



Fooled by facts: quantifying anchoring bias through a large-scale experiment

Taha Yasseri^{1,2,3,4} · Jannie Reher³

Received: 29 April 2021 / Accepted: 30 December 2021 / Published online: 13 January 2022
© The Author(s) 2022

Abstract

Through a large-scale online field experiment, we provide new empirical evidence for the presence of the anchoring bias in people's judgement due to irrational reliance on a piece of information that they are initially given. The comparison of the anchoring stimuli and respective responses across different tasks reveals a positive, yet complex relationship between the anchors and the bias in participants' predictions of the outcomes of events in the future. Participants in the treatment group were equally susceptible to the anchors regardless of their level of engagement, previous performance, or gender. Given the strong and ubiquitous influence of anchors quantified here, we should take great care to closely monitor and regulate the distribution of information online to facilitate less biased decision making.

Keywords Anchoring bias · Quasi-experiment · Social influence · Information processing

Introduction

Heuristics are mental shortcuts that enable us to arrive at solutions to complex tasks or problems with minimal effort [25, 31]. However, as has been shown by Tversky and Kahneman [30] these shortcuts come at a cost: to be able to quickly solve a problem, certain information will be simplified, some ignored, and estimations will be made, thus increasing the likelihood of systematic errors in decisions. Commonly referred to as cognitive biases, these errors are the result of non-rational information

✉ Taha Yasseri
taha.yasseri@ucd.ie

¹ School of Sociology, University College Dublin, Belfield, Dublin D04 V1W8, Ireland

² Geary Institute for Public Policy, University College Dublin, Belfield, Dublin D04 N9Y1, Ireland

³ Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

⁴ Alan Turing Institute for Data Science and AI, 96 Euston Rd, London NW1 2DB, UK

processing [11, 15]. Examples of cognitive bias include the anchoring effect (that is, the influence on decisions of the first piece of information encountered), the availability heuristic (where estimates of the probability of an outcome depend on ease of access to that information), the framing effect (the presentation of identical information in different ways), and confirmation bias (a focus on information that supports a pre-existing position), among others. Given the black-box nature of the algorithms that drive searching and selecting relevant information on the Internet [22], it is left to large Internet corporations managing those algorithms to decide which information is “valuable” and therefore displayed to users [24]. Recent experiments by Nikolov et al. [19] and Kramer et al. [13] show that depending on what online platform is used, people may be exposed to extremely different information, which can, in turn, result in the development of “social bubbles” or even changes in their emotional state.

This paper examines one of these effects: anchoring. Tversky and Kahneman [30] observed that in situations where people make estimates or predictions (e.g. the value of a car), the resulting judgement tends to be similar to a previously encountered value (e.g. what the salesperson is offering). The term ‘anchoring’ can therefore be understood as people’s tendency to rely heavily on these prior values (or ‘anchors’) when making decisions. This effect is pervasive and robust in a variety of experimental settings (see [1, 3, 9, 15] and real-world contexts, including in courtroom sentencing [6], in negotiations [10, 29], in financial market decisions [5], in property pricing [20], and in judging the probability of the outbreak of a nuclear war [23]. Typically, anchoring bias occurs when numeric anchors are provided, although some research has also investigated the effect of non-numeric anchors [14].

Most experiments investigating anchoring effects have been limited in their design and implementation. Often experiments have been limited to single questions examining participants general knowledge. For example, subjects have been asked to estimate the percentage of African countries in the United Nations [30], the length of the Mississippi river [12], Gandhi’s age at death [28], and the number of calories in a strawberry [16]. Few of these questions, however, seem particularly engaging or representative of everyday situations. With the assumption that research on social psychology should inform and shape policy and regulations towards information systems (such as social media), it is important to bring the experiments as close as possible to daily life scenarios where those policies would apply. Moreover, participants were often incentivised to take part through monetary reimbursement or course credit. Such a setting does not encourage participants to concentrate fully on the task at hand, given pay is frequently fixed-rate and non-performance-related. The general notion of WEIRD (western, educated, industrialised, rich and democratic) participants and the issues of validity in psychology experiments are present in most of the previous work on anchoring bias too [18].

In recent years, participants were often recruited from online portals like Amazon Mechanical Turk [16, 27] adds further to the artificiality of the experiments. Many experiments only recruited between 30 and 50 participants, which limits the statistical power of the experiments and the generalizability of the results. Lastly, most of these experiments did not systematically quantify and generalise the anchoring effect by diversifying and analysing a large range of questions at the same time

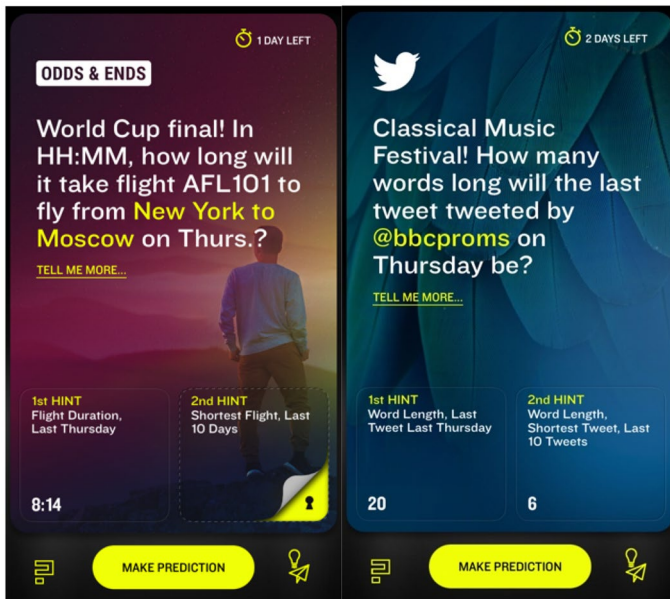


Fig. 1 Screenshots of two example questions during the experiment on the Play the Future mobile application. The first hint is provided by default. The second hint has been unlocked in the right screenshot

or adequately control for subjects already knowing the correct answer to a general knowledge question.

We address these shortcomings through a large-scale online field experiment undertaken on a mobile game (“Play the Future”) which “gives players the power to predict the everyday outcomes of the world’s most popular brands and events for points, prestige and prizes”. Players make numeric predictions regarding the outcome of certain natural, economic, social, sports and entertainment events such as the weather, stock prices, and flight times (Fig. 1), to receive points and accuracy feedback for each prediction, which they can then compare with other users on the app. We know that predictions about future events rely on a complex process in which prior knowledge is used to determine the likelihood of a given event. However, due to time or knowledge constraints, people rely on heuristics or mental shortcuts when making predictions—so-called “bounded rationality” [26]. By systematically manipulating numeric hints provided as part of the game (i.e. manipulating information that is potentially relevant for the judgement process), we examine the influence of anchoring bias on the decisions made by our experimental subjects. As discussed above, anchoring bias is a known and documented phenomenon, our contribution here, therefore, is the provision of a more accurate and detailed image of the phenomenon using a more methodologically sophisticated design.

For the experiment, we designed 62 questions (and corresponding hints) to be provided to the gamers (see Table 1 for example questions). The subjects who answered our questions were not explicitly recruited to the experiment and did not know they were part of the experiment. For each question, two hints are provided

Table 1 Examples of experiment questions

Question	Control group	Treatment groups		Result
		Hint 2 Group A	Hint 2 Group B	
In HH:MM, how long will it take flