even advice that is far away from the participant's initial estimate is often rated as and treated as honest.

7.1.2.3 Discussion

This experiment followed up on the findings in Experiment 5§7.1.1.3 by presenting participants with a context in which advisors were more or less likely to be benevolent, rather than having participants form a relationship with an advisor who was more or less likely to be benevolent. This shift from relationship to context means that the experiment is slightly closer to scenario 1 in our evolutionary models§6.2.

Whereas in Experiment 5 we saw that both the advice and its source mattered for how much participants followed advice, and we saw a suggestion of the same in Experiment 6. In both cases participants were more likely to appraise advice as helpful (and to follow it accordingly) when the source of the advice was trustworthy. In terms of our simulations, this suggests that the effect of context may increase egocentric discounting more through increasing the frequency with which judges wholly disregard advice than through decreasing the extent to which they take advice uniformly across all interactions. Put simply, they are more on their guard for bad advice, but once they accept advice is good they may treat it fairly normally.

This differs somewhat from the effect of building a relationship with a single advisor. In this case, while advice is judged similarly warily when it might be deceptive, even advice judged as Honest is accepted much more hesitantly. Both ignoring advice more frequently and accepting advice more tentatively appear the same when the advice-taking is represented using an average over multiple trials. Overall, the results of Experiments 5 and 6 illustrate that people respond flexibly to changes in the likely benevolence of advice, and that this can affect their relationships with individual advisors.

7.2 Noise in the advice

The second scenario§6.3 explored in the evolutionary models added noise to the advice agents received and demonstrated that this provided an evolutionary pressure towards egocentric discounting. 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. This scenario is not explored in behavioural experiments because its conclusions are well supported by existing literature. Specifically, the literature demonstrates 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. This demonstration is supplemented with a more comprehensive model of the utility of advice according to the ability to identify the more expert opinion, the relative accuracy of the advisor and judge, and the independence of the opinions (Soll and Larrick 2009). A more complete account of the effects of advisor expertise was provided earlier§5.2.3.1.

Views of relative expertise may also be affected by self-enhancement bias wherein people typically assess their own abilities to be above average (Brown 1986). We expect that perception of relative expertise 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 (Gino and Moore 2007). Taken together, then, people are likely to consider themselves more able on a given task than an arbitrary advisor, and consequently that they are likely to down-weight advice relative to their own initial estimate. This behaviour is supported by normative models which show biasing towards the better estimator (in this case the judge) is the optimal strategy (Soll and Larrick 2009; Mahmoodi et al. 2015). As discussed in scenario 2§6.3.3, 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, whereas advisors do not.

7.3 Confidence mapping

The third scenario§6.4 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 versus 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.

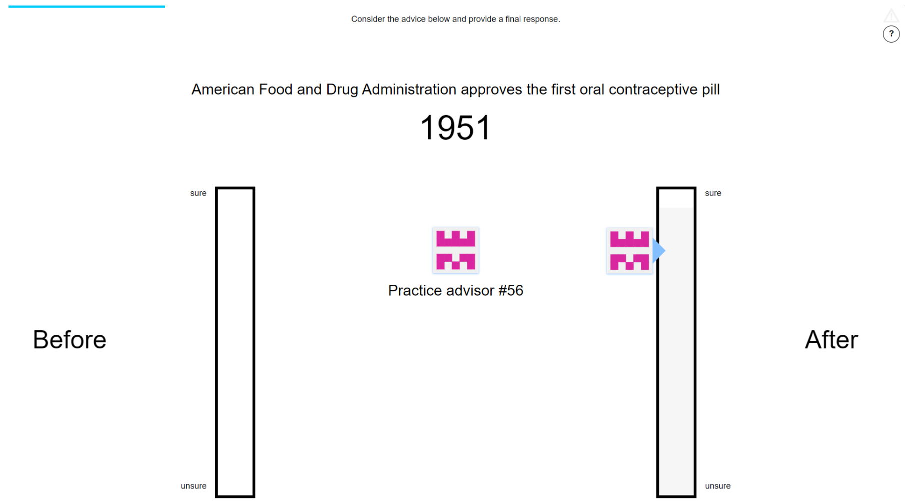
We attempted to investigate this in two ways. Firstly, we performed an experiment where participants received advice over blocks of trials from either the same advisor on each trial or a different advisor on each trial (Experiment 7§7.3.1). Repeated interactions with a single advisor should allow a participant to understand the advisor's confidence mapping, whereas if the advisor changes on every trial the participants instead have to interpret the advice in a more generic manner. Secondly, we familiarised participants with two advisors who differed in their use of the confidence scale, and then tested whether participants responded differently based on whether they could identify which of those two advisors was providing advice (Experiment 8§7.3.2).

To look ahead briefly, neither experiment provided evidence consistent with the hypothesis that familiarity with an advisor's confidence calibration leads to greater trust in that advisor's advice. We suggest that this may be a failure in experimental design rather than reliable evidence of the absence of this feature of advice-taking behaviour.

7.3.1 Experiment 7: individuality as a cue to confidence

The expression of confidence is highly idiosyncratic (Ais et al. 2016; Navajas et al. 2017). Given this, we hypothesised that the same advice may be treated differently depending upon whether or not it was labelled as coming from repeated interactions with a single individual or from a series of one-off interactions with different individuals. Where the advice comes from the same individual over and over again, participants should become more aware of that individual's confidence calibration (the relationship between their confidence and the probability they are correct), and should thus be able to better discriminate higher-quality (more well calibrated) advice.

In this experiment the effect of individuality is confounded with familiarity, a feature that we expect will increase the palatability of advice. While participants are learning about the advisors, and learning more about the confidence calibration of the individual advisor than that of the group members, they are also becoming more familiar with the individual advisor. Nevertheless, this is a useful experiment because it is capable of indicating the absence of an effect: if the manipulation proves unable to produce greater influence from the individual advisor it will do so despite rather than because of the familiarity confound. If the experiment demonstrates higher influence for the individual advisor then further experiments can be run to attempt to isolate confidence calibration knowledge from mere exposure.



Consider the advice below and provide a final response.

American Food and Drug Administration approves the first oral contraceptive pill

1951

Figure 7.10: Advice confidence.
Advisors' markers shifted position vertically to indicate the confidence with which the advice was given. As with the participants' own responses, higher placement on the column indicated higher confidence.

7.3.1.1 Open scholarship practices

This experiment was not preregistered because it was highly exploratory. The experiment data are available in the `esmData` package for R (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from `https://github.com/oxacclab/ExploringSocialMetacognition/blob/4d8aace6b864c4465cacd8579ec1abd870dd65d2/ACBin/ce.html`.

7.3.1.2 Method

Procedure 33 participants each completed 40 trials over 5 blocks of the binary version of the Dates task§2.1.3.2. In these experiments, advisors gave advice that included confidence. The advisor's confidence in their response was indicated by the height of the advisor's avatar on the indicated bar (Figure 7.10).

Participants started with 1 block of 7 trials that contained no advice. All trials in this section included feedback for all participants indicating whether or not their response was correct.

Participants then did 3 trials with a practice advisor. They also received feedback on these trials. They were informed that they would "get advice about the correct answer" and that the "advisors indicate their confidence in a similar way to you, by placing their marker higher on the scale if they are more sure".

Participants then performed 4 blocks of trials that constituted the main experiment. These blocks came in pairs, with each pair containing one block of 10 trials that included feedback followed by one block of 5 trials that did not include feedback. The first block served to allow the participants to familiarise themselves with the confidence calibration of their advisor in the individual case, or to experience the range of advisors available in the group case. The feedback on these trials allowed participants to recognise the usefulness of the advice, and potentially to observe its relationship between higher confidence and higher accuracy. The second block allowed us to test the influence of advice without further feedback learning effects. 1 trial in the first block in each pair was an attention check trial. Each pair of blocks had a different advisor presentation.

In one of the pairs of blocks, the advice was always labelled as coming from the same advisor. In the other pair of blocks, the advice on each trial was labelled as coming from a different advisor. The order that the participants encountered these presentations in was counterbalanced.

Advice profiles Despite differences in labelling, the advisors were identical in terms of how they actually produced advice because we were interested in whether, when there was the same thing to learn (calibration of advice confidence) it would be learned when interacting with a single advisor but not when interacting with a group of advisors. The advice was determined on each trial by selecting a point drawn from a normal distribution around the correct answer with standard deviation of 8 years. This point served as the advisor's internal estimate of the correct answer. The advisor then gave the appropriate answer according to whether or not the internal estimate was before or after the anchor year, and linearly scaled their confidence based on the distance between those two years. The advisor was maximally confident where the difference between the internal estimate and the anchor year was at least 20 years.

Table 7.3: Participant exclusions for Experiment 7

Reason	Participants excluded
Multiple attempts	0
Not enough changes	3
Too many outlying trials	3
Total excluded	**6**
Total remaining	**27**

7.3.1.3 Results

Exclusions Participants (total n = 33) were excluded for a number of reasons: having fewer than 10 no-feedback trials which took less than 60s to complete, or providing final decisions which were the same as the initial estimate on more than 90% of trials. Table 7.3 shows the number of participants excluded and the reasons why. Participants who failed attention checks would have been immediately dropped, but no participants did so.

The final participant list consists of 27 participants who completed an average of 28.00 trials each.

Task performance We present an overview of participants' performance in the task in Figures 7.11 and 7.12 and detail some key features using statistical tests. The key analysis of our hypothesis, exploring the effect of advisor presentation, is presented below§7.3.1.3.

A first analysis assessed differences participants' accuracy, using a within-participant ANOVA with factors of answer (initial estimate versus final decision) and advisor (Consistent individual versus Group member). Accuracy during the Familiarisation trials (Figure 7.11) was higher for final decisions than for initial estimates (F(1,26) = 31.69, p < .001; $\mathrm{M}_{\mathrm{FinalDecision}}$ = 0.72 [0.68, 0.76], $\mathrm{M}_{\mathrm{InitialEstimate}}$ = 0.58 [0.51, 0.64]), so participants' accuracy improved after advice. Accuracy was also higher for the Consistent individual advisor than for the Group members (F(1,26) = 6.35, p = .018; $\mathrm{M}_{\mathrm{ConsistentIndividual}}$ = 0.69 [0.63, 0.75], $\mathrm{M}_{\mathrm{GroupMember}}$ = 0.60 [0.55, 0.66]). There was no significant effect of advisor presentation on improvement (F(1,26) = 2.81, p = .106; $\mathrm{M}_{\mathrm{Improvement|ConsistentIndividual}}$ = 0.10 [0.04,

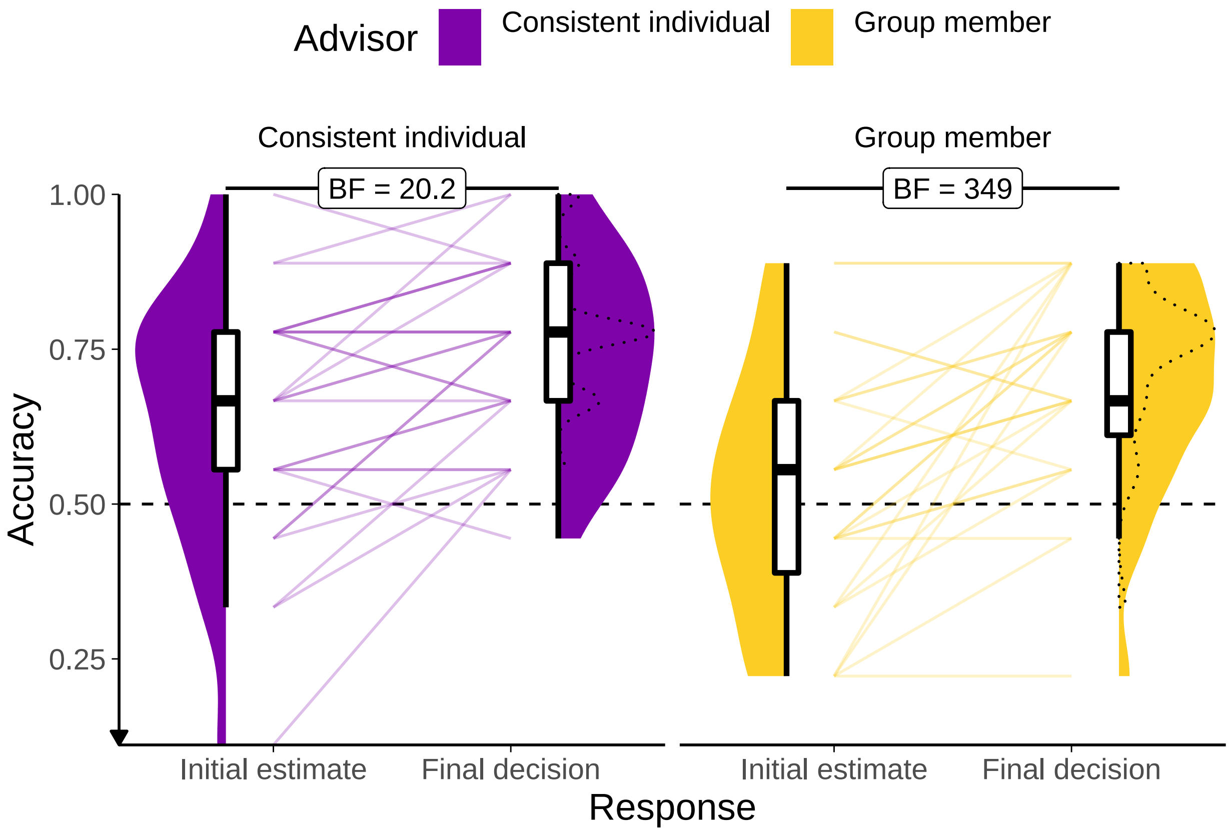


Figure 7.11: Response accuracy for Experiment 7.
Faint lines show individual participant means, for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.
Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

0.16], $\mathrm{M}_{\mathrm{Improvement|GroupMember}}$ = 0.18 [0.10, 0.26]). This peculiar pattern was driven by a significant difference in participants' accuracy rates between advisors on their initial decisions, i.e. before they had seen advice ($\mathrm{BF}_{\mathrm{H1:H0}}$ = 14.4; $\mathrm{M}_{\mathrm{ConsistentIndividual}}$ = 0.64 [0.56, 0.72], $\mathrm{M}_{\mathrm{GroupMember}}$ = 0.51 [0.44, 0.59]). It is not clear why this occurred; it was likely the result of random chance.

A second ANOVA was run to explore the effects of advisor influence, using factors of advisor (Consistent individual versus Group member) and stage (Familiarisation versus Test). The influence of advisors might differ according to whether participants

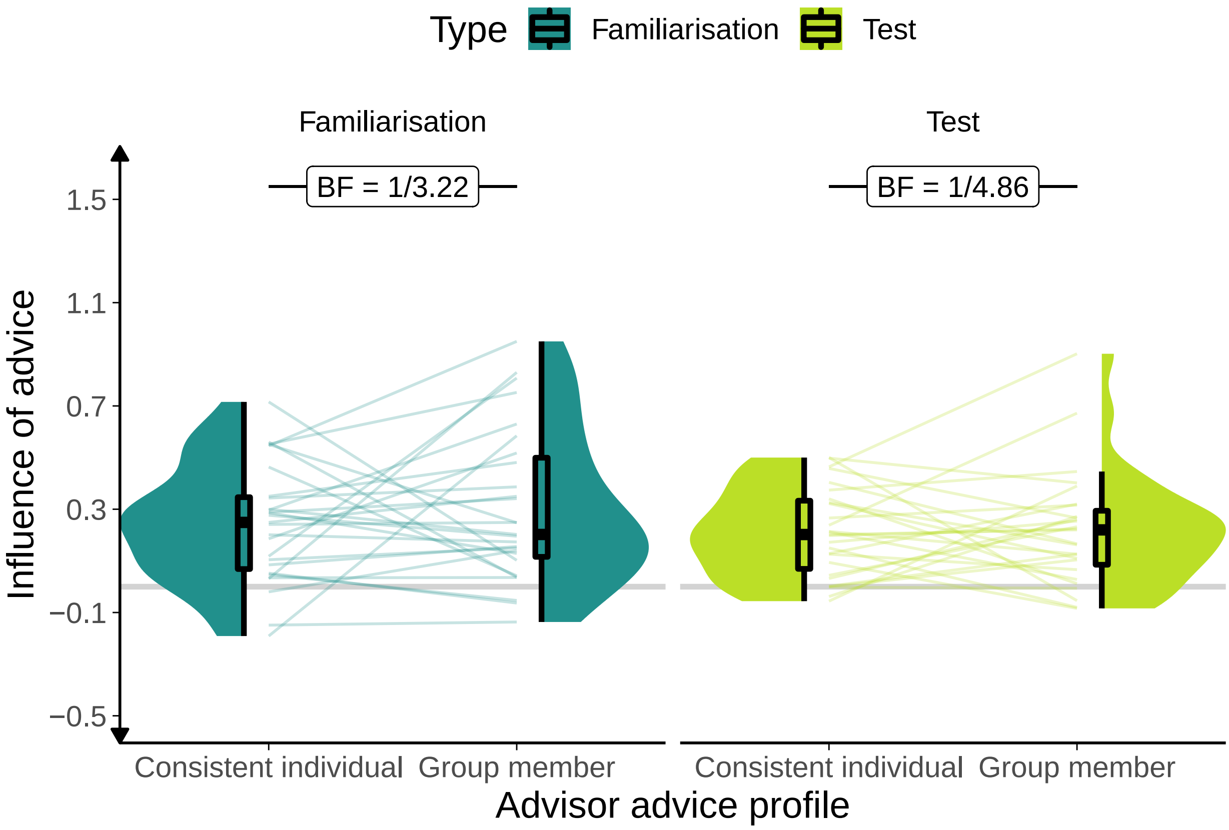


Figure 7.12: Influence for Experiment 7.
Participants' weight on the advice for advisors in the Familiarisation and Test stages of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

had become familiar with them, hence separating out the initial Familiarisation trials (that contained feedback) from the later Test trials (that did not) might provide some evidence of learning. No significant effects were found for influence (Figure 7.12). Neither the advisor presentation (F(1,26) = 0.61, p = .442; $M_{\text{ConsistentIndividual}}$ = 0.23 [0.16, 0.29], $M_{\text{GroupMember}}$ = 0.26 [0.17, 0.35]), nor whether the trial was in the Familiarisation stage (with feedback) or the test stage (without feedback) appeared to be systematically linked to the influence of the advice (F(1,26) = 3.62, p = .068; $M_{\text{Familiarisation}}$ = 0.27 [0.19, 0.35], M_{Test} = 0.21 [0.15, 0.27]). There was also no indication that the influence of the advisors' advice changed differently between stages (F(1,26) = 0.64, p = .432; $M_{\text{Individual-Group|Familiarisation}}$ = -0.06 [-0.20, 0.07], $M_{\text{Individual-Group|Test}}$ = -0.01 [-0.10, 0.09]).

Advisor performance The advice is generated probabilistically using the same rules for both advisor presentations. The advisors' advice is not dependent on the participants' initial estimates, but it is useful nevertheless to compare the advisors' agreement and accuracy rates. The advisors were demonstrably similar in their overall agreement rates ($\text{BF}_{\text{H1:H0}}$ = 1/4.74; $\text{M}_{\text{ConsistentIndividual}}$ = 0.59 [0.52, 0.66], $\text{M}_{\text{GroupMember}}$ = 0.60 [0.53, 0.68]), but not so in their accuracy rates ($\text{BF}_{\text{H1:H0}}$ = 1/2.24; $\text{M}_{\text{ConsistentIndividual}}$ = 0.78 [0.74, 0.82], $\text{M}_{\text{GroupMember}}$ = 0.74 [0.69, 0.79]). They were not demonstrably *different* in their accuracy rates, however, so the randomisation worked sufficiently well to allow the results to be interpreted.

Effect of advisor presentation The hypothesis for this study was that, in the key Test trials where participants did not receive feedback, the advisor presented as a consistent individual would be more influential than the advisor presented as different members of a group on each trial. As illustrated in the right-hand panel of Figure 7.12, influence was equivalent between the advisor presentations on the Test trials ($t(26)$ = -0.14, p = .886, d = 0.03, $\text{BF}_{\text{H1:H0}}$ = 1/4.86; $\text{M}_{\text{ConsistentIndividual}}$ = 0.21 [0.14, 0.28], $\text{M}_{\text{GroupMember}}$ = 0.22 [0.13, 0.30]). From this we can conclude that the manipulation was not effective in differentiating the advisors or that such differences as were produced do not affect the influence of advice.

7.3.1.4 Discussion

This experiment provided evidence against the prediction that participants would take more advice where they could learn about an individual advisor's confidence calibration. We presented participants with the same advice, labelled either as always coming from the same advisor or as coming from a different advisor on each trial. We had expected that presenting participants with a single advisor would allow them to learn about that advisor's ability to perform the task, and how well their confidence was calibrated (i.e. how strong was the relationship between increased confidence and increased probability of being correct). However, contrary to our key prediction, we did not observe differences in the influence of advice as a function of the presentation of the advisor.

Two reasons why we failed to detect the effects may be that the manipulation was not powerful enough or that there is no systematic relationship between calibration knowledge (and familiarity more generally) and advice influence. The manipulation may not have been strong enough because participants may have paid relatively little attention to the identity of the advisor and focused more on the content of the advice. Participants may also have learned about the calibration distributions just the same for individual and group advisors: statistically the experience is the same and some research has suggested that statistical mechanisms such as associative learning may be sufficient to explain social learning phenomena (Behrens et al. 2008; FeldmanHall and Dunsmoor 2019). It is also possible, though we suspect less likely given prior evidence of sensitivity to calibration (Pescetelli and Yeung 2021), that the participants were simply unable to distinguish the advisors even in principle, or that any distinction they did make failed to translate into differential influence.

7.3.2 Experiment 8: identifiability of advice

The results of Experiment 7 were not promising. Consequently, we used a different manipulation in an attempt to place the calibration of the advice in the forefront of participants' experience. To this end, we familiarised participants with two advisors whose accuracy, probability of agreement, and confidence calibration were identical, but whose use of the confidence scale differed. A Low confidence advisor used the bottom two-thirds of the confidence scale, while a High confidence advisor used the top two-thirds.

On Key trials, participants frequently saw advice where the confidence lay in the middle third of the scale. On some of these trials, the advice was labelled as coming from a specific advisor. If the advice came from the Low confidence advisor it would represent high confidence advice, and imply a high probability that it was correct. If the advice came from the High confidence advisor, it would represent low confidence advice, and imply a low probability that it was correct. On the remainder of these trials, the advice was labelled as coming from an unknown one of those two advisors. We expected that advice that was labelled in this way would be treated differently relative to advice that was identifiable because it was impossible for participants to appropriately interpret the confidence (and therefore probability of correctness) of the advice.

7.3.2.1 Open scholarship practices

This experiment was not preregistered because it was exploratory. The experiment data are available in the `esmData` package for R (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from `https://github.com/oxacclab/ExploringSocialMetacognition/blob/9ac26116e348f8c4ff5617a8e68710b889bc7e08/ACBin/ck.html`.

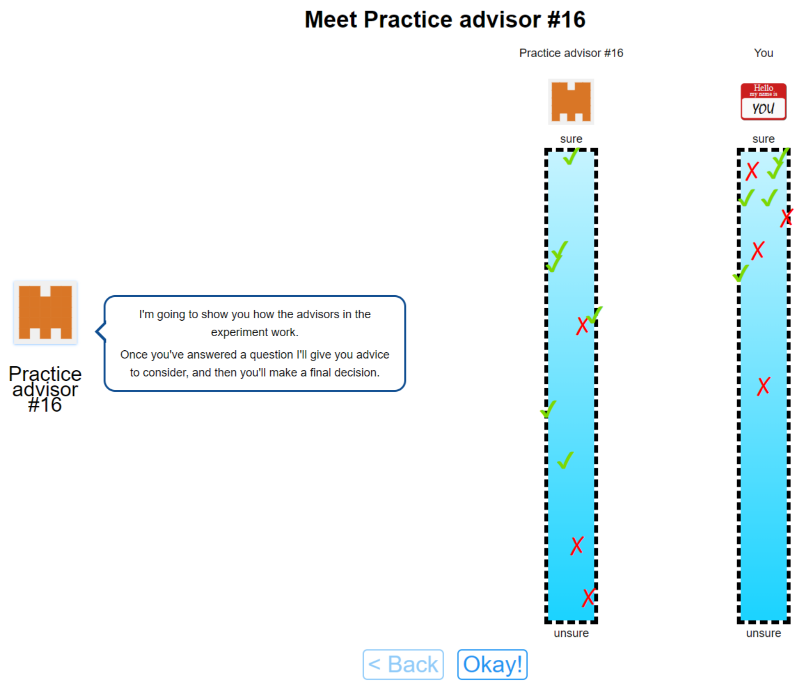
7.3.2.2 Method

Procedure 48 participants each completed 67 trials over 5 blocks of the binary version of the Dates task§2.1.3.2. The procedure for filling in responses and the way the advice was represented was the same as described for Experiment 7§7.3.1.

Participants started with 1 block of 10 trials that contained no advice, to allow them to familiarise themselves with the task. All trials in this section included feedback for all participants indicating whether or not their response was correct.

Participants then did 9 trials with a practice advisor to get used to receiving advice. They also received feedback on these trials. They were informed that they would "get advice about the correct answer" and that the "advisors indicate their confidence in a similar way to you, by placing their marker higher on the scale if they are more sure".

Before performing trials with an advisor, participants saw the advisor's avatar alongside a speech bubble in which the advisor introduced themself. On the same screen, participants saw a scorecard for that advisor (Figure 7.13). Scorecards were introduced to the participant after the initial practice (without advice).



Meet Practice advisor #16

I'm going to show you how the advisors in the experiment work.

Once you've answered a question I'll give you advice to consider, and then you'll make a final decision.

Figure 7.13: Advice scorecard for Experiment 8.
The scorecards indicated the relationship between confidence and accuracy, allowing participants to see how well calibrated they or an advisor was.

Participants then performed 3 blocks of trials that constituted the main experiment. In the first 2 blocks participants were familiarised with their advisors over 10 trials each (1 of which was an attention check). The order in which participants encountered the advisors was counterbalanced, and participants received feedback during these Familiarisation trials.

Finally, participants performed 1 block of 28 Test trials that did not include feedback about the correct answer. Before this block, participants were reminded again about their advisors, and shown the introduction screen with the scorecards for each advisor again. The scorecards were shown individually, and then displayed on the same screen next to one another, and the participants invited to "compare them with one another". During these Test trials, advice was sometimes labelled as coming from a specific one of the advisors encountered during the Familiarisation trials, and sometimes the advice was labelled as coming from an unspecified one of those advisors.

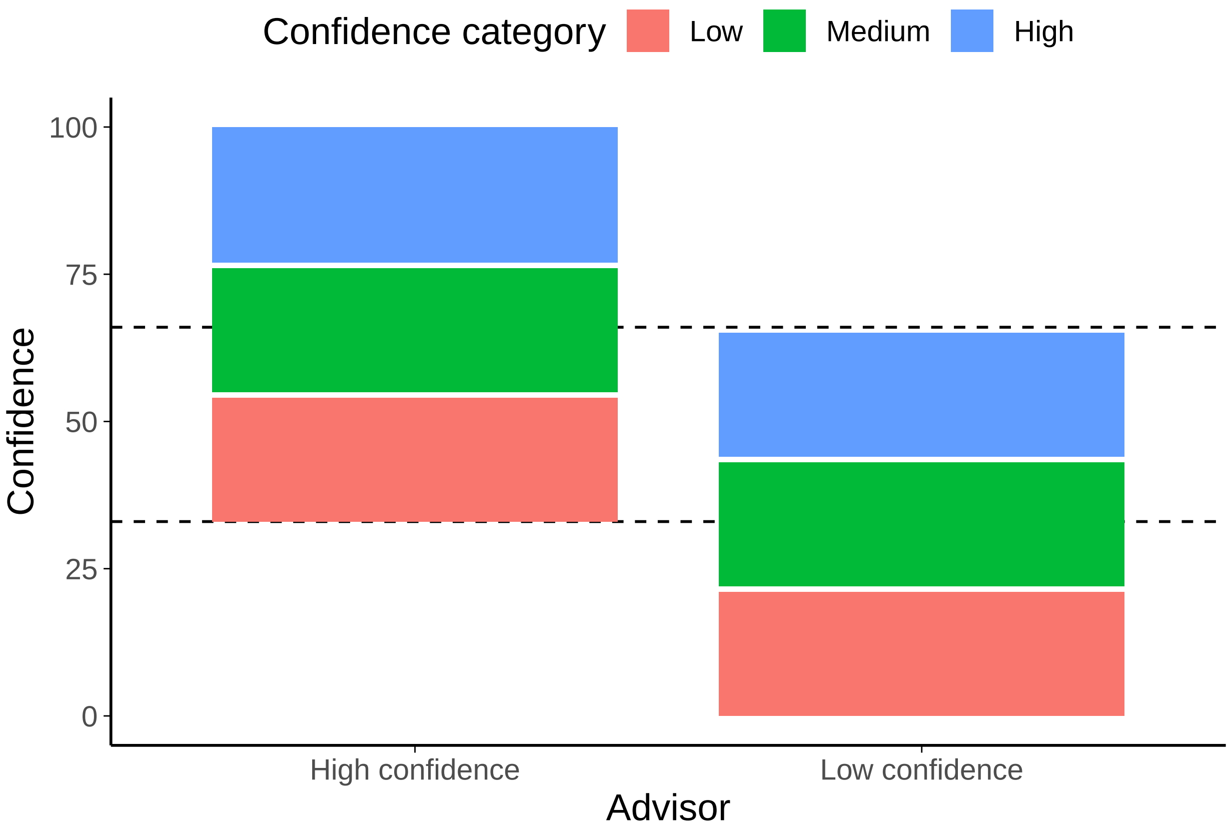


Figure 7.14: Advice distribution map for advisors in Experiment 8.
The High confidence equates to 100% correct, Medium to 66%, and Low to 33%. The dashed lines show the overlap area between the minimum confidence for the High confidence advisor and the maximum confidence for the Low confidence advisor. Note that the advisors are not equally correct on average in this zone: the Low confidence advisor is substantially more accurate because they are idiosyncratically more confident.

Advice profiles There are two advisors in this experiment, a High confidence advisor and a Low confidence advisor. Before the Familiarisation block with each advisor, the advisors introduced themselves with a statement about their advice-giving style. The High confidence advisor stated "I'm confident and outgoing. I tend to make up my mind and stick to it." The Low confidence advisor stated "I'm cautious and careful. I try to keep an open mind even when I've made my choice."

The Advice profiles (Figure 7.14) for the two advisors were balanced in terms of the information provided by each advisor: both advisors had a subjectively high confidence at which they were 100% accurate, a medium confidence range within which they were 66% accurate, and a low accuracy range in which they were

33% accurate. Crucially, these ranges span different parts of the confidence scale; the Low confidence advisor never expressed confidence above 66%, and the High confidence advisor never expressed confidence below 33%.

The advisor's advice during the Familiarisation phase was choreographed to provide all participants with an exact insight into the underlying calibration. Each advisor offered three estimates in each of their three confidence zones. All of the high-confidence advice was correct, two of the medium-confidence, and one of the low-confidence.

During the Test phase, advisors provided advice while either labelled unambiguously, as they had been during the Familiarisation phase, with their avatar and advisor number, or ambiguously, with a hybrid avatar and question marks instead of the advisor number. Each advisor offered 7 pieces of advice while labelled unambiguously, and 7 while labelled ambiguously. Each of these sets of 7 advice had the following structure: three trials in the third of the confidence scale unique to the advisor that were correct, two trials in the ambiguous third of the confidence scale where the advisor agreed with the participant's initial estimate, and two trials in the ambiguous third where the advisor disagreed with the participant's initial estimate.

The decision to include advice outside the ambiguous third of the confidence scale (between the dashed lines in Figure 7.14)) was taken with the aim of ensuring participants' experience of the advice during the Test phase felt natural. Unfortunately, this meant that the advisors either had to degrade their calibration or they had to have different accuracy rates (either overall or within the ambiguous third of the confidence scale). Between these two lemmas we opted for the former, meaning that during the Test phase the Low confidence advisor's calibration changed such that all low-confidence advice was accurate.

The most important feature of this complex advice profile construction is that there are an equal number of trials that have advice with confidence in the ambiguous range for each advisor and for each presentation (ambiguous or identifiable). Within these trials, agreement is also tightly controlled meaning that each advisor-presentation combination agrees on half the trials and disagrees on half

Table 7.4: Participant exclusions for Experiment 8

Reason	Participants excluded
Multiple attempts	2
Not enough changes	13
Study incomplete	12
Too many outlying trials	0
Total excluded	**15**
Total remaining	**33**

the trials. These 16 Key trials are central to testing our hypothesis that participants will be less influenced by advice when the source is ambiguous, so it is important that they are as balanced as possible for agreement and number.

7.3.2.3 Results

Exclusions Participants (total $n = 48$) were excluded for a number of reasons: having fewer than 16 Key trials which took less than 60s to complete, or providing final decisions which were the same as the initial estimate on more than 90% of trials. Table 7.4 shows the number of participants excluded and the reasons why.[1] Participants who failed attention checks do not reach this stage of the experiment.

The final participant list consists of 33 participants who completed an average of 45.91 trials each. Many participants were excluded using our typical exclusion criteria, but the results of the main statistical tests are the same when participants who seldom altered their responses are left in the sample, even where they did not complete the entire study.

Task performance We first characterise the participants' performance on the task. We investigated participant accuracy in the Familiarisation trials using a 2x2 ANOVA with within-participant factors of answer time (initial estimates versus final decisions) and advisor (High confidence versus Low confidence). Participants increased their accuracy (Figure 7.15) from initial estimate to final decision (F(1,32) = 29.53, $p < .001$; $\mathrm{M}_{\mathrm{FinalDecision}} = 0.75$ [0.71, 0.79], $\mathrm{M}_{\mathrm{InitialEstimate}} = 0.66$ [0.60, 0.71]). There was no evidence of a difference between the advisors in either overall

[1] Note that no participants were excluded only for failing to finish the study.

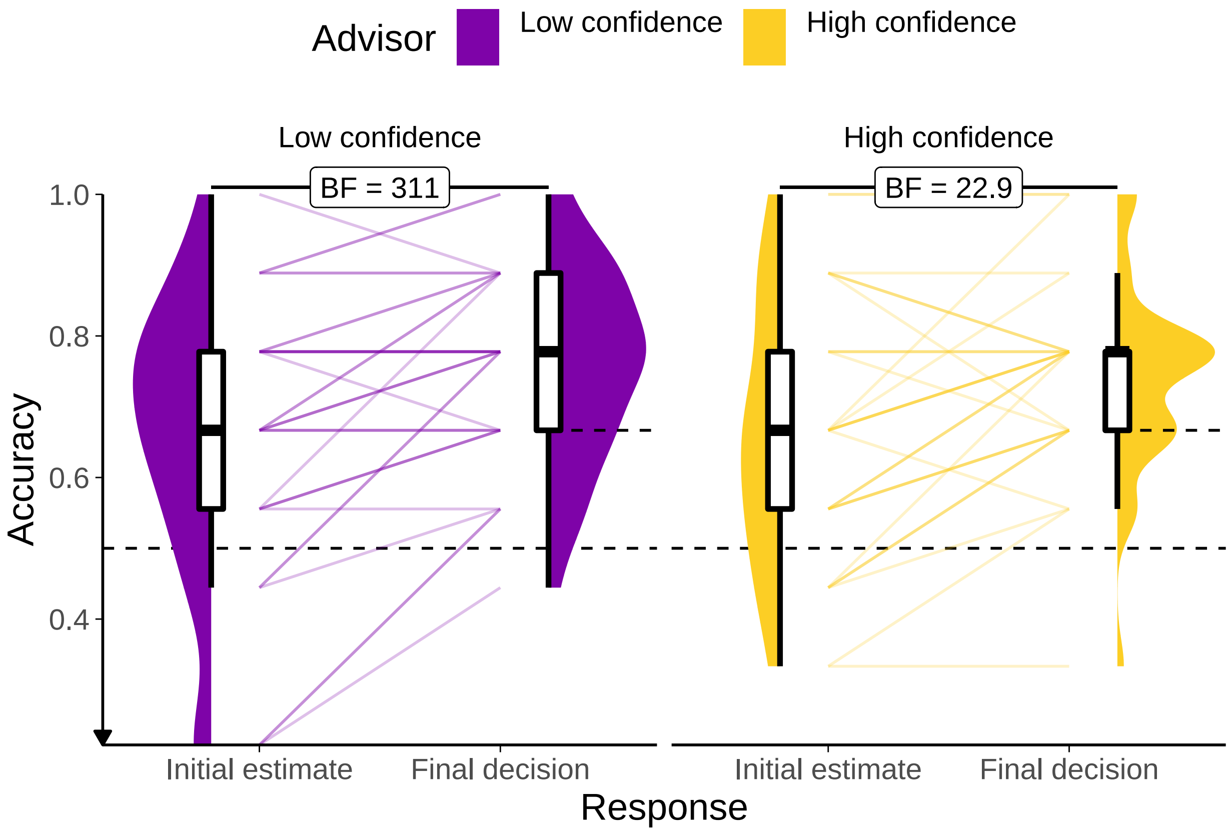


Figure 7.15: Response accuracy for Experiment 8.
Faint lines show individual participant means, for which the violin and box plots show the distributions. The full width dashed line indicates chance performance, and the half width dashed lines indicate advisor accuracy. Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

accuracy (F(1,32) = 0.09, p = .770; $\mathrm{M}_{\mathrm{HighConfidence}}$ = 0.70 [0.65, 0.75], $\mathrm{M}_{\mathrm{LowConfidence}}$ = 0.71 [0.65, 0.76]) or in the increase in accuracy following advice (F(1,32) = 0.28, p = .600; $\mathrm{M}_{\mathrm{Improvement|HighConfidence}}$ = 0.08 [0.03, 0.13], $\mathrm{M}_{\mathrm{Improvement|LowConfidence}}$ = 0.10 [0.06, 0.15]).

We also investigated the participants' confidence on Familiarisation trials using a 2x2 ANOVA with factors of answer time (initial estimates versus final decisions) and initial estimate correctness (initial estimate was correct versus initial estimate was incorrect). This analysis allows us to get a sense of whether their confidence

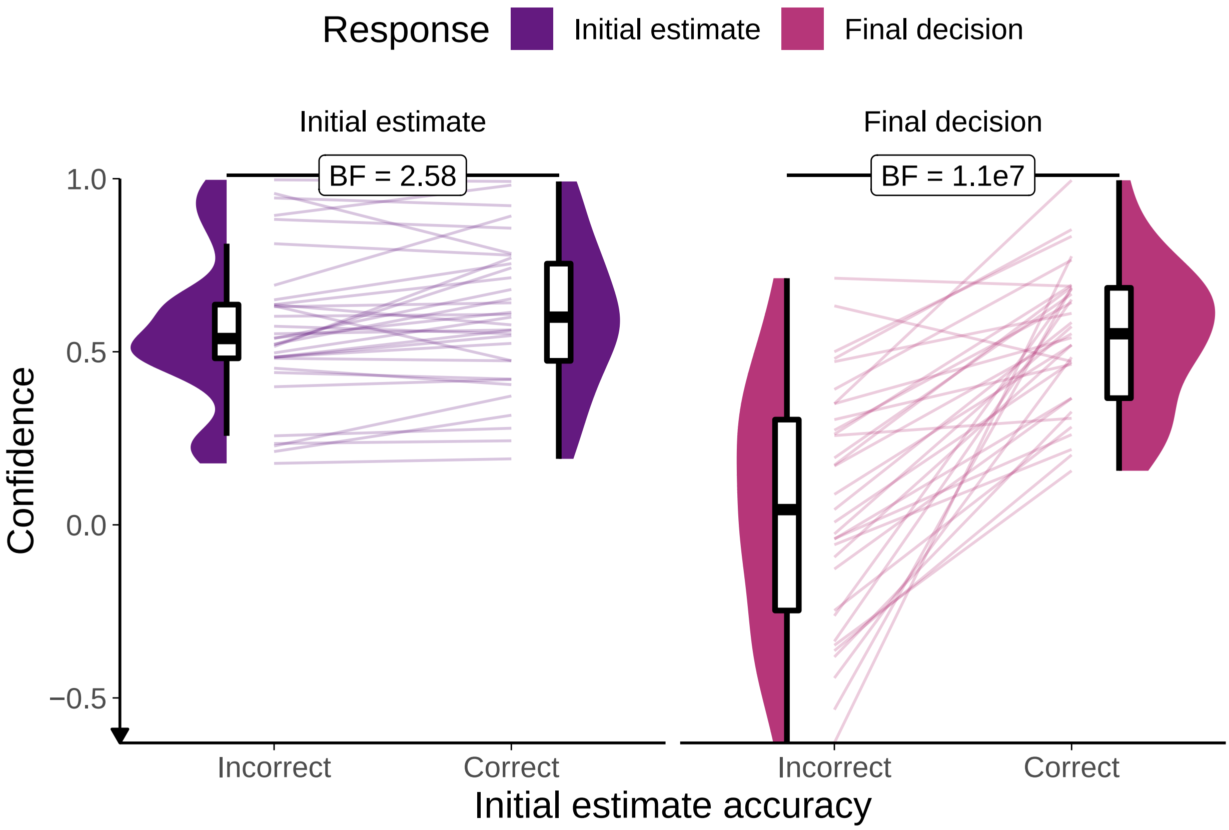


Figure 7.16: Confidence for Experiment 8.
Faint lines show individual participant means, for which the violin and box plots show the distributions. Final confidence is negative where the answer side changes. Theoretical range of confidence scores is initial: [0,1]; final: [-1,1].

changes systematically after encountering advice and whether it does so more when their initial decision was correct. Participants were more confident (Figure 7.16) in their responses when their initial estimate was correct (F(1,32) = 63.93, p < .001; $\mathrm{M}_{\mathrm{Correct}}$ = 0.57 [0.51, 0.64], $\mathrm{M}_{\mathrm{Incorrect}}$ = 0.31 [0.23, 0.39]). They were systematically less confident on final decisions compared to initial estimates (F(1,32) = 46.58, p < .001; $\mathrm{M}_{\mathrm{InitialEstimate}}$ = 0.58 [0.51, 0.66], $\mathrm{M}_{\mathrm{FinalDecision}}$ = 0.30 [0.21, 0.38]), and that confidence decrease was larger when the initial estimate was incorrect (F(1,32) = 72.89, p < .001; $\mathrm{M}_{\mathrm{Decrease|Correct}}$ = 0.06 [-0.01, 0.13], $\mathrm{M}_{\mathrm{Decrease|Incorrect}}$ = 0.51 [0.38, 0.64]). This indicates that participants were adjusting their confidence in a rational manner. There is always more scope for adjusting confidence downwards than upwards (because of the nature of the reporting scale), so it is not strange

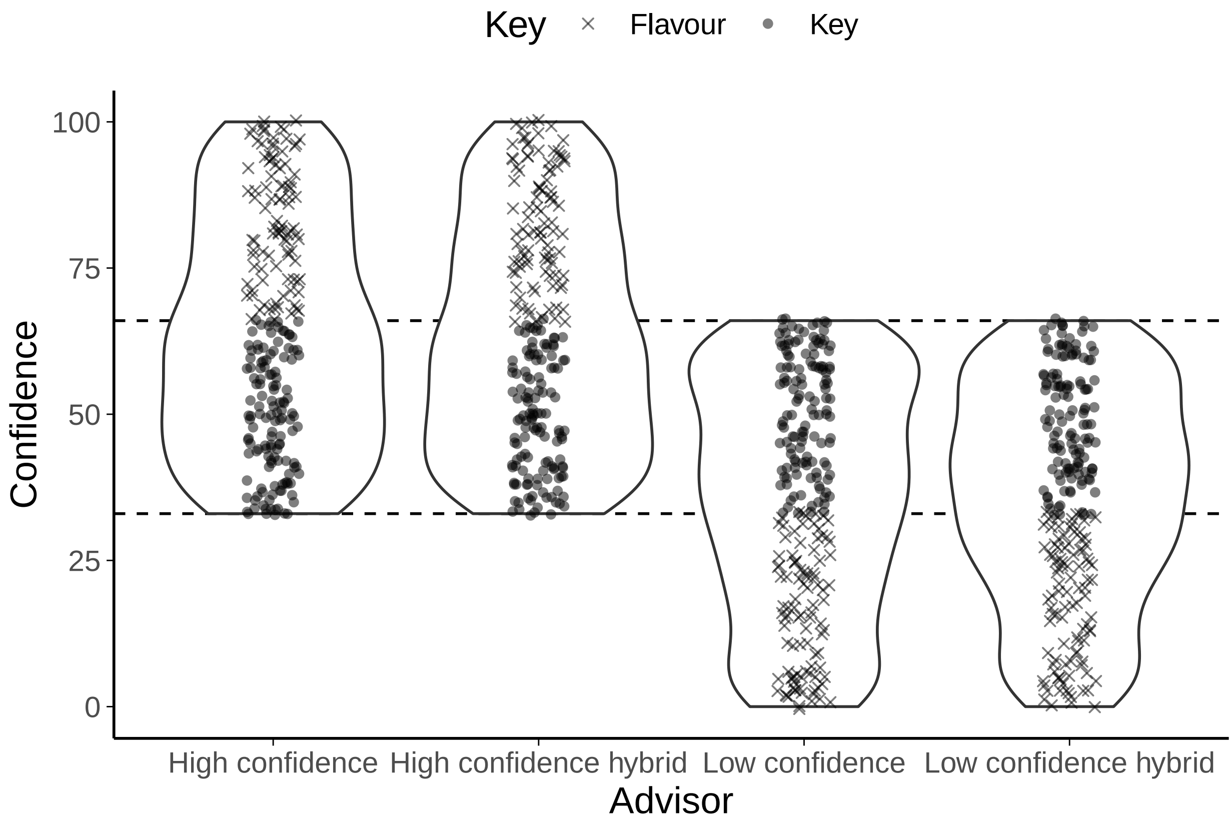


Figure 7.17: Key trials in Experiment 8.
Each point represents a single trial for a single participant. The outlines show the overall distribution of trials. Trials were matched for agreement and confidence across advisor presentation and advisor identity. The Flavour trials were included to avoid the feeling of the task changing dramatically from the Familiarisation blocks to the Test block. The Key trials occur where advice could potentially come from either advisor. If the advice came from the High confidence advisor it represented a low confidence response; it if came from the Low confidence advisor it represented a high confidence response. If the advisor came from an ambigous (hybrid) source, the participant would not know whether it represented a high or low confidence response.

to see that the average confidence scores are lower for the final decisions than for the initial estimates.

Effect of advisor presentation The main hypothesis in this experiment is that advice will be treated differently if it comes from an identifiable as opposed to an ambiguous source. We thus expect that the High confidence advisor's advice will be more influential when the source is ambiguous, because when participants know that the advice comes from the High confidence advisor, they should also know

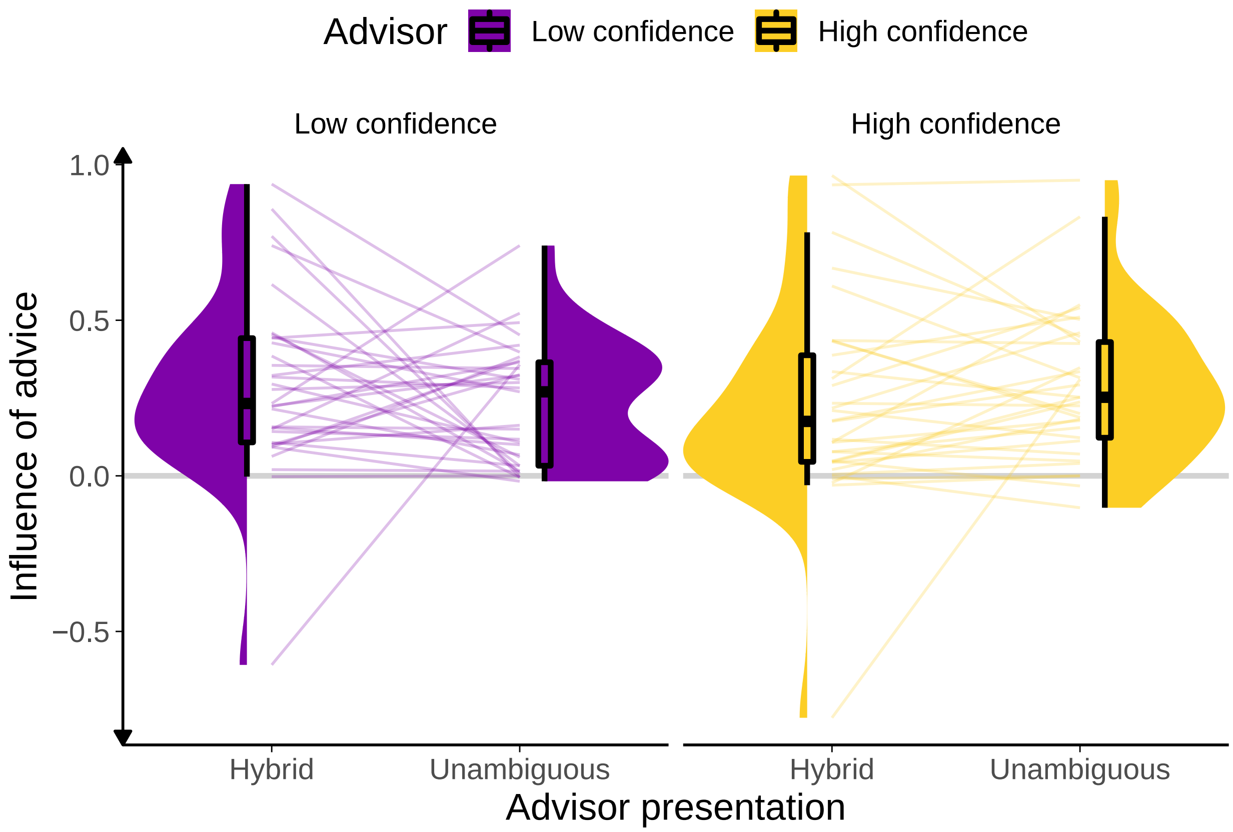


Figure 7.18: Advice influence in Experiment 8.
Participants' weight on the advice for advisors in the Test stage of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

that the confidence score represents that advisor's low confidence advice (which is less likely to be correct). The opposite prediction is made for the Low confidence advisor. These predictions are assessed on Key trials. Key trials occur in that band of confidence where the advice could have come from either advisor (Figure 7.17). Where advice is given with very high objective confidence it is not likely to come from the Low confidence advisor; likewise, where advice is given with very low objective confidence it is not likely to come from the High confidence advisor. It is only in that region where the advisors' use of the confidence scale overlaps that the Hybrid presentation is genuinely ambiguous.

Our predictions were not borne out in the data. A 2x2 ANOVA of influence scores with factors of Advisor type (High versus Low confidence) by presentation

(Unambiguous versus Hybrid) indicated no significant effects. Furthermore, the numerical direction of effects was against our prediction. As expected, there was no main effect to indicate that Hybrid presentation would be more or less influential than Unambiguous presentation (F(1,32) = 0.00, p = .961; $\mathrm{M}_{\mathrm{Unambiguous}}$ = 0.26 [0.19, 0.32], $\mathrm{M}_{\mathrm{Hybrid}}$ = 0.26 [0.16, 0.36]), and no main effect of advisor to indicate that, absent Unambiguous presentation, one advisor was more influential than the other (F(1,32) = 0.02, p = .896; $\mathrm{M}_{\mathrm{LowConfidence}}$ = 0.26 [0.20, 0.32], $\mathrm{M}_{\mathrm{HighConfidence}}$ = 0.26 [0.17, 0.34]). However, contrary to predictions, the interaction is not significant (F(1,32) = 3.91, p = .057; $\mathrm{M}_{\mathrm{Unambiguous\text{-}Hybrid|LowConfidence}}$ = -0.06 [-0.19, 0.06], $\mathrm{M}_{\mathrm{Unambiguous\text{-}Hybrid|HighConfidence}}$ = 0.06 [-0.04, 0.16]), and in fact goes in the opposite direction numerically to the one we would expect. The Unambiguous presentation of the Low confidence advisor is *less* influential, and the Unambiguous presentation of the High confidence advisor *more* influential than their Hybrid presentation counterparts.

Predictors of advice influence It is quite possible that participants failed to adequately grasp the calibration of the advisors, despite the lengths that we went to in order to highlight the importance of this feature. Indeed, a paired T-test of the influence of the advisors in Key trials where they were presented Unambiguously indicated that the High confidence advisor was numerically *more* influential than the Low confidence advisor ($t(32)$ = 1.33, p = .191, d = 0.26, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/2.39; $\mathrm{M}_{\mathrm{HighConfidence}}$ = 0.28 [0.20, 0.37], $\mathrm{M}_{\mathrm{LowConfidence}}$ = 0.23 [0.16, 0.30]). Statistically, there was not enough evidence to conclude whether or not the effect was present or absent.

This suggested that participants were simply adopting a heuristic that more confident advice was more likely to be correct – entirely ignoring the calibration of the advisors. Indeed, linear mixed modelling indicated that, when predicting the influence on Key trials from the advisor, presentation, advisor-by-presentation interaction, and advice confidence (with random intercepts for participant), only advice confidence was a significant predictor (β = 0.01 [0.01, 0.02], df = 509.64,

$p < .001$, $\mathrm{BF}_{\mathrm{H1:H0}} = 2.5\mathrm{e}4$). The other main predictors were both demonstrably absent ($\beta_{\text{High confidence advisor}}$ = -0.05 [-0.18, 0.07], df = 490.18, p = .390, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/5.12; $\beta_{\text{Unambiguous presentation}}$ = -0.08 [-0.20, 0.05], df = 490.11, p = .228, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/5.12), and there was not sufficient evidence to adjudicate on the presence of an interaction ($\beta_{\text{High confidence advisor, Unambiguous presentation}}$ = 0.13 [-0.04, 0.31], df = 490.14, p = .142, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 1/1.78).

7.3.2.4 Discussion

We attempted to manipulate participants' ability to use their knowledge of advisors' idiosyncratic confidence expression to perform our Dates estimation task. We were unable to demonstrate that participants did so, and it appeared from our linear modelling that participants paid more attention to the advice than its source. The predicted interaction between advisor confidence and advisor presentation (labelled versus hybrid) was not statistically significant, and indeed trended in the opposite direction to that hypothesised. It appeared from the numerical patterns, where participants were *more* influenced by moderate confidence advice if it was clearly labelled as coming from the High confidence advisor (despite therefore representing lower subjective confidence), and vice-versa for the Low confidence advisor, that participants had developed a mild positive association between the more confident advisor and more confident advice, treating that advisor as more trustworthy even when they expressed a moderate level of confidence.

There are several features of this experiment that may have led to the underwhelming conclusion. First, the advisors' gravatar images may be harder for participants to differentiate than human faces. Second, without offering an incentive beyond the innate appeal of the task, participants may not have been motivated to absorb the complex confidence calibration information presented to them, especially where the time cost to do so is not offset by performance-based payment. Third, the somewhat artificial nature of the task may limit participants' interest in the character of the advisors. Fourth, it may be quite difficult to overcome the widespread heuristic that higher confidence advice is more likely to be correct (Soll and Larrick 2009;

Moussaïd et al. 2013; Bang et al. 2014; Pulford et al. 2018; Price and Stone 2004). Adding additional information beyond the simple introductory speech bubble may have helped to alleviate this issue.

In short, we were either unable to get participants to attend to advisors' confidence calibration or people do not use confidence calibration information about advisors to modulate the influence of their advice. Whether or not additional appeal and character for the advisors would improve participants' sensitivity to the advisors' use of the confidence scale, these results cast doubt on a the practical applications of our model of advice-taking.

7.4 General discussion

Simulations in the previous chapter§6 suggested that egocentric discounting might arise as a rational response to context, and identified some contextual features that would make egocentric discounting rational. The experiments presented in this chapter tested this idea, based on the assumption that priming these reasons should lead to changes in egocentric discounting. Taken together, the results of these 4 experiments provided mixed support for this idea. The results of the direct benevolence manipulation in Experiments 5 and 6 show that people are sensitive to the motivations of their advisors, and exercise appropriate caution where those motivations may mean the advice is misleading. Experiment 5 also demonstrated 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 (Soll and Larrick 2009; Yaniv and Kleinberger 2000; Rakoczy et al. 2015; Schultze, Mojzisch, and Schulz-Hardt 2017), this supports the idea from the models that advice-taking can be flexibly 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 such as relying on cultural norms about the meanings of metacognitive terms (Roseano et al. 2016; Dhami and Wallsten 2005; but see Wallsten et al. 1986) are used instead of complex calculations. The lack of flexibility on the time-scale of a behavioural experiment does not, of course, negate the possibility of flexibility on the time-scale of a human interpersonal relationship (perhaps there are a certain number of people whose confidence mappings we can track, as suggested by Robin Dunbar's Social Brain Hypothesis (Dunbar 2010)[2]). Nor does it negate the possibility of a genetic or cultural evolutionary mechanism producing advice discounting as a protection against unknown confidence mapping, although it is questionable whether advice has been occurring in human societies for long enough to allow the former to take place. We have a plausible explanation of egocentric discounting that 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 hyper-priors held by participants that taking advice is risky for various reasons.

We saw in the first part of this thesis§3 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.

This work presents a picture of egocentric discounting that portrays it as a feature, not a bug. Overall, manipulations that affect the degree to which people take advice have the effects predicted by a normative theory: more advice is taken from advisors who are better at the task (e.g. Yaniv and Kleinberger 2000; Rakoczy et al. 2015; Sniezek, Schrah, and Dalal 2004; Soll and Mannes 2011); more confident advice is weighted more highly (Pulford et al. 2018; Soll and Larrick 2009; Moussaïd et al. 2013; Bang et al. 2014; Price and Stone 2004) (and people show a positive

[2] Note that this hypothesis has been viewed increasingly sceptically in recent years, as neatly covered by Lindenfors, Wartel, and Lind (2021)

association between confidence and accuracy (Pulford et al. 2018)); advisors with whom judges have stronger relationships are trusted more (van Swol and Sniezek 2005; Sniezek and van Swol 2001); and advisors with an incentive to mislead are trusted less (Bonner and Cadman 2014). With *changes* in advice-taking accounted for well by the normative model, only the base rate stands in need of explanation: the egocentric discounting phenomenon that people take less advice than they should. I explain this base rate reduction by appealing to the wider context of participants' personal, cultural, and evolutionary histories. Egocentric discounting is irrational only where the judge can be sure the advice will be, on average, as good as their own initial estimate. It is our contention that this assumption is simply not true in everyday life, and there is no reason why these background assumptions should be waived within the confines of a Judge-Advisor System, especially in the context of psychology experiments (which are notorious for deceiving participants).

The behavioural experiments extending the results of our evolutionary computational simulations were messier than we had hoped. The apparently obvious result that people take less advice from advisors who may sometimes mislead them proved quite difficult to demonstrate in practice – participants in early experiments did not appear to attend very much to the source of advice and it required a heavy-handed manipulation alongside splitting experience with the advisors into separate blocks to produce the effect seen in simulation. The difficulty we encountered would cast serious doubt on the amount of support we can derive from our experiments, except that the results proved robust once established: pre-registered replications showed very similar patterns in independent samples.

There is one major feature of advice-taking that has received relatively little investigation, yet that would, if real, be entirely inexplicable in terms of evolutionarily optimal hyper-priors. This is the finding that people tend to consider all advisory estimates they receive as a single, multifaceted opinion, which is then integrated with their own individual opinion§5.2.2.4. Crucially, this seems to happen no matter how many advisory estimates are supplied (Hütter and Ache 2016; Yonah and Kessler 2021). Some studies have found the opposite – rational weighting of advice according to the size of the group that produces it: Park et al. (2017) showed that a Bayesian model that took group size into account outperformed other models of advice-taking, although linear mixed modelling of their behavioural data did not show a group size effect except as an interaction with opinion differences; Martino et al. (2017) found that participants' ratings of on-line shopping products were influenced by other users' reviews, although they did not present evidence as to how closely social influence scaled with the number of reviewers. A rational approach to egocentric discounting might attempt, as we have above, to explain how it is sensible to afford less weight to another's opinion compared to one's own, but it cannot explain how it is sensible to afford the same weight to one's opinion whether integrating it with the opinions of one advisor or of several advisors. At best, we may be able to offer arguments that plead for different interpretations of advice from multiple estimates: many presentations may represent multiple estimates as a single estimate, and ones that do not may impose a cognitive load that leads to participants ignoring some information; or perhaps people sometimes interpret a mixture of answers as representing a general confusion and down-weight the group estimates as if they were advisory estimates given with very low confidence. Advice-taking is sensitive to particulars of task, presentation, and context, and the question of the extent to which multiple pieces of advice are weighted individually versus as a single piece of advice is also likely sensitive to those factors.

8 Conclusion

This thesis presented work on two questions: whether people take too much advice from those who agree with them; and why people do not take enough advice in general. Behavioural experiments using a Judge-Advisor System extended previous work (Pescetelli and Yeung 2021) on advice-taking to the domain of advisor choice, showing that people can use an advisor's agreement as a proxy for the advisor's accuracy where direct feedback about accuracy is not available. Computational simulations explored the implications of this tendency for the structure of social networks, replicating the echo chamber formation and polarisation found in similar simulations (Pescetelli and Yeung 2021; Madsen, Bailey, and Pilditch 2018) but suggesting that modelling heterogeneous agents reduced the speed and extent of these effects. The second arm of the thesis argued that egocentric discounting – taking advice less seriously than a normative model would predict – is not as irrational as it is often portrayed. Computation simulations showed that egocentric discounting was an adaptive trait in environments where advice was sometimes misleading, where advisors put less effort into making decisions than judges, and even where honest and diligent advice was communicated with a slightly different scaling of confidence. These simulations were complemented with a series of behavioural experiments investigating whether short-term changes in an advice-taking context would produce the discounting behaviour seen in the simulations. The results of these experiments were mixed: reducing the risk of deceptive advice increased advice-taking, but we found no evidence that familiarisation with an advisor's confidence calibration increased advice-taking.

The first of our questions is motivated by a concern that people are trapped in echo chambers – hearing only their own opinions parroted back to them and ignoring other perspectives. The argument is that people form social connections on the basis of similarities with others (homophily), and that their behaviour becomes increasingly similar to those they interact with. These age-old tendencies become pathological in the hyper-fluid social environment of on-line social networks: the pool of potential connections is much wider, meaning criteria for similarity go up; and the friction of making and breaking connections is much lower. Overall, as Sunstein (2018) persuasively argued, the on-line social world produces pockets of self-reinforcing opinion that trends to extremes because they exclude dissenting voices. In the last two decades there has been a great deal of discussion about the extent to which this picture is true (Garrett 2009a; Colleoni, Rozza, and Arvidsson 2014; Barberá 2015; Weeks, Ksiazek, and Holbert 2016; Sunstein 2002).

We found some evidence that the basic processes necessary for this to occur exist using behavioural experiments: people do seem to prefer to hear from advisors who are likely to support their initial opinion in the absence of reliable feedback, but this tendency did not show clearly when accurate and agreeing advisors were contrasted directly. Similar work using media sources rather than ostensibly human advisors has likewise shown mixed results: people seem to have a moderate tendency to choose sources likely to agree with their perspective when given a direct choice (Kobayashi and Ikeda 2009; Marquart 2016; Hart et al. 2009), but do not seem to deliberately avoid sources likely to disagree (Jang 2014; Hart et al. 2009). Viewing agreement as a positive attribute, while pathologised in this discourse, is not inherently misguided. We follow others (Soll and Larrick 2009) in highlighting that, for advisors and judges who perform better than chance, their answers overlap more the better they are at the task. Likewise, when consulting media sources, the framing placed on events and policies will ring more true for a person where it matches their own framing: a newspaper that talks about ‘riots’ will probably appear to be giving a more truthful account to a police officer, while one that speaks of ‘protests’ will probably appear more truthful to an activist.

Partisan media and homophilic social networks are long-standing concerns (Freedman 1965; Sears and Freedman 1967)[1], so a key question is whether, as argued by Sunstein (2018), on-line social media exacerbates the effects due to its presenting people with a multitude of potential social connections, activated by the click of a button. Our computational simulations presented agents with a similar frictionless environment – they were free to associate with which other agents they chose and incurred no cost to do so. These simulations reproduced the basic phenomenon of echo chamber formation seen in previous work (Pescetelli and Yeung 2021; Madsen, Bailey, and Pilditch 2018), but including the kind of heterogeneity in advice-taking and -seeking behaviour observed in our participant population worked to dilute the effect, slowing them down and making the divergence into two camps less complete.

There are further reasons to question whether experimental and simulation evidence adequately supports the idea of on-line social networks as engines of division and polarisation rather than tools for connecting the world. The interactions considered in this thesis and similar work are information transfers along social connections. These bare-bones representations are abstractions of much richer experiences that are conducted by specific individuals within particular contexts (discussed more below§8.1). Any number of aspects of these wider and richer experiences could mollify or reverse the simple effects of information. Secondly, on-line communities are not divorced from the real world – most have both on-line and off-line meetings, and most on-line communities are joined through pre-existing social connections (Hui and Buchegger 2009), with off-line connections being the dominant entry points (michelle boyd 2008). The aspects of on-line communities that are troublesome are similarly troublesome in off-line communities: the problematic echo chambers that make headlines, Q-Anon for example, share more features with

[1] The Pamphlet wars (*Pamphlet Wars* 2020) are an example from the 16th and 17th centuries.

other cults and conspiracy theorist networks from the off-line days than they do with the on-line social networks of the average contemporary netizen.[2] Despite the intuitive appeal of arguments and simulations showing echo chambers as a dominant social phenomenon, it remains highly plausible that echo chambers are nothing new, and that the opportunities granted by greater mixing on-line do not increase their prevalence or strength.

The second question is a puzzle that exemplifies the ongoing debate in psychology over whether people are rational[3] or whether their apparent rationality is an illusion created by a motley collection of heuristics. I certainly do not claim to have the last word in this debate, but this thesis does show that egocentric discounting, hitherto a good example of irrational cognition, may quite readily be seen as rational when the frame of reference is appropriately adjusted. Computational simulations showed that egocentric discounting is adaptive when there is the potential for advice to be deceptive, shiftless, or even slightly misunderstood. There is always this potential. Behavioural experiments added to a wealth of other experiments in the literature showing that advice-taking follows rational patterns once the base rate reduction (egocentric discounting) is explained. Our experiments on confidence calibration were unsuccessful in producing the results predicted by the simulations, though we cannot determine whether they were methodologically unsuccessful or whether they were unsuccessful because the simulated behaviour does not emerge in practice on a time-scale appropriate to behavioural experimental investigation.

The perspective of egocentric discounting as irrational is based on a normative model: the *discounting* is relative to an optimal level of advice weighting. I argue that the normative model is not so much incorrect as it is inappropriate. It requires that certain assumptions are met: there will be no lying; advisors will try as hard as the judge, and be roughly as good as the judge; the advice will be perfectly understood; and the advisor's expressions of confidence are interpreted as intended.

[2] *Netizen* is a neologism for an inter*net* cit*izen*.

[3] The debate is older than psychology as a discipline – Jonathan Swift for example wrote to Alexander Pope that humans should be defined not as "animal rationale" (a rational animal) but instead as "rationis capax" (capable of rationality) (Swift 1801, Dr. Swift to Mr. Pope, 29 September 1725).

If any of these are not met, the judge will be over-reliant on advice. I argue that these assumptions are never fully met.

The appeal of the normative model on one hand is matched on the other by the strength of the intuition that one should not equally weight advice with one's own opinion. This intuition may be culturally bound (indeed little English language research has been done on advice-taking outside Western Educated Industrialised Rich and Demographic populations (Henrich, Heine, and Norenzayan 2010) apart from Mahmoodi et al. (2015)), but it appears to be widely shared by both scientists and laypeople with whom I have discussed my research. This cultural norm advocating discounting exists alongside other norms that place strong emphasis on fairness and equality, including in integrating estimates to produce group decisions (Mahmoodi et al. 2015). There may be other cultural norms surrounding more proximal explanations of egocentric discounting that may also play a role, such as norms surrounding accountability for one's actions.

Overall, this work contributes to a picture of advice-taking and -seeking as remarkable quotidian. We highlight that there are reasonable motivations for trusting agreement over disagreement in the absence of feedback, and where tasks become more difficult the strength of this preference subsides. We illustrate how the heterogeneity that is sometimes lamented as the reason for social divisions appears to also prevent these divisions from wholly dichotomising opinions. We demonstrate that the guarded approach people show to advice-taking in general can be seen as rational even without invoking the presence of cheats, liars, and free-loaders.

8.1 Generalisability, limitations, and open questions

Despite the variety of methods and experimental tasks used in this work, it is necessarily limited in scope, and consequently there remains a deep concern about generalisability. Advice is the exchange of social information, and this thesis has focussed far more on the information than the social. Each instance of advice-taking is a meeting of particular people exchanging particular information with particular goals within a broader social context. Reading an advertisement is a very different situation from agonising with an old friend about their marital concerns, and both of these are very different from sharing a short message on a social media platform about planned building developments in your local area or reading expert reviews of food processors.

In many cases, advice comes not as a bald recommendation or a statement with mere confidence attached, but as a constructed argument with reasons and rhetoric. Indeed, reasons are more persuasive than recommendations with confidence where they are appropriate to the task at hand (Trouche, Sander, and Mercier 2014), as is often the case. Advice also serves several important social tasks: seeking advice can be a mark of respect and having advice taken an inherent pleasure (Hertz and Bahrami 2018); giving advice can be a form of dominance (See et al. 2011); and taking advice can be a way of diluting responsibility for decisions (Harvey and Fischer 1997).

These interactions are what build up the social relationships modelled in our simulations, and they do not change greatly in the on-line social arena. Like the off-line world, the on-line world is full of the same people pursuing the same goals; they may have slightly different opportunities and affordances, but they are in service of the same ends. The implications for real social networks of our simulations that are constructed from a narrow, information-centric view of advice-taking should be interpreted with caution – as should the simulations of others (Song and Boomgaarden 2017; Madsen, Bailey, and Pilditch 2018).

Our work is limited even within the information part of social information. We examined situations that present decision-makers with constrained choices concerning their advisors and variation only in the trust dimension of (perceived) ability. When real people make real choices about where to seek advice they are not constrained in this way, and their choice may be based on all three of the trust dimensions identified by Mayer, Davis, and Schoorman (1995). Furthermore, even within the dimension of ability, a great many different kinds of cues may be available beyond past personal experience. We may have been recommended to consult with someone by another person we trust, as happens when accountants acquire work through word-of-mouth. We might be consulting someone who has a reputation for expertise, like a rambler who has done a particular hike we want to try, or who belongs to a relevant professional organisation, like a counsellor who is a member of the British Association for Counselling and Psychotherapy. We may make inferences about expertise from how closely a potential advisor fits our stereotypes: do they dress appropriately or, more problematically, do they have the "right" accent, gender, or skin tone (Hollingshead and Fraidin 2003; Arnold, Crawford, and Khalifa 2016). How these features are combined with personal experience with an advisor is an important question because many of these aspects are constant, or change on different time-scales to the personal experience studied in this work, and thus may work to establish and maintain divisions that are difficult to break down using personal experience.

The method of comparison based on the many factors contributing to perceptions of ability is also likely to differ according to context. In our work, we gave participants in behavioural experiments pairs of advisors to choose between, and in simulation we extended this direct comparison process to a wider selection of potential advisors. In real choices of where to get advice, however, people are unlikely to dispassionately weigh up a variety of options and select the one that excels most according to their own criteria. Firstly, even in apparently simple multi-item choice problems, people's relative preferences for options varies according to the other options available (Becker and McClintock 1967). Secondly, unlike in our simulations and experiments, the opportunity cost to acquiring advice is usually low: people generally have the option to consult as many advisors as they have time to. Thirdly, the cost of consulting sources in terms of time, effort, or social risk is likely to be a factor that differs between sources. As with the relative importance of personal experience with an advisor to other factors, depending upon the extent of these considerations governing advisor choice, and depending upon their tendency to exacerbate or reduce homophily, the results of our computational simulations of network dynamics might be considerably different if they included these features.

Our work on egocentric discounting appears more robustly generalisable to the phenomenon of interest. Although the broader contextual considerations highlighted by our evolutionary simulations 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. The conclusion is therefore not that the normative model should be abandoned or even extended in general use, rather a recommendation that researchers avoid implying that deviation from the normative model is 'irrational', mistaken, or even the result of a 'satisficing' heuristic (Radner 1975).

Our theory of egocentric discounting's underlying rationality is not incontrovertible. In fact, the theory is poorly positioned to explain people's apparent tendency to treat multiple advisory estimates as a single instance of advice, and to discount it as if it represented only a single alternative opinion (Hütter and Ache 2016; Yonah and Kessler 2021), or at least to increase the weighting of it much less than would be expected as the number of contributing estimates increases (Park et al. 2017). The presentation of the advisory estimates may be an issue: presenting a single figure may understandably lead people to treat it as the product of a single advisory estimate; and requiring people to keep many suggestions in mind will likely overwhelm them and lead to cognitive shortcuts.

Another puzzle for our account is the hard-easy effect of advice: when given difficult tasks (such as our Dates task), people take more advice than when given easy tasks (Gino and Moore 2007). This is perhaps mediated by confidence. The normative model makes the prediction that what matters is the relative difficulty of the task for the judge and the advisor, not the absolute difficulty for the judge, and thus does not explain this effect. This puzzle is resolved by looking at the *perception* of relative difficulty, rather than the relative difficulty itself. People typically over-estimate their ability relative to others for easy tasks, while under-estimating their ability relative to others for difficult tasks (Kruger 1999). This behaviour may not be rational in the way that I have argued discounting is in general, but it may be an inescapable consequence of confidence (Hilbert 2012).

8.2 Key directions for future research

Behavioural experiments in this thesis provided some evidence to support the idea that people use agreement as a proxy for accuracy when they do not have access to feedback. This details of this picture require elucidation – for example it would be informative to know more precisely how task difficulty alters people's updating of trust in advisors. Extrapolating from the Pescetelli and Yeung (2021) theory of confidence-mediation, trust is updated more (in the absence of feedback) where the judge is most confident. For difficult decisions, the judge usually has very low confidence in their initial decisions and consequently we would expect advisor trust to update very little. Despite this, our experiments showed a very wide range of preference strengths and directions when participants were offered choices between advisors, suggesting that perhaps people did develop preferences (although they may also have had those preferences from the outset). It would be a useful extension of the Pescetelli and Yeung (2021) theory to explore how it interacts with very difficult decisions, especially because advice-taking is very high for difficult tasks (Gino and Moore 2007) and it is intuitive to think that the act of trusting (taking advice) should lead to greater trust in future.

The behavioural experiments used either binary decisions or continuous estimates as the task. In these tasks agreement can be readily quantified. Perhaps the majority of decisions for which advice is sought in the real world are not of this kind: they are much richer qualitative decisions with multiple options where advice is seldom a simple endorsement or specific estimate: “Where should I go on holiday?”; “Should I try watching this new drama programme?”; “What's the best way to get rust off a vintage cheese grater?”. Even the second of these, most like the tasks faced by participants in the behavioural experiments, tend not to have simple answers like “yes” or “no”, but qualified answers accompanied by non-verbal confidence cues: “I liked it, but you don’t like things that are too dark, right?”; “Definitely, but you have to watch until episode 3 to get into it.”; “I thought [some other programme] was better.”. It is important to investigate whether similar ‘agreement’ effects operate in these more complex interactions. It is plausible, for instance, that agreement is actually an instance of endorsement or positivity, and that these are responsible for increasing trust more than agreement per se.

This work replicated other models of polarisation and echo chamber formation (Pescetelli and Yeung 2021; Madsen, Bailey, and Pilditch 2018) and indicated that the heterogeneity seen in the participants of the behavioural experiments worked to slow the pace and limit the extent of those effects. Whether that heterogeneity needs to be a stable feature of agents deserves investigation, especially in light of the suggestion from our data that there was as much heterogeneity within participants as between them. Secondly, although these models make a strong theoretical case for polarisation and echo chamber formation in social networks, more work needs to be done to investigate whether these effects actually occur in practice. Analysis of data from on-line social networks has been equivocal on the subject (Colleoni, Rozza, and Arvidsson 2014; Barberá 2015; Weeks, Ksiazek, and Holbert 2016; Hui and Buchegger 2009), and this is an area where further research is needed.

This thesis presented a view of egocentric discounting that sees it as a rational response to ubiquitous issues in advice-taking: whether advice is well-intentioned; the quality of the advice; and the interpretability of advice. These issues are akin to generic issues in information exchange – how accurately information can be decoded from a signal and the utility of that information. Following this, a prediction from our view of egocentric discounting is that information exchange between any organisms, not just advice-giving humans, will show similar ‘discounting’ where the full utility of salient information cannot be used because it would lead to over-reliance on less useful or misunderstood information. More work needs to be done to properly demarcate the similarities and differences between advice-taking and other information exchange domains.

8.3 Summary

This work explored two questions: whether people take too much advice from those who agree with them; and why people do not take enough advice in general. To the first, I can say that people may well do this, but that the extent and consequences of their doing so are probably over-estimated. The natural variation in people's tendencies, and the wealth of other factors that contribute to the ways in which people interact when exchanging information, mean that simple but persuasive models of frictionless social network dynamics may not accurately capture relevant dynamics in real on-line social networks. To the second question, I answer that people *do* take enough advice… we just misrepresent how much 'enough' should be. When wider contexts are taken into account, the mystery fades, leaving the 'normative' model to act as a reference rather than an ideal.

Appendices

A
Appendix 1: Description of unanalysed data

A.0.0.1 Experiment 0

This branch of experiments was the core of the Dates task. As such, most experimentation and piloting happened on this branch, meaning that there were many versions of the study where data were collected for which analysis is not included here.

A.0.0.2 Experiment 1A

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 estimate 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.

A.0.0.3 Experiments 1B and 2B

Early versions included a bug which prevented feedback from being shown during the Familiarisation phase even to participants in the Feedback condition. 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.

A.0.0.4 Experiment 3A

Pilot data were collected to ensure the study functioned properly, and so the data are not analysed with regard to the hypotheses. The v1 Mixed design was conducted and run as a proper experiment in which participants learned about both advisors simultaneously (preregistered at https://osf.io/5z2fp), but there were no effects in the data.

A.0.0.5 Experiment 3B

The first version included a bug in which the advisor choice options were not recorded, making it difficult or impossible to work out which trials included a choice of advisor. Data for these participants could be included in a study that was agnostic about advisor choice.

A.0.0.6 Experiment 4A

The first version had a bug which meant that the advisors were identical in the Familiarisation phase. Both the first and second versions had a bug in which advisors who were supposed to agree with the participant's initial estimate gave the correct answer rather than agreeing. These participants' data could be included in an analysis which used a participant's actual experience of advice to predict their advice-taking and advisor choice behaviour, provided appropriate care was taken to reconstruct the advice data from the raw values instead of relying on the reported summaries.

A.0.0.7 Experiment 5

Several versions of this study were run in the course of developing the manipulation. The initial version had a bug that prevented the groups from being visually distinct. Later versions introduced a clearer manipulation, equivalent rather than genuinely misleading advice for the Sometimes misleading advisor, and rating of advice deceptiveness, respectively.

A.0.0.8 Experiment 6

Two short pilot versions of this experiment were run, and participants' data were not analysed due to bugs in the experiment code.

B

Appendix 2: Advice-taking in the Dates task

The advice-taking framework underlying this thesis§1.3 has some empirical support from previous work in our lab (Pescetelli and Yeung 2021). In those experiments, participants completed a perceptual decision-making task within a Judge-Advisor System. Participants provided a judgement (including a confidence rating) concerning which of two briefly presented boxes of dots contained more dots. They then received advice on the answer from an advisor and provided a final decision, again including a confidence judgement.

Pescetelli and Yeung (2021) showed that participants who received feedback on their decisions, which could in turn be used to evaluate the advisors, showed larger shifts in their confidence in the direction of advice from more accurate advisors. Participants who did not receive feedback showed similarly higher influence from advisors who tended to agree more often (i.e. whose advice indicated the same side that participants chose in their initial estimate). Both effects, accuracy and agreement, were present regardless of feedback, although the agreement effect was far more pronounced in the no feedback group than the feedback group, and the accuracy effect was more pronounced in the feedback than the no feedback group. Participants' subjective assessments of the trustworthiness of the advisors roughly[1] followed their behaviour: advisors who were more influential were rated as more trustworthy.

We aimed to replicate these results in a new paradigm. This new paradigm, the Dates task, was newly-implemented for this project. The Dates task was designed with several appealing features in mind: it was to be more similar to the tasks performed by participants in previous advice-taking experiments; it was to be shorter for participants to complete, and suitable for on-line delivery; and it was supposed to be more engaging for participants. We hoped this replication would accomplish two objectives: pilot our novel implementation and replicate the key advice-taking features reported by Pescetelli and Yeung (2021) using a different experimental task.

For this replication, we chose to directly contrast the tendency to prefer agreeing advisors over accurate ones where feedback was withheld with the tendency to prefer accurate advisors over agreeing ones where feedback was available.

B.1 Experiment 0: extending results to the Dates task

B.1.0.1 Open scholarship practices

This experiment was preregistered at `https://osf.io/fgmdw`. This is a replication of a study of identical design that produced the same results. The data for both this and the original study can be obtained from the `esmData` R package (Jaquiery 2021c). A snapshot of the state of the code for running the experiment at the time the experiment was run can be obtained from `https://github.com/oxacclab/ExploringSocialMetacognition/blob/f90b6f9266a901211a4ddb7b5ee1de1c74e8df57/ACv2/index.html`.

B.1.0.2 Method

37 participants each completed 42 trials over 3 blocks of the continuous version of the Dates task§2.1.3.2. On each trial, participants were presented with an historical event that occurred on a specific year between 1900 and 2000. They were asked to

[1] In some cases the results were only numerically compatible, and not statistically significant.

drag one of three markers onto a timeline to indicate the date range within which they thought the event occurred. The three markers each had different widths, and each marker had point value associated with it, with wider markers worth fewer points. The markers were 1, 3, and 9 years wide, being worth 27, 9, and 3 points respectively. Participants then received advice indicating a region of the timeline in which the advisor suggested the event occurred. This advice came from one of two advisors (as defined in detail below): one characterised by high objective accuracy and one characterised by a high degree of agreement with the participant's initial judgement. Participants could then mark a final response in the same manner as their original response, and could choose a different marker width if they wished.

Participants started with 1 block of 10 trials that contained no advice to allow them to familiarise themselves with the task. All trials in this section included feedback for all participants indicating whether or not the participant's response was correct.

This block ended with 2 trials with a practice advisor to get used to receiving advice. Participants also received feedback on these trials. They were informed that they would "get advice on the answers you give" and that the feedback they received would "tell you about how well the advisor does, as well as how well you do". Before starting the main experiment they were told that they would receive advice from multiple advisors and that "advisors might behave in different ways, and it's up to you to decide how useful you think each advisor is, and to use their advice accordingly".

Participants then performed 2 blocks of trials that constituted the main experiment. In each of these blocks participants had a single advisor for 14 trials, plus 1 attention check.

Participants were assigned randomly into one of two conditions that determined whether or not they received feedback during the main experiment. The first order in which advisors were encountered was counterbalanced between participants and across conditions. For each advisor, participants saw the advisor's advice on 14 trials. Advice was probabilistic: on each trial there was an 80% chance the advisor gave advice according to the advice profile (detailed below). Otherwise, the advisors issued the same kind of advice as one another, chosen to neither agree with the participant's answer nor indicate the correct answer. This "Off-brand" advice was used to control for the effects of advice when the influence of advice was the dependent variable. These trials are the key ones for analysis because they allow us to assess the influence of each advisor's advice, while matching the overall properties of that advice (i.e. objective accuracy and nearness to the participant's initial estimate).

Advice profiles The High accuracy and High agreement advisor profiles defined marker placements based on the timeline based on the correct answer and the participant's initial estimate respectively. Both advisors used markers that spanned 7 years, and both placed the markers in a normal distribution around the target point with a standard deviation of 5 years, but they differed in the central 'target point' around which this distribution was centred.. The target point for the High accuracy advisor was the correct answer, whereas the target point for the High agreement advisor was the participant's initial estimate. Neither advisor ever placed their marker exactly on the midpoint of the participant's marker (because doing so means the Weight on Advice statistic is undefined).

Each advice trial had a 20% chance of being designated "Off-brand". On these trials advisors neither indicated the correct answer nor agreed with the participant. This was achieved by picking a target point of the participant's answer reflected around the correct answer. Thus, if the centre of the participant's initial estimate was 1955, and the correct answer was 1945, the target point would be 1935. In some cases there was not enough of the scale left once this reflection had taken place. If the centre of a participant's initial estimate was 1960, and the correct answer was 1990, the target point could not be 2020, because the scale only went up to 2010. Where this occurred, advisors issued answers that were twice as wrong as the participant's answer, in the same direction. In this case, that would mean the new target point was 1930.

Table B.1: Participant exclusions for Dates task advice influence experiment

	Condition		
Reason	Feedback	No feedback	Total
Attention check	5	1	**6**
Unfinished	1	1	**2**
Too many outlying trials	0	1	**1**
Missing offbrand trial data	0	1	**1**
Total excluded	**5**	**3**	**8**
Total remaining	**19**	**18**	**37**

B.1.0.3 Results

Exclusions In line with the preregistration, participants' data were excluded from analysis where they failed attention checks, failed to complete the entire experiment, or had more than 2 outlying trials. Outlying trials were calculated after excluding participants who failed to complete the experiment, and were defined as trials for which the total trial time was greater than 3 standard deviations away from the mean of all trials from all participants. Table B.1 shows the number of participants excluded for each of the reasons, broken down by experimental condition.

A browser compatibility issue in this study meant that any participants completing the study using the Safari family of browsers had to be excluded because the advice was not presented appropriately.

Task performance Participants had lower error on final decisions than on their initial estimates (F(1,28) = 102.57, p < .001; $\text{M}_{\text{Initial}}$ = 16.45 [14.51, 18.39], M_{Final} = 11.26 [9.72, 12.81]), indicating improved performance after seeing advice. They also had lower error on their answers with the High accuracy advisor (main effect of advisor across both responses: F(1,28) = 44.75, p < .001; $\text{M}_{\text{HighAgreement}}$ = 16.89 [14.70, 19.09], $\text{M}_{\text{HighAccuracy}}$ = 10.82 [9.24, 12.41]). As expected, there was an interaction: participants reduced their error much more following advice from the High accuracy advisor (F(1,28) = 74.50, p < .001; $\text{M}_{\text{Reduction|HighAgreement}}$ = 1.06 [0.45, 1.67], $\text{M}_{\text{Reduction|HighAccuracy}}$ = 9.32 [7.38, 11.25]; Figure B.1).

Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Confidence can be measured

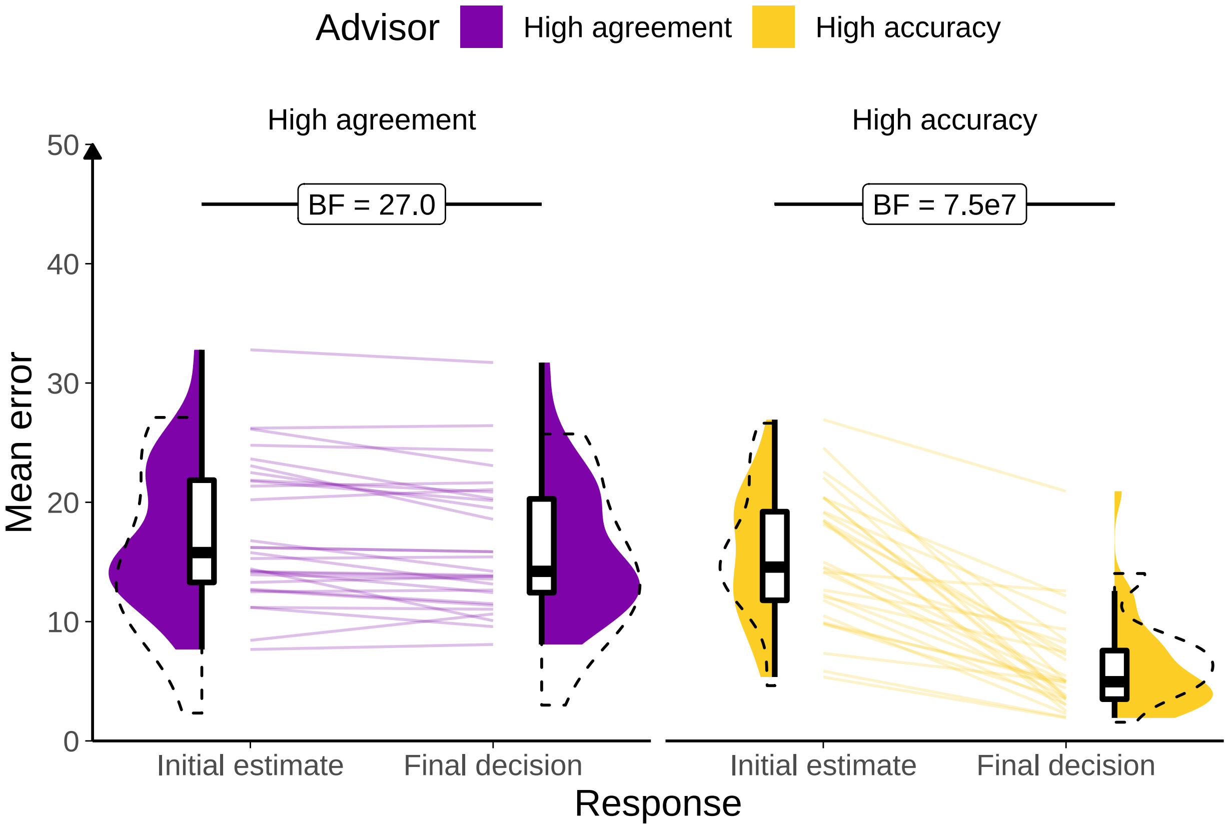


Figure B.1: Response error for Experiment 0.
Faint lines show individual participant mean error (the absolute difference between the participant's response and the correct answer), for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show data from the original study which this is a replication. The dependent variable is error, the distance between the correct answer and the participant's answer; lower values represent better performance. The theoretical limit for error is around 100.

by the width of the marker selected by the participant. Where participants are more confident in their response, they can maximise the points they receive by selecting a thinner marker. Where participants are unsure, they can maximise their chance of getting the answer correct by selecting a wider marker. Participants' error was lower for each marker width in final decisions than initial estimates (Figure B.2). For both initial estimates and final decisions, error was higher for wider markers than for narrower ones.

Advisor performance The advice is generated probabilistically so it is important to check that the advice experienced by the participants matched the experience

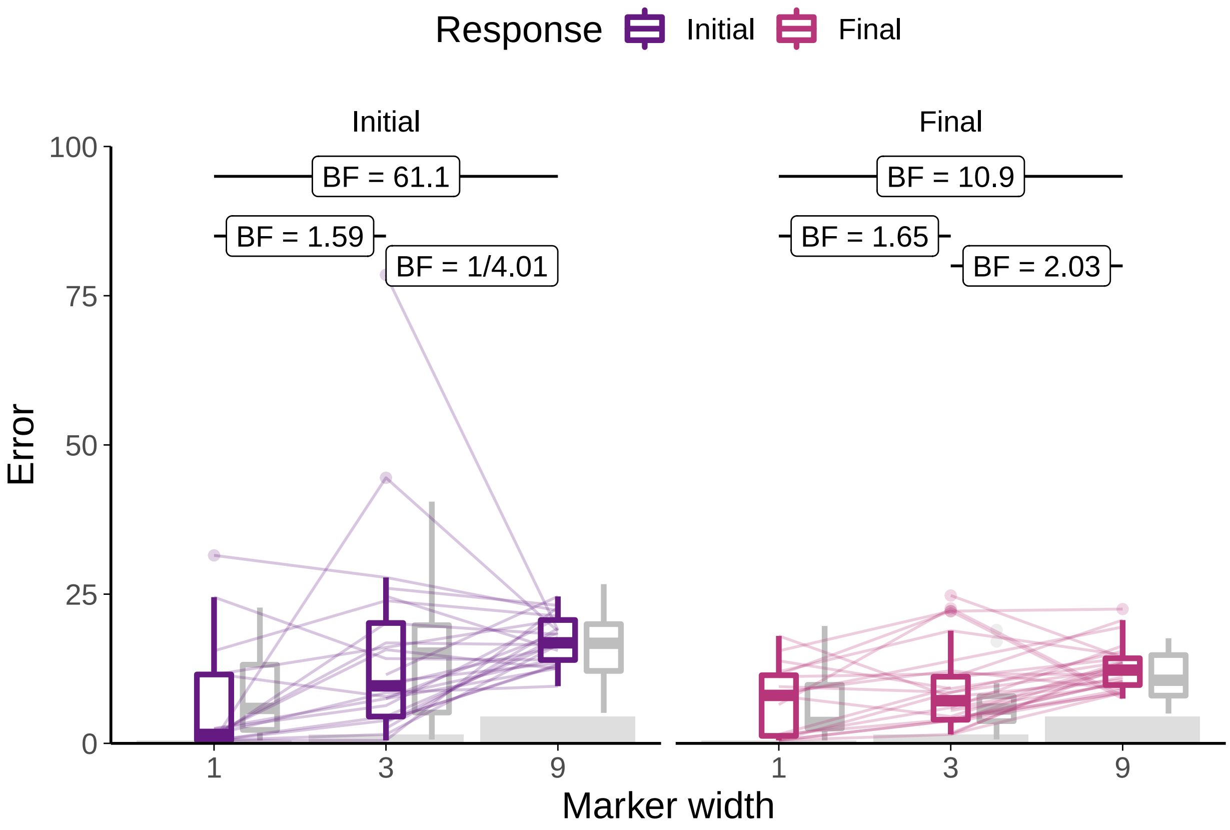


Figure B.2: Error by marker width for Experiment 0.
Faint lines show individual participant mean error (distance from the centre of the participant's marker to the correct answer) for each width of marker used, and box plots show the distributions. Some participants did not use all markers, and thus not all lines connect to each point on the horizontal axis. Grey box plots show data from the original experiment. The faint black points indicate outliers. Grey bars show half of the marker width: mean error scores within this range mean the marker covers the correct answer.

we designed. On average, the High accuracy advisor had lower error than the High agreement advisor ($t(28)$ = -12.95, p < .001, d = 3.37, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 3.0e10; $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 3.88 [3.49, 4.27], $\mathrm{M}_{\mathrm{HighAgreement}}$ = 16.99 [14.93, 19.05]), and their advice was further away from the participants' initial estimates than the High agreement advisor's ($t(28)$ = 9.66, p < .001, d = 2.00, $\mathrm{BF}_{\mathrm{H1:H0}}$ = 5.0e7; $\mathrm{M}_{\mathrm{HighAccuracy}}$ = 17.80 [15.58, 20.02], $\mathrm{M}_{\mathrm{HighAgreement}}$ = 7.77 [6.24, 9.29]). 29/29 (100.00%) participants experienced the High accuracy advisor as having lower average error than the High agreement advisor, and 27/29 (93.10%) participants experienced the High agreement advisor as offering advice closer to their initial estimates than the High accuracy advisor. Overall, this indicates that the manipulation was implemented as planned.

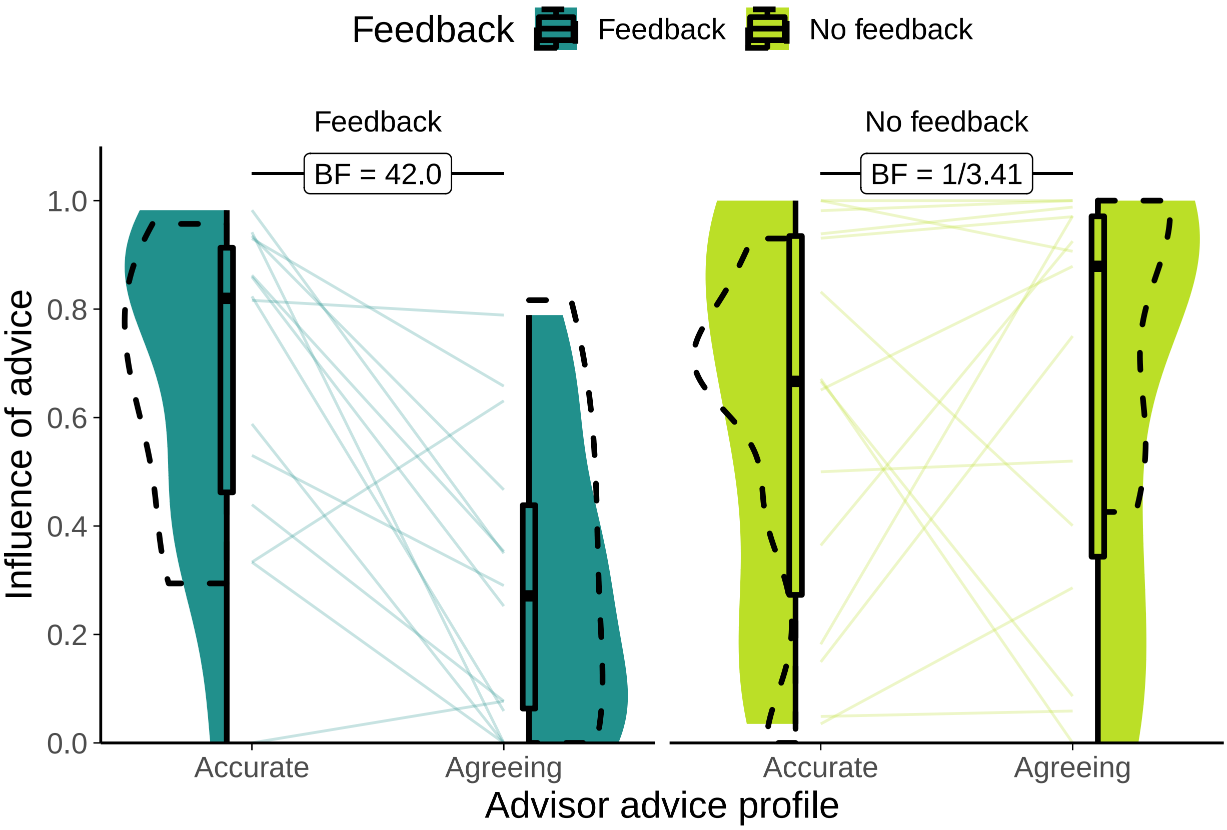


Figure B.3: Dates task advisor influence for High accuracy/agreement advisors. Shows the influence of the advice of the advisors. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The dashed outline shows the distribution of participant means in the original study of which this is a replication.

Hypothesis test There were systematic differences in the influence of advice on the Key trials where the advice itself was balanced between advisors (Figure B.3. A 2x2 mixed ANOVA of Advisor (within) by Feedback (between) indicated that the Accurate advisor was more influential than the Agreeing advisor (F(1,27) = 5.10, p = .032; $\mathrm{M}_{\mathrm{Accurate}}$ = 0.63 [0.51, 0.76], $\mathrm{M}_{\mathrm{Agreeing}}$ = 0.47 [0.33, 0.62]), and this was more extreme in the Feedback than the No feedback condition (F(1,27) = 9.72, p = .004; $\mathrm{M}_{\mathrm{Accurate\text{-}Agreeing|Feedback}}$ = 0.38 [0.19, 0.58], $\mathrm{M}_{\mathrm{Accurate\text{-}Agreeing|NoFeedback}}$ = -0.05 [-0.28, 0.17]). There was no significant difference between the Feedback and No feedback groups overall (F(1,27) = 2.02, p = .167; $\mathrm{M}_{\mathrm{Feedback}}$ = 0.48 [0.34, 0.61], $\mathrm{M}_{\mathrm{NoFeedback}}$ = 0.62 [0.45, 0.80]). As shown in the Bayesian statistics on the graph (Figure B.3), there was good evidence that the advisors were differently influential in the Feedback condition ($t(13) = 4.27$, $p < .001$, $d = 1.34$, $\mathrm{BF}_{\mathrm{H1:H0}} = 42.0$; $\mathrm{M}_{\mathrm{Accurate|Feedback}} = 0.67$ [0.50, 0.84], $\mathrm{M}_{\mathrm{Agreeing|Feedback}} = 0.29$ [0.13, 0.44]), and adequate evidence that they were not different in the No feedback condition ($t(14) = -0.50$, $p = .626$, $d = 0.14$, $\mathrm{BF}_{\mathrm{H1:H0}} = 1/3.41$; $\mathrm{M}_{\mathrm{Accurate|!Feedback}} = 0.60$ [0.40, 0.80], $\mathrm{M}_{\mathrm{Agreeing|!Feedback}} = 0.65$ [0.44, 0.86]).

B.1.1 Discussion

The purpose of this experiment was twofold: to pilot a novel implementation of a Judge-Advisor System paradigm; and to determine whether the key advice-taking effects identified by Pescetelli and Yeung (2021) are visible using a different experimental task. The first of these was successful, the second was more mixed.

We were able to provide advice to participants in an interface that they could understand, and to vary advice according to the key dimensions of accuracy and agreement (similarity) as desired. This was a pleasing result because the experiment was substantially shorter than our other approach, making it more enjoyable for participants and cheaper for us, and we had less precise control over participants' answers, making it harder to gauge how participants would experience the task. The behaviour of participants given feedback illustrated that they attended to the advice and they discriminated between advice that provided information and advice that provided support but no information.

The latter result provides a conceptual replication of the results found by Pescetelli and Yeung (2021), especially combined with the observation that where feedback is unavailable, people do not appear to distinguish between useful and supportive advice. Nevertheless, the results are not wholly consistent with an account of advice-taking in which people use agreement to evaluate advisors in the absence of feedback. Under such an account, we would expect participants in the No feedback condition to have shown a greater susceptibility to advice from the Agreeing advisor, but this did not happen.

One explanation for this discrepancy may be that it is possible that different people have differing preferences for agreeing versus non-redundant advice, particularly where the task is hard and participants are lacking information with which to judge their own and others' performance (as in the No Feedback condition). Another explanation for the equivalence of the influence of the advisors in the No feedback condition may be a consequence of relatively high levels of advice influence overall, producing a ceiling effect (the numerical advantage for the Agreeing advisor is as predicted by the theory). The relatively high levels of advice influence are a feature of the Dates task; the questions are difficult for most participants and consequently participants are likely to take more advice (Gino and Moore 2007; Yonah and Kessler 2021). A third possible explanation for the equivalence of influence between the advisors is that participants may not have had enough exposure to the advisors to properly learn about the value of their advice. The limitation on exposure is another feature of the Dates task: while the Dots task provides participants with many tens of trials in which to update their assessment of advisors, the Dates task provides a level of exposure more similar to normal social interaction (although not necessarily very similar).

Like its counterpart in the Dots task (Pescetelli and Yeung 2021), the design of this experiment 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. Nevertheless, the results show 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.

The results were only partially compatible with those of Pescetelli and Yeung (2021). In the Feedback condition, we saw, as expected, a much greater influence of advice from the Accurate advisor. In the No feedback condition, however, the influence of the advisors appeared to be equivalent. We did demonstrate that this paradigm is capable of measuring influence differences.

C

Appendix 3: Literate programming environment information

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.1252
[2] LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] parallel  stats     graphics  grDevices datasets  utils     methods
[8] base
```

```
other attached packages:
 [1] lmerTest_3.1-3        lme4_1.1-27.1         withr_2.4.2
 [4] igraph_1.2.6          adviseR_1.0.0         ez_4.4-0
 [7] patchwork_1.1.1       broom_0.7.9           ggtext_0.1.1
[10] ggpmisc_0.4.3         ggpp_0.4.2            ggridges_0.5.3
[13] see_0.6.7             magrittr_2.0.1        BayesFactor_0.9.12-4.2
[16] Matrix_1.3-4          coda_0.19-4           prettyMD_1.0.1
[19] esmData_1.0.0         kableExtra_1.3.4      glue_1.4.2
[22] forcats_0.5.1         stringr_1.4.0         dplyr_1.0.7
[25] purrr_0.3.4           readr_2.0.2           tidyr_1.1.4
[28] tibble_3.1.5          ggplot2_3.3.5         tidyverse_1.3.1
[31] knitr_1.36

loaded via a namespace (and not attached):
  [1] minqa_1.2.4         colorspace_2.0-2    ellipsis_0.3.2
  [4] rio_0.5.27          htmlTable_2.2.1     markdown_1.1
  [7] base64enc_0.1-3     fs_1.5.0            gridtext_0.1.4
 [10] rstudioapi_0.13     farver_2.1.0        MatrixModels_0.5-0
 [13] fansi_0.5.0         mvtnorm_1.1-2       lubridate_1.7.10
 [16] xml2_1.3.2          splines_4.1.0       Formula_1.2-4
 [19] jsonlite_1.7.2      nloptr_1.2.2.2      lsr_0.5.1
 [22] cluster_2.1.2       dbplyr_2.1.1        png_0.1-7
 [25] compiler_4.1.0      httr_1.4.2          backports_1.2.1
 [28] Ckmeans.1d.dp_4.3.3 assertthat_0.2.1    fastmap_1.1.0
 [31] cli_3.0.1           htmltools_0.5.2     quantreg_5.86
 [34] tools_4.1.0         gtable_0.3.0        reshape2_1.4.4
 [37] Rcpp_1.0.7          carData_3.0-4       cellranger_1.1.0
 [40] vctrs_0.3.8         svglite_2.0.0       nlme_3.1-153
 [43] conquer_1.0.2       insight_0.14.4      xfun_0.26
 [46] rbibutils_2.2.3     openxlsx_4.2.4      rvest_1.0.1
 [49] lifecycle_1.0.1   renv_0.14.0         gtools_3.9.2
 [52] MASS_7.3-54       scales_1.1.1        hms_1.1.1
 [55] SparseM_1.81      RColorBrewer_1.1-2  yaml_2.2.1
 [58] curl_4.3.2        gridExtra_2.3       pbapply_1.5-0
 [61] rpart_4.1-15      latticeExtra_0.6-29 stringi_1.7.4
 [64] highr_0.9         checkmate_2.0.0     boot_1.3-28
 [67] zip_2.2.0         Rdpack_2.1.2        rlang_0.4.11
 [70] pkgconfig_2.0.3   systemfonts_1.0.2   matrixStats_0.61.0
 [73] evaluate_0.14     lattice_0.20-45     htmlwidgets_1.5.4
 [76] labeling_0.4.2    tidyselect_1.1.1    plyr_1.8.6
 [79] bookdown_0.24     R6_2.5.1            Hmisc_4.5-0
 [82] generics_0.1.0    DBI_1.1.1           pillar_1.6.3
 [85] haven_2.4.3       foreign_0.8-81      mgcv_1.8-37
 [88] nnet_7.3-16       survival_3.2-13     abind_1.4-5
 [91] modelr_0.1.8      crayon_1.4.1        car_3.0-11
 [94] utf8_1.2.2        tzdb_0.1.2          rmarkdown_2.11
 [97] jpeg_0.1-9        grid_4.1.0          readxl_1.3.1
[100] data.table_1.14.2 reprex_2.0.1        digest_0.6.28
[103] webshot_0.5.2     numDeriv_2016.8-1.1 munsell_0.5.0
[106] viridisLite_0.4.0
```

Works Cited

20th Century (2012). In: *Oxford Reference*. HistoryWorld. URL: https://www.oxfordreference.com/view/10.1093/acref/9780191735639.timeline.0001 (visited on 08/13/2021).

Ais, Joaquín, Ariel Zylberberg, Pablo Barttfeld, and Mariano Sigman (2016). "Individual Consistency in the Accuracy and Distribution of Confidence Judgments". In: *Cognition* 146 (Supplement C), pp. 377–386. DOI: 10.1016/j.cognition.2015.10.006. URL: http://www.sciencedirect.com/science/article/pii/S0010027715300846 (visited on 10/18/2017).

Alexander, J. McKenzie (2021). "Evolutionary Game Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2021/entries/game-evolutionary/ (visited on 09/09/2021).

Arnold, Noelle Witherspoon, Emily R. Crawford, and Muhammad Khalifa (2016). "Psychological Heuristics and Faculty of Color: Racial Battle Fatigue and Tenure/Promotion". In: *The Journal of Higher Education* 87.6, pp. 890–919. DOI: 10.1353/jhe.2016.0033. URL: https://muse.jhu.edu/article/634189 (visited on 09/23/2021).

Azaria, Amos, Ya'akov Gal, Sarit Kraus, and Claudia V. Goldman (2016). "Strategic Advice Provision in Repeated Human-Agent Interactions". In: *Autonomous Agents and Multi-Agent Systems* 30.1, pp. 4–29. DOI: 10.1007/s10458-015-9284-6.

Bahrami, Bahador, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith (2010). "Optimally Interacting Minds". In: *Science* 329.5995, pp. 1081–1085. JSTOR: 40803021.

Bang, Dan, Riccardo Fusaroli, Kristian Tylén, Karsten Olsen, Peter E. Latham, Jennifer Y. F. Lau, Andreas Roepstorff, Geraint Rees, Chris D. Frith, and Bahador Bahrami (2014). "Does Interaction Matter? Testing Whether a Confidence Heuristic Can Replace Interaction in Collective Decision-Making". In: *Consciousness and Cognition* 26, pp. 13–23. DOI: 10.1016/j.concog.2014.02.002. URL: http://www.sciencedirect.com/science/article/pii/S1053810014000324 (visited on 01/05/2017).

Barberá, Pablo (2015). "How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S." In: *Job Market Paper, New York University*, p. 46.

Baron, Robert S., Sieg I. Hoppe, Chuan Feng Kao, Bethany Brunsman, Barbara Linneweh, and Diane Rogers (1996). "Social Corroboration and Opinion Extremity". In: *Journal of Experimental Social Psychology* 32.6, pp. 537–560. DOI: 10.1006/jesp.1996.0024. URL:

https://www.sciencedirect.com/science/article/pii/S0022103196900244 (visited on 09/16/2021).

Becker, G M and C G McClintock (1967). "Value: Behavioral Decision Theory". In: *Annual Review of Psychology* 18.1, pp. 239–286. DOI: 10.1146/annurev.ps.18.020167.001323. URL: http://www.annualreviews.org/doi/10.1146/annurev.ps.18.020167.001323 (visited on 09/23/2021).

Behrens, Timothy E. J., Laurence T. Hunt, Mark W. Woolrich, and Matthew F. S. Rushworth (2008). "Associative Learning of Social Value". In: *Nature* 456.7219, pp. 245–249. DOI: 10.1038/nature07538. URL: http://www.nature.com/articles/nature07538 (visited on 05/22/2019).

Boldt, Annika and Nicholas Yeung (2015). "Shared Neural Markers of Decision Confidence and Error Detection". In: *Journal of Neuroscience* 35.8, pp. 3478–3484. DOI: 10.1523/JNEUROSCI.0797-14.2015. pmid: 25716847. URL: https://www.jneurosci.org/content/35/8/3478 (visited on 09/25/2021).

Bonabeau, Eric (2002). "Agent-Based Modeling: Methods and Techniques for Simulating Human Systems". In: *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl 3), pp. 7280–7287. DOI: 10.1073/pnas.082080899. pmid: 12011407. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC128598/ (visited on 10/09/2017).

Bonner, Bryan L. and Brian D. Cadman (2014). "Group Judgment and Advice-Taking: The Social Context Underlying CEO Compensation Decisions". In: *Group Dynamics-Theory Research and Practice* 18.4, pp. 302–317. DOI: 10.1037/gdn0000011.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro (2017). "Greater Internet Use Is Not Associated with Faster Growth in Political Polarization among US Demographic Groups". In: *Proceedings of the National Academy of Sciences* 114.40, pp. 10612–10617. DOI: 10.1073/pnas.1706588114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1706588114 (visited on 09/10/2021).

Bramson, Aaron, Patrick Grim, Daniel J. Singer, Steven Fisher, William Berger, Graham Sack, and Carissa Flocken (2016). "Disambiguation of Social Polarization Concepts and Measures". In: *The Journal of Mathematical Sociology* 40.2, pp. 80–111. DOI: 10.1080/0022250X.2016.1147443. URL: https://doi.org/10.1080/0022250X.2016.1147443 (visited on 09/16/2021).

Brown, Jonathon D. (1986). "Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments". In: *Social Cognition* 4.4, pp. 353–376. DOI: http://dx.doi.org/10.1521/soco.1986.4.4.353. URL: https://www.proquest.com/docview/848858905/abstract/3BACA0DEE9424747PQ/1 (visited on 08/25/2021).

Bruch, Elizabeth and Jon Atwell (2015). "Agent-Based Models in Empirical Social Research". In: *Sociological Methods & Research* 44.2, pp. 186–221. DOI: 10.1177/0049124113506405. URL: http://journals.sagepub.com/doi/10.1177/0049124113506405 (visited on 10/10/2017).

Byrne, Kaileigh A., Thomas P. Tibbett, Lauren N. Laserna, Adrienne R. Carter-Sowell, and Darrell A. Worthy (2016). "Ostracism Reduces Reliance on Poor Advice from

Others during Decision Making". In: *Journal of Behavioral Decision Making* 29.4, pp. 409–418. DOI: 10.1002/bdm.1886.

Cardoso, Felipe Maciel, Sandro Meloni, Andre Santanche, and Yamir Moreno (2017). "Topical Homophily in Online Social Systems". In: URL: https://arxiv.org/abs/1707.06525 (visited on 11/07/2018).

Carlebach, Nomi and Nicholas Yeung (2020). "Subjective Confidence Acts as an Internal Cost-Benefit Factor When Choosing between Tasks." In: *Journal of Experimental Psychology: Human Perception and Performance* 46.7, pp. 729–748. DOI: 10.1037/xhp0000747. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000747 (visited on 09/25/2021).

Charles, Lucie and Nicholas Yeung (2019). "Dynamic Sources of Evidence Supporting Confidence Judgments and Error Detection." In: *Journal of Experimental Psychology: Human Perception and Performance* 45.1, pp. 39–52. DOI: 10.1037/xhp0000583. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000583 (visited on 09/25/2021).

Collaboration, Open Science (2015). "Estimating the Reproducibility of Psychological Science". In: *Science* 349.6251, aac4716. DOI: 10.1126/science.aac4716. pmid: 26315443. URL: http://science.sciencemag.org/content/349/6251/aac4716 (visited on 11/21/2017).

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (2014). "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter". In: *Journal of Communication* 64.2, pp. 317–332. DOI: 10.1111/jcom.12084. URL: http://doi.wiley.com/10.1111/jcom.12084 (visited on 01/05/2018).

Collins, Peter J., Ulrike Hahn, Ylva von Gerber, and Erik J. Olsson (2018). "The Bi-directional Relationship between Source Characteristics and Message Content". In: *Frontiers in Psychology* 9. DOI: 10.3389/fpsyg.2018.00018. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00018/full (visited on 09/02/2020).

Colombo, Matteo and Stephan Hartmann (2015). "Bayesian Cognitive Science, Unification, and Explanation". In: *The British Journal for the Philosophy of Science*, axv036. DOI: 10.1093/bjps/axv036. URL: http://bjps.oxfordjournals.org/lookup/doi/10.1093/bjps/axv036 (visited on 09/22/2020).

Criado Perez, Caroline (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. 1st ed. London: Random House.

Dana, Jason and Daylian M Cain (2015). "Advice versus Choice". In: *Current Opinion in Psychology* 6, pp. 173–176. DOI: 10.1016/j.copsyc.2015.08.019. URL: http://linkinghub.elsevier.com/retrieve/pii/S2352250X15002158 (visited on 08/22/2018).

Dandekar, Pranav, Ashish Goel, and David T. Lee (2013). "Biased Assimilation, Homophily, and the Dynamics of Polarization". In: *Proceedings of the National Academy of Sciences* 110.15, pp. 5791–5796. DOI: 10.1073/pnas.1217220110. pmid: 23536293. URL: http://www.pnas.org/content/110/15/5791 (visited on 11/13/2018).

Davies, Stephen (2017). "The Twin Impact of Homophily and Accessibility on Ideological Polarization". In: *Proceedings of the 2017 International Conference of The*

Computational Social Science Society of the Americas. CSS 2017: CSSSA's Annual Conference on Computational Social Science. Santa Fe NM USA: ACM, pp. 1–8. DOI: 10.1145/3145574.3145586. URL: https://dl.acm.org/doi/10.1145/3145574.3145586 (visited on 09/16/2021).

De Waal, F. B. M. (2014). *The Bonobo and the Atheist: In Search of Humanism among the Primates*. New York: Norton.

Dhami, Mandeep K. and Thomas S. Wallsten (2005). "Interpersonal Comparison of Subjective Probabilities: Toward Translating Linguistic Probabilities". In: *Memory & Cognition* 33.6, pp. 1057–1068. DOI: 10.3758/BF03193213. URL: http://link.springer.com/10.3758/BF03193213 (visited on 08/25/2021).

Diamond, Jared (1998). *Guns, Germs & Steel*. 3rd ed. London: Vintage.

Dietvorst, Berkeley J. and Uri Simonsohn (2019). "Intentionally "Biased": People Purposely Use to-Be-Ignored Information, but Can Be Persuaded Not To." In: *Journal of Experimental Psychology: General* 148.7, pp. 1228–1238. DOI: 10.1037/xge0000541. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000541 (visited on 11/04/2020).

Duggins, Peter (2017). "A Psychologically-Motivated Model of Opinion Change with Applications to American Politics". In: *Journal of Artificial Societies and Social Simulation* 20.1, p. 13. DOI: 10.18564/jasss.3316. arXiv: 1406.7770. URL: http://arxiv.org/abs/1406.7770 (visited on 07/16/2021).

Dunbar, Robin (2010). *How Many Friends Does One Person Need?: Dunbar's Number and Other Evolutionary Quirks*. Faber & Faber. 310 pp. Google Books: gQ_MFDc_F4kC.

East, Robert, Mark D. Uncles, Jenni Romaniuk, and Wendy Lomax (2016). "Improving Agent-Based Models of Diffusion". In: *European Journal of Marketing; Bradford* 50.3/4, pp. 639–646. URL: https://search.proquest.com/docview/1823080285/abstract/6000096DD4C0431DPQ/1 (visited on 10/10/2017).

Ernst, Marc O. and Martin S. Banks (2002). "Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion". In: *Nature* 415.6870 (6870), pp. 429–433. DOI: 10.1038/415429a. URL: https://www.nature.com/articles/415429a (visited on 09/22/2020).

FeldmanHall, Oriel and Joseph E. Dunsmoor (2019). "Viewing Adaptive Social Choice Through the Lens of Associative Learning". In: *Perspectives on Psychological Science* 14.2, pp. 175–196. DOI: 10.1177/1745691618792261. URL: http://journals.sagepub.com/doi/10.1177/1745691618792261 (visited on 08/14/2019).

Festinger, Leon (1957). *A Theory of Cognitive Dissonance*. Evanston, Ill: Row, Peterson. 291 pp.

Fetsch, Christopher R., Alexandre Pouget, Gregory C. DeAngelis, and Dora E. Angelaki (2012). "Neural Correlates of Reliability-Based Cue Weighting during Multisensory Integration". In: *Nature Neuroscience* 15.1, pp. 146–154. DOI: 10.1038/nn.2983. URL: http://www.nature.com.ezproxy.sussex.ac.uk/neuro/journal/v15/n1/abs/nn.2983.html (visited on 01/05/2017).

Fleming, Stephen M. and Hakwan C. Lau (2014). "How to Measure Metacognition". In: *Frontiers in Human Neuroscience* 8. DOI: 10.3389/fnhum.2014.00443. pmid:

25076880. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4097944/ (visited on 10/31/2017).

Fleming, Stephen M., Jihye Ryu, John G. Golfinos, and Karen E. Blackmon (2014). "Domain-Specific Impairment in Metacognitive Accuracy Following Anterior Prefrontal Lesions". In: *Brain* 137.10, pp. 2811–2822. DOI: 10.1093/brain/awu221. URL: https://academic.oup.com/brain/article/137/10/2811/2847667/Domain-specific-impairment-in-metacognitive (visited on 08/22/2017).

Freedman, Jonathan L. (1965). "Confidence, Utility, and Selective Exposure: A Partial Replication." In: *Journal of Personality and Social Psychology* 2.5, pp. 778–780. DOI: 10.1037/h0022670. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0022670 (visited on 07/27/2021).

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

Gao, Wenjun, Lin Qiu, Chi-yue Chiu, and Yiyin Yang (2015). "Diffusion of Opinions in a Complex Culture System: Implications for Emergence of Descriptive Norms". In: *Journal of Cross-Cultural Psychology* 46.10, pp. 1252–1259. DOI: 10.1177/0022022115610212. URL: https://doi.org/10.1177/0022022115610212 (visited on 10/10/2017).

Garrett, R. Kelly (2009a). "Echo Chambers Online?: Politically Motivated Selective Exposure among Internet News Users". In: *Journal of Computer-Mediated Communication* 14.2, pp. 265–285. DOI: 10.1111/j.1083-6101.2009.01440.x. URL: http://doi.wiley.com/10.1111/j.1083-6101.2009.01440.x (visited on 01/05/2018).

— (2009b). "Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate". In: *Journal of Communication* 59.4, pp. 676–699. DOI: 10.1111/j.1460-2466.2009.01452.x. URL: http://doi.wiley.com/10.1111/j.1460-2466.2009.01452.x (visited on 01/05/2018).

Gino, Francesca (2008). "Do We Listen to Advice Just Because We Paid for It? The Impact of Advice Cost on Its Use". In: *Organizational Behavior and Human Decision Processes* 107.2, pp. 234–245. DOI: 10.1016/j.obhdp.2008.03.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0749597808000435 (visited on 01/09/2018).

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". In: *Journal of Personality and Social Psychology* 102.3, pp. 497–512.

Gino, Francesca and Don A. Moore (2007). "Effects of Task Difficulty on Use of Advice". In: *Journal of Behavioral Decision Making* 20.1, pp. 21–35. DOI: 10.1002/bdm.539. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.539 (visited on 05/15/2018).

Grönlund, Kimmo, Kaisa Herne, and Maija Setälä (2015). "Does Enclave Deliberation Polarize Opinions?" In: *Political Behavior* 37.4, pp. 995–1020. DOI: 10.1007/s11109-015-9304-x. URL: https://doi.org/10.1007/s11109-015-9304-x (visited on 09/16/2021).

Haddara, Nadia and Dobromir Rahnev (2020). *The Impact of Feedback on Perceptual Decision Making and Metacognition: Reduction in Bias but No Change in Sensitivity.*

preprint. PsyArXiv. DOI: 10.31234/osf.io/p8zyw. URL: https://osf.io/p8zyw (visited on 06/03/2020).

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill (2009). "Feeling Validated versus Being Correct: A Meta-Analysis of Selective Exposure to Information." In: *Psychological Bulletin* 135.4, pp. 555–588. DOI: 10.1037/a0015701. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0015701 (visited on 07/27/2021).

Harvey, Nigel and Ilan Fischer (1997). "Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility". In: *Organizational Behavior and Human Decision Processes* 70.2, pp. 117–133. DOI: 10.1006/obhd.1997.2697. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749597897926972 (visited on 08/04/2020).

Harvey, Nigel and Clare Harries (2004). "Effects of Judges' Forecasting on Their Later Combination of Forecasts for the Same Outcomes". In: *International Journal of Forecasting* 20.3, pp. 391–409. DOI: 10.1016/j.ijforecast.2003.09.012. URL: http://www.sciencedirect.com/science/article/pii/S0169207003001109 (visited on 08/04/2020).

Hauser, Tobias U., Michael Moutoussis, Peter Dayan, and Raymond J. Dolan (2017). "Increased Decision Thresholds Trigger Extended Information Gathering across the Compulsivity Spectrum". In: *Translational Psychiatry* 7.12 (12), pp. 1–10. DOI: 10.1038/s41398-017-0040-3. URL: https://www.nature.com/articles/s41398-017-0040-3 (visited on 09/06/2021).

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). "The Weirdest People in the World?" In: *Behavioral and Brain Sciences* 33.2-3, pp. 61–83. DOI: 10.1017/S0140525X0999152X. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/weirdest-people-in-the-world/BF84F7517D56AFF7B7EB58411A554C17 (visited on 12/03/2020).

Hertz, Uri and Bahador Bahrami (2018). "Intrinsic Value of Social Influence over Others". In: DOI: 10.31234/osf.io/6jm7t. URL: https://psyarxiv.com/6jm7t/ (visited on 10/10/2018).

Hertz, Uri, Stefano Palminteri, Silvia Brunetti, Cecilie Olesen, Chris D. Frith, and Bahador Bahrami (2017). "Neural Computations Underpinning the Strategic Management of Influence in Advice Giving". In: *Nature Communications* 8.1, p. 2191. DOI: 10.1038/s41467-017-02314-5. URL: https://www.nature.com/articles/s41467-017-02314-5 (visited on 03/09/2018).

Heyes, Cecilia, Dan Bang, Nicholas Shea, Chris D. Frith, and Stephen M. Fleming (2020). "Knowing Ourselves Together: The Cultural Origins of Metacognition". In: *Trends in Cognitive Sciences* 24.5, pp. 349–362. DOI: 10.1016/j.tics.2020.02.007. URL: http://www.sciencedirect.com/science/article/pii/S1364661320300590 (visited on 09/07/2020).

Hilbert, Martin (2012). "Toward a Synthesis of Cognitive Biases: How Noisy Information Processing Can Bias Human Decision Making." In: *Psychological Bulletin* 138.2, pp. 211–237. DOI: 10.1037/a0025940. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0025940 (visited on 09/24/2021).

Hollingshead, Andrea B. and Samuel N. Fraidin (2003). “Gender Stereotypes and Assumptions about Expertise in Transactive Memory”. In: *Journal of Experimental Social Psychology* 39.4, pp. 355–363. DOI: 10.1016/S0022-1031(02)00549-8. URL: https://www.sciencedirect.com/science/article/pii/S0022103102005498 (visited on 09/23/2021).

Hui, Pan and Sonja Buchegger (2009). “Groupthink and Peer Pressure: Social Influence in Online Social Network Groups”. In: *2009 International Conference on Advances in Social Network Analysis and Mining*. 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 53–59. DOI: 10.1109/ASONAM.2009.17.

Humphries, Mark D. and Kevin Gurney (2008). “Network ‘Small-World-Ness’: A Quantitative Method for Determining Canonical Network Equivalence”. In: *PLOS ONE* 3.4, e0002051. DOI: 10.1371/journal.pone.0002051. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002051 (visited on 07/26/2021).

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

Jackson, Joshua Conrad, David Rand, Kevin Lewis, Michael I. Norton, and Kurt Gray (2017). “Agent-Based Modeling: A Guide for Social Psychologists”. In: *Social Psychological and Personality Science* 8.4, pp. 387–395. DOI: 10.1177/1948550617691100. URL: https://doi.org/10.1177/1948550617691100 (visited on 10/10/2017).

Jacowitz, Karen E. and Daniel Kahneman (1995). “Measures of Anchoring in Estimation Tasks:” in: *Personality and Social Psychology Bulletin*. DOI: 10.1177/01461672952111004. URL: https://journals.sagepub.com/doi/10.1177/01461672952111004 (visited on 08/04/2020).

Jang, S. Mo (2014). “Challenges to Selective Exposure: Selective Seeking and Avoidance in a Multitasking Media Environment”. In: *Mass Communication and Society* 17.5, pp. 665–688. DOI: 10.1080/15205436.2013.835425. URL: https://doi.org/10.1080/15205436.2013.835425 (visited on 07/27/2021).

Jaquiery, Matt (2021a). *Exploring Social Metacognition Simulation Data*. Zenodo. DOI: 10.5281/zenodo.5543918. URL: https://zenodo.org/record/5543918 (visited on 10/01/2021).

— (2021b). *Oxacclab/adviseR: Thesis*. Zenodo. DOI: 10.5281/zenodo.5543799. URL: https://zenodo.org/record/5543799 (visited on 10/01/2021).

— (2021c). *Oxacclab/esmData: Thesis*. Zenodo. DOI: 10.5281/zenodo.5543803. URL: https://zenodo.org/record/5543803 (visited on 10/01/2021).

— (2021d). *Oxacclab/EvoEgoBias: Thesis*. Zenodo. DOI: 10.5281/zenodo.5543811. URL: https://zenodo.org/record/5543811 (visited on 10/01/2021).

Jones, Gregory Todd (2007). “Agent-Based Modelling: Use with Necessary Caution”. In: *American Journal of Public Health* 97.5, pp. 780–781. DOI: 10.2105/AJPH.2006.109058. pmid: 17395853. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1854870/ (visited on 10/10/2017).

Kerr, Norbert L. (1998). “HARKing: Hypothesizing After the Results Are Known”. In: *Personality and Social Psychology Review* 2.3, pp. 196–217. DOI: 10.1207/s15327957pspr0203_4. URL:

http://journals.sagepub.com/doi/10.1207/s15327957pspr0203_4 (visited on 04/21/2020).

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

Knobloch-Westerwick, Silvia (2015). “The Selective Exposure Self- and Affect-Management (SESAM) Model: Applications in the Realms of Race, Politics, and Health”. In: *Communication Research* 42.7, pp. 959–985. DOI: 10.1177/0093650214539173. URL: http://journals.sagepub.com/doi/10.1177/0093650214539173 (visited on 07/27/2021).

Kobayashi, Tetsuro and Ken’ichi Ikeda (2009). “Selective Exposure in Political Web Browsing: Empirical Verification of ‘Cyber-Balkanization’ in Japan and the USA”. In: *Information, Communication & Society* 12.6, pp. 929–953. DOI: 10.1080/13691180802158490. URL: http://www.tandfonline.com/doi/abs/10.1080/13691180802158490 (visited on 01/05/2018).

Körding, Konrad P., Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams (2007). “Causal Inference in Multisensory Perception”. In: *PLOS ONE* 2.9, e943. DOI: 10.1371/journal.pone.0000943. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000943 (visited on 09/22/2020).

Kruger, Justin (1999). “Lake Wobegon Be Gone! The "Below-Average Effect" and the Egocentric Nature of Comparative Ability Judgments”. In: *Journal of personality and social psychology* 77.2, pp. 221–232. DOI: 10.1037/0022-3514.77.2.221.

Lawrence, Barbara S. and Neha Parikh Shah (2020). “Homophily: Measures and Meaning”. In: *Academy of Management Annals* 14.2, pp. 513–597. DOI: 10.5465/annals.2018.0147. URL: https://journals.aom.org/doi/full/10.5465/annals.2018.0147 (visited on 09/15/2021).

Liberman, Varda, Julia A. Minson, Christopher J. Bryan, and Lee Ross (2012). “Naïve Realism and Capturing the “Wisdom of Dyads””. In: *Journal of Experimental Social Psychology* 48.2, pp. 507–512. DOI: 10.1016/j.jesp.2011.10.016. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022103111002599 (visited on 05/15/2018).

Lindenfors, Patrik, Andreas Wartel, and Johan Lind (2021). “‘Dunbar’s Number’ Deconstructed”. In: *Biology Letters* 17.5, p. 20210158. DOI: 10.1098/rsbl.2021.0158. URL: https://royalsocietypublishing.org/doi/full/10.1098/rsbl.2021.0158 (visited on 08/25/2021).

Lockwood, Patricia L. and Miriam C. Klein-Flügge (2020). “Computational Modelling of Social Cognition and Behaviour—a Reinforcement Learning Primer”. In: *Social Cognitive and Affective Neuroscience*. DOI: 10.1093/scan/nsaa040. URL: https://academic.oup.com/scan/advance-article/doi/10.1093/scan/nsaa040/5813717 (visited on 05/20/2020).

Lord, Charles G., Lee Ross, and Mark R. Lepper (1979). “Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence”. In: *Journal of Personality and Social Psychology* 37.11, pp. 2098–2109. DOI: 10.1037/0022-3514.37.11.2098.

Lyngs, Ulrik (2019). *Oxforddown: An Oxford University Thesis Template for r Markdown.* DOI: 10.5281/zenodo.3484682. URL: https://github.com/ulyngs/oxforddown.

MacLeod, A. and S. Pietravalle (2017). "Communicating Risk: Variability of Interpreting Qualitative Terms". In: *EPPO Bulletin* 47.1, pp. 57–68. DOI: 10.1111/epp.12367. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/epp.12367 (visited on 09/09/2021).

Madsen, Jens Koed, Richard M. Bailey, and Toby D. Pilditch (2018). "Large Networks of Rational Agents Form Persistent Echo Chambers". In: *Scientific Reports* 8.1 (1), p. 12391. DOI: 10.1038/s41598-018-25558-7. URL: https://www.nature.com/articles/s41598-018-25558-7 (visited on 03/10/2021).

Mahmoodi, Ali, Dan Bang, Karsten Olsen, Yuanyuan Aimee Zhao, Zhenhao Shi, Kristina Broberg, Shervin Safavi, Shihui Han, Majid Nili Ahmadabadi, Chris D. Frith, Andreas Roepstorff, Geraint Rees, and Bahador Bahrami (2015). "Equality Bias Impairs Collective Decision-Making across Cultures". In: *Proceedings of the National Academy of Sciences* 112.12, pp. 3835–3840. DOI: 10.1073/pnas.1421692112. pmid: 25775532. URL: http://www.pnas.org/content/112/12/3835 (visited on 10/18/2017).

Marquart, Franziska (2016). "Selective Exposure in the Context of Political Advertising: A Behavioral Approach Using Eye-Tracking Methodology". In: p. 20.

Martino, Benedetto De, Sebastian Bobadilla-Suarez, Takao Nouguchi, Tali Sharot, and Bradley C. Love (2017). "Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability". In: *Journal of Neuroscience* 37.25, pp. 6066–6074. DOI: 10.1523/JNEUROSCI.3880-16.2017. pmid: 28566360. URL: https://www.jneurosci.org/content/37/25/6066 (visited on 09/22/2021).

MATLAB (2017). The Mathworks, Inc. URL: https://uk.mathworks.com/products/matlab.html.

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.

McClain, William (2016). "A Pathway Forwards for the Social Capital Metaphor". In: *Review of Social Economy* 74.2, pp. 109–128. DOI: 10.1080/00346764.2015.1089106. URL: http://dx.doi.org/10.1080/00346764.2015.1089106 (visited on 10/10/2017).

McCormick, Meghan P., Elise Cappella, Diane L. Hughes, and Emily K. Gallagher (2015). "Feasible, Rigorous, and Relevant: Validation of a Measure of Friendship Homophily for Diverse Classrooms". In: *The Journal of Early Adolescence* 35.5-6, pp. 817–851. DOI: 10.1177/0272431614547051. URL: http://journals.sagepub.com/doi/10.1177/0272431614547051 (visited on 09/15/2021).

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. DOI: 10.1146/annurev.soc.27.1.415. URL: https://doi.org/10.1146/annurev.soc.27.1.415 (visited on 11/07/2018).

Mezic, Igor, Paul J. Gruenewald, Dennis M. Gorman, and Jadranka Mezic (2007). "Mezic et al. Respond". In: *American Journal of Public Health* 97.5, pp. 781–782. DOI: 10.2105/AJPH.2007.109710. pmid: null. URL:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1854862/ (visited on 10/10/2017).

Michelle boyd, danah (2008). "Taken out of Context: American Teen Sociality in Networked Publics". PhD thesis. United States – California: University of California, Berkeley. 395 pp. URL: https://www.proquest.com/docview/304697239/abstract/3EFB697FB09F4A52PQ/1 (visited on 09/23/2021).

Minson, Julia A., Varda Liberman, and Lee Ross (2011). "Two to Tango: Effects of Collaboration and Disagreement on Dyadic Judgment". In: *Personality and Social Psychology Bulletin* 37.10, pp. 1325–1338. DOI: 10.1177/0146167211410436. URL: http://journals.sagepub.com/doi/10.1177/0146167211410436 (visited on 07/22/2020).

Minson, Julia A. and Jennifer S. Mueller (2012). "The Cost of Collaboration: Why Joint Decision Making Exacerbates Rejection of Outside Information". In: *Psychological Science* 23.3, pp. 219–224. DOI: 10.1177/0956797611429132. URL: http://journals.sagepub.com/doi/10.1177/0956797611429132 (visited on 05/15/2018).

Moorman, Robert H., Gerald L. Blakely, and Todd C. Darnold (2018). "Understanding How Perceived Leader Integrity Affects Follower Trust: Lessons From the Use of Multidimensional Measures of Integrity and Trust". In: *Journal of Leadership & Organizational Studies* 25.3, pp. 277–289. DOI: 10.1177/1548051817750544. URL: http://journals.sagepub.com/doi/10.1177/1548051817750544 (visited on 09/15/2021).

Morey, Richard D. and Jeffrey N. Rouder (2015). *BayesFactor: Computation of Bayes Factors for Common Designs*. URL: https://CRAN.R-project.org/package=BayesFactor.

Moses-Payne, Madeleine E., Max Rollwage, Stephen M. Fleming, and Jonathan P. Roiser (2019). "Postdecision Evidence Integration and Depressive Symptoms". In: *Frontiers in Psychiatry* 10. DOI: 10.3389/fpsyt.2019.00639. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00639/full (visited on 01/15/2020).

Moussaïd, Mehdi, Juliane E. Kämmer, Pantelis P. Analytis, and Hansjörg Neth (2013). "Social Influence and the Collective Dynamics of Opinion Formation". In: *PLOS ONE* 8.11, e78433. DOI: 10.1371/journal.pone.0078433. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078433 (visited on 07/22/2020).

Navajas, Joaquin, Chandni Hindocha, Hebah Foda, Mehdi Keramati, Peter E. Latham, and Bahador Bahrami (2017). "The Idiosyncratic Nature of Confidence". In: *Nature Human Behaviour* 1.11, p. 810. DOI: 10.1038/s41562-017-0215-1. URL: https://www.nature.com/articles/s41562-017-0215-1 (visited on 11/27/2018).

Nelson, Jacob L. and James G. Webster (2017). "The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience". In: *Social Media + Society* 3.3, p. 205630511772931. DOI: 10.1177/2056305117729314. URL: http://journals.sagepub.com/doi/10.1177/2056305117729314 (visited on 04/13/2018).

Nosek, Brian A. and Yoav Bar-Anan (2012). "Scientific Utopia: I. Opening Scientific Communication". In: *Psychological Inquiry* 23.3, pp. 217–243. DOI: 10/gcsk27. URL: https://doi.org/10.1080/1047840X.2012.692215 (visited on 03/03/2020).

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl (2012). "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability". In: *Perspectives on Psychological Science* 7.6, pp. 615–631. DOI: 10/f4fc2k. URL: https://doi.org/10.1177/1745691612459058 (visited on 03/03/2020).

Nowak, Andrzej, Jacek Szamrej, and Bibb Latané (1990). "From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact." In: *Psychological Review* 97.3, p. 362. URL: http://psycnet.apa.org/journals/rev/97/3/362/ (visited on 10/11/2017).

Önkal, Dilek, Sinan M. Gönül, Paul Goodwin, Mary Thomson, and Esra Öz (2017). "Evaluating Expert Advice in Forecasting: Users' Reactions to Presumed vs. Experienced Credibility". In: *International Journal of Forecasting* 33.1, pp. 280–297. DOI: 10.1016/j.ijforecast.2015.12.009. URL: http://www.sciencedirect.com/science/article/pii/S0169207016300061 (visited on 07/21/2020).

Önkal, Dilek, Paul Goodwin, Mary Thomson, Sinan M. Gönül, and Andrew Pollock (2009). "The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments". In: *Journal of Behavioral Decision Making* 22.4, pp. 390–409. DOI: 10.1002/bdm.637.

Palanski, Michael E. and Francis J. Yammarino (2007). "Integrity and Leadership:: Clearing the Conceptual Confusion". In: *European Management Journal* 25.3, pp. 171–184. DOI: 10.1016/j.emj.2007.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0263237307000400 (visited on 09/15/2021).

Pamphlet Wars (2020). In: *Wikipedia*. URL: https://en.wikipedia.org/w/index.php?title=Pamphlet_wars&oldid=974602198 (visited on 09/24/2021).

Park, Barum (2018). "How Are We Apart? Continuity and Change in the Structure of Ideological Disagreement in the American Public, 1980–2012". In: *Social Forces* 96.4, pp. 1757–1784. DOI: 10.1093/sf/sox093. URL: https://academic.oup.com/sf/article/96/4/1757/4781058 (visited on 08/31/2021).

Park, Seongmin A., Sidney Goïame, David A. O'Connor, and Jean-Claude Dreher (2017). "Integration of Individual and Social Information for Decision-Making in Groups of Different Sizes". In: *PLOS Biology* 15.6, e2001958. DOI: 10.1371/journal.pbio.2001958. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2001958 (visited on 09/22/2021).

Perrett, Stuart (2021). "A Divided Kingdom? Variation in Polarization, Sorting, and Dimensional Alignment among the British Public, 1986–2018". In: *The British Journal of Sociology* n/a.n/a. DOI: 10.1111/1468-4446.12873. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-4446.12873 (visited on 08/31/2021).

Pescetelli, Niccolò, Anna-Katharina Hauperich, and Nicholas Yeung (2021). "Confidence, Advice Seeking and Changes of Mind in Decision Making". In: *Cognition* 215, p. 104810. DOI: 10.1016/j.cognition.2021.104810. URL: https://www.sciencedirect.com/science/article/pii/S0010027721002298 (visited on 07/29/2021).

Pescetelli, Niccolò and Nicholas Yeung (2021). "The Role of Decision Confidence in Advice-Taking and Trust Formation." In: *Journal of Experimental Psychology: General* 150.3, pp. 507–526. DOI: 10.1037/xge0000960. arXiv: 1809.10453. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000960 (visited on 07/20/2021).

Pinter, Brad and Anthony G. Greenwald (2011). "A Comparison of Minimal Group Induction Procedures". In: *Group Processes & Intergroup Relations* 14.1, pp. 81–98. DOI: 10.1177/1368430210375251. URL: http://journals.sagepub.com/doi/10.1177/1368430210375251 (visited on 03/20/2019).

Price, Paul C. and Eric R. Stone (2004). "Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic". In: *Journal of Behavioral Decision Making* 17.1, pp. 39–57. DOI: 10.1002/bdm.460. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.460 (visited on 09/22/2021).

Pulford, Briony D., Andrew M. Colman, Eike K. Buabang, and Eva M. Krockow (2018). "The Persuasive Power of Knowledge: Testing the Confidence Heuristic." In: *Journal of Experimental Psychology: General* 147.10, p. 1431. DOI: 10.1037/xge0000471. URL: https://psycnet.apa.org/fulltext/2018-41338-001.pdf (visited on 09/22/2021).

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Rabbie, Jacob M. and Murray Horwitz (1969). "Arousal of Ingroup-Outgroup Bias by a Chance Win or Loss". In: *Journal of Personality and Social Psychology* 13.3, pp. 269–277. DOI: 10.1037/h0028284.

Rader, Christina A., Richard P. Larrick, and Jack B. Soll (2017). "Advice as a Form of Social Influence: Informational Motives and the Consequences for Accuracy". In: *Social and Personality Psychology Compass* 11.8, n/a–n/a. DOI: 10.1111/spc3.12329. URL: http://onlinelibrary.wiley.com/doi/10.1111/spc3.12329/abstract (visited on 01/12/2018).

Radner, Roy (1975). "Satisficing". In: *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*. Ed. by G. I. Marchuk. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 252–263. DOI: 10.1007/978-3-662-38527-2_34. URL: https://doi.org/10.1007/978-3-662-38527-2_34 (visited on 09/23/2021).

Rakoczy, Hannes, Christoph Ehrling, Paul L. Harris, and Thomas Schultze (2015). "Young Children Heed Advice Selectively". In: *Journal of Experimental Child Psychology* 138, pp. 71–87. DOI: 10.1016/j.jecp.2015.04.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S002209651500106X (visited on 02/01/2019).

Renault, Jérôme, Eilon Solan, and Nicolas Vieille (2013). "Dynamic Sender–Receiver Games". In: *Journal of Economic Theory* 148.2, pp. 502–534. DOI: 10.1016/j.jet.2012.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0022053113000161 (visited on 09/10/2021).

Richards, Andrew (2015). "University of Oxford Advanced Research Computing". In: DOI: 10.5281/zenodo.22558. URL: https://zenodo.org/record/22558#.W79A1PbTVhE (visited on 10/11/2018).

Ronayne, David and Daniel Sgroi (2018). "When Good Advice Is Ignored: The Role of Envy and Stubbornness". In: WERP 1150. DOI: 10.22004/ag.econ.269082.

Roseano, Paolo, Montserrat González, Joan Borràs-Comes, and Pilar Prieto (2016). "Communicating Epistemic Stance: How Speech and Gesture Patterns Reflect Epistemicity and Evidentiality". In: *Discourse Processes* 53.3, pp. 135–174. DOI: 10.1080/0163853X.2014.969137. URL: http://dx.doi.org/10.1080/0163853X.2014.969137 (visited on 05/16/2017).

Rouault, Marion, Peter Dayan, and Stephen M. Fleming (2019). "Forming Global Estimates of Self-Performance from Local Confidence". In: *Nature Communications* 10.1, p. 1141. DOI: 10.1038/s41467-019-09075-3. URL: https://www.nature.com/articles/s41467-019-09075-3 (visited on 06/16/2019).

Rouault, Marion, Tricia Seow, Claire M. Gillan, and Stephen M. Fleming (2018). "Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance". In: *Biological Psychiatry*. Translating Biology to Treatment in Schizophrenia 84.6, pp. 443–451. DOI: 10.1016/j.biopsych.2017.12.017. URL: http://www.sciencedirect.com/science/article/pii/S0006322318300295 (visited on 06/16/2019).

Sah, Sunita, Don A. Moore, and Robert J. MacCoun (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. DOI: 10.1016/j.obhdp.2013.02.001. URL: http://www.sciencedirect.com/science/article/pii/S074959781300023X (visited on 05/09/2017).

Schkade, David, Cass R. Sunstein, and Reid Hastie (2010). "When Deliberation Produces Extremism". In: *Critical Review* 22.2-3, pp. 227–252. DOI: 10.1080/08913811.2010.508634. URL: http://www.tandfonline.com/doi/abs/10.1080/08913811.2010.508634 (visited on 09/16/2021).

Schmuck, Desiree, Miriam Tribastone, Joerg Matthes, Franziska Marquart, and Eva Maria Bergel (2020). "Avoiding the Other Side? An Eye-Tracking Study of Selective Exposure and Selective Avoidance Effects in Response to Political Advertising." In: *Journal Of Media Psychology-Theories Methods And Applications* 32.3, pp. 158–164. DOI: 10.1027/1864-1105/a000265. URL: https://lirias.kuleuven.be/retrieve/603662 (visited on 07/27/2021).

Schönbrodt, Felix D. and Eric-Jan Wagenmakers (2018). "Bayes Factor Design Analysis: Planning for Compelling Evidence". In: *Psychonomic Bulletin & Review* 25.1, pp. 128–142. DOI: 10.3758/s13423-017-1230-y. URL: https://doi.org/10.3758/s13423-017-1230-y (visited on 10/01/2021).

Schul, Yaacov and Noam Peni (2015). "Influences of Distrust (and Trust) on Decision Making". In: *Social Cognition* 33.5, pp. 414–435. DOI: 10.1521/soco.2015.33.5.414.

Schultze, Thomas, Andreas Mojzisch, and Stefan Schulz-Hardt (2017). "On the Inability to Ignore Useless Advice: A Case for Anchoring in the Judge-Advisor-System". In:

Experimental Psychology 64.3, pp. 170–183. DOI: 10.1027/1618-3169/a000361. URL: http://econtent.hogrefe.com/doi/10.1027/1618-3169/a000361 (visited on 01/09/2018).

Schultze, Thomas, Anne-Fernandine Rakotoarisoa, and Stefan Schulz-Hardt (2015). "Effects of Distance between Initial Estimates and Advice on Advice Utilization". In: *Judgment and Decision Making* 10.2, p. 28.

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., Elizabeth W. Morrison, Naomi B. Rothman, and Jack B. Soll (2011). "The Detrimental Effects of Power on Confidence, Advice Taking, and Accuracy". In: *Organizational Behavior and Human Decision Processes* 116.2, pp. 272–285. DOI: 10.1016/j.obhdp.2011.07.006. URL: http://www.sciencedirect.com/science/article/pii/S0749597811000975 (visited on 10/18/2017).

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11, pp. 1359–1366. DOI: 10.1177/0956797611417632. URL: http://journals.sagepub.com/doi/10.1177/0956797611417632 (visited on 11/21/2017).

Smith, Eliot R. and Frederica R. Conrey (2007). "Agent-Based Modeling: A New Approach for Theory Building in Social Psychology". In: *Personality and Social Psychology Review* 11.1, pp. 87–104. DOI: 10.1177/1088868306294789. URL: https://doi.org/10.1177/1088868306294789 (visited on 10/10/2017).

Sniezek, Janet A., Gunnar E. Schrah, and Reeshad S. Dalal (2004). "Improving Judgement with Prepaid Expert Advice". In: *Journal of Behavioral Decision Making* 17.3, pp. 173–190. DOI: 10.1002/bdm.468. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.468 (visited on 05/15/2018).

Sniezek, Janet A. and Lyn M. van Swol (2001). "Trust, Confidence, and Expertise in a Judge-Advisor System". In: *Organizational Behavior and Human Decision Processes* 84.2, pp. 288–307. DOI: 10.1006/obhd.2000.2926. URL: http://www.sciencedirect.com/science/article/pii/S0749597800929261 (visited on 10/18/2017).

Soll, Jack B. and Richard P. Larrick (2009). "Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.3, pp. 780–805. DOI: 10.1037/a0015145. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0015145 (visited on 05/17/2018).

Soll, Jack B. and Albert E. Mannes (2011). "Judgmental Aggregation Strategies Depend on Whether the Self Is Involved". In: *International Journal of Forecasting* 27.1, pp. 81–102. DOI: 10.1016/j.ijforecast.2010.05.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169207010000877 (visited on 05/15/2018).

Song, Chen, Ryota Kanai, Stephen M. Fleming, Rimona S. Weil, D. Samuel Schwarzkopf, and Geraint Rees (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. DOI: 10.1016/j.concog.2010.12.011. URL: http://www.sciencedirect.com/science/article/pii/S1053810010002709 (visited on 08/22/2017).

Song, Hyunjin and Hajo G. Boomgaarden (2017). “Dynamic Spirals Put to Test: An Agent-Based Model of Reinforcing Spirals Between Selective Exposure, Interpersonal Networks, and Attitude Polarization”. In: *Journal of Communication* 67.2, pp. 256–281. DOI: 10.1111/jcom.12288. URL: http://onlinelibrary.wiley.com/doi/10.1111/jcom.12288/abstract (visited on 10/10/2017).

Steinhauser, Marco and Nicholas Yeung (2010). “Decision Processes in Human Performance Monitoring”. In: *Journal of Neuroscience* 30.46, pp. 15643–15653. DOI: 10.1523/JNEUROSCI.1899-10.2010. pmid: 21084620. URL: https://www.jneurosci.org/content/30/46/15643 (visited on 09/25/2021).

Sunstein, Cass R. (2002). *Republic.Com*. Princeton University Press. 252 pp. Google Books: 07AG9TxDJdgC.

— (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. 332 pp. Google Books: nVBLDwAAQBAJ.

Svenson, Ola (1981). “Are We All Less Risky and More Skillful than Our Fellow Drivers?” In: *Acta Psychologica* 47.2, pp. 143–148. DOI: 10.1016/0001-6918(81)90005-6. URL: https://linkinghub.elsevier.com/retrieve/pii/0001691881900056 (visited on 08/04/2020).

Swift, Jonathan (1801). “Letters Between Jonathan Swift and Alexander Pope – Complete”. In: *The Works of the Rev. Jonathan Swift, Volume 14*. Ed. by John Nichols.

Timeline of the 19th Century (2021). In: *Wikipedia*. URL: https://en.wikipedia.org/w/index.php?title=Timeline_of_the_19th_century&oldid=1029982440 (visited on 08/13/2021).

Timeline of the 20th Century (2021). In: *Wikipedia*. URL: https://en.wikipedia.org/w/index.php?title=Timeline_of_the_20th_century&oldid=1037994474 (visited on 08/13/2021).

Tost, Leigh Plunkett, Francesca Gino, and Richard P. Larrick (2012). “Power, Competitiveness, and Advice Taking: Why the Powerful Don’t Listen”. In: *Organizational Behavior and Human Decision Processes* 117.1, pp. 53–65. DOI: 10.1016/j.obhdp.2011.10.001.

Tragedy of the Commons (2021). In: *Wikipedia*. URL: https://en.wikipedia.org/w/index.php?title=Tragedy_of_the_commons&oldid=1046798003 (visited on 10/01/2021).

Trouche, Emmanuel, Petter Johansson, Lars Hall, and Hugo Mercier (2018). “Vigilant Conservatism in Evaluating Communicated Information”. In: *PLOS ONE* 13.1. Ed. by Alexander N. Sokolov, e0188825. DOI: 10.1371/journal.pone.0188825. URL: http://dx.plos.org/10.1371/journal.pone.0188825 (visited on 05/15/2018).

Trouche, Emmanuel, Emmanuel Sander, and Hugo Mercier (2014). “Arguments, More than Confidence, Explain the Good Performance of Reasoning Groups.” In: *Journal of Experimental Psychology: General* 143.5, pp. 1958–1971. DOI: 10.1037/a0037099. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0037099 (visited on 09/22/2021).

Van Overwalle, Frank and Francis Heylighen (2006). “Talking Nets: A Multiagent Connectionist Approach to Communication and Trust between Individuals.” In:

Psychological Review 113.3, pp. 606–627. DOI: 10.1037/0033-295X.113.3.606. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.113.3.606 (visited on 10/11/2017).

Van Swol, Lyn M. (2011). "Forecasting Another's Enjoyment versus Giving the Right Answer: Trust, Shared Values, Task Effects, and Confidence in Improving the Acceptance of Advice". In: *International Journal of Forecasting* 27.1, pp. 103–120. DOI: 10.1016/j.ijforecast.2010.03.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169207010000415 (visited on 10/15/2018).

Van Swol, Lyn M. and Janet A. Sniezek (2005). "Factors Affecting the Acceptance of Expert Advice". In: *British Journal of Social Psychology* 44.3, pp. 443–461. DOI: 10.1348/014466604X17092. URL: https://onlinelibrary.wiley.com/doi/abs/10.1348/014466604X17092 (visited on 07/27/2020).

Wallsten, Thomas S, David V Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth (1986). "Measuring the Vague Meanings of Probability Terms". In: *Journal of Experimental Psychology: General* 115, pp. 348–365.

Wang, Xiuxin and Xiufang Du (2018). "Why Does Advice Discounting Occur? The Combined Roles of Confidence and Trust". In: *Frontiers in Psychology* 9, p. 2381. DOI: 10.3389/fpsyg.2018.02381.

Watts, Duncan and Steven Strogatz (1998). "Collective Dynamics of Small-World Networks". In: 393, p. 3.

Weeks, Brian E., Thomas B. Ksiazek, and R. Lance Holbert (2016). "Partisan Enclaves or Shared Media Experiences? A Network Approach to Understanding Citizens' Political News Environments". In: *Journal of Broadcasting & Electronic Media* 60.2, pp. 248–268. DOI: 10.1080/08838151.2016.1164170. URL: https://doi.org/10.1080/08838151.2016.1164170 (visited on 07/27/2021).

Wickham, Hadley (2021). *Tidyverse: Easily Install and Load the Tidyverse*. manual. URL: https://CRAN.R-project.org/package=tidyverse.

Yaniv, Ilan (2004). "Receiving Other People's Advice: Influence and Benefit". In: *Organizational Behavior and Human Decision Processes* 93.1, pp. 1–13. DOI: 10.1016/j.obhdp.2003.08.002.

Yaniv, Ilan and Shoham Choshen-Hillel (2012). "Exploiting the Wisdom of Others to Make Better Decisions: Suspending Judgment Reduces Egocentrism and Increases Accuracy". In: *Journal of Behavioral Decision Making* 25.5, pp. 427–434. DOI: 10.1002/bdm.740. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.740 (visited on 05/15/2018).

Yaniv, Ilan and Eli Kleinberger (2000). "Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation". In: *Organizational Behavior and Human Decision Processes* 83.2, pp. 260–281. DOI: 10.1006/obhd.2000.2909. URL: http://linkinghub.elsevier.com/retrieve/pii/S0749597800929091 (visited on 05/15/2018).

Yaniv, Ilan and Maxim Milyavsky (2007). "Using Advice from Multiple Sources to Revise and Improve Judgments". In: *Organizational Behavior and Human Decision Processes* 103.1, pp. 104–120. DOI: 10.1016/j.obhdp.2006.05.006.

Yonah, Merav and Yoav Kessler (2021). ""They Don't Know Better than I Do": People Prefer Seeing for Themselves Over Using the Wisdom of Crowds in Perceptual

Decision Making”. In: *Journal of Cognition* 4.1 (1), p. 30. DOI: 10.5334/joc.173. URL: http://www.journalofcognition.org/articles/10.5334/joc.173/ (visited on 08/25/2021).

Zajkowski, Wojciech and Jiaxiang Zhang (2021). “Within and Cross-Domain Effects of Choice-Induced Bias”. In: DOI: 10.31234/osf.io/vzqsw. URL: https://psyarxiv.com/vzqsw/ (visited on 06/30/2021).

Zylberberg, Ariel, Daniel M. Wolpert, and Michael N. Shadlen (2017). “Counterfactual Reasoning Underlies the Learning of Priors in Decision Making”. In: *bioRxiv*, p. 227421. DOI: 10.1101/227421. URL: https://www.biorxiv.org/content/early/2017/11/30/227421 (visited on 01/22/2018).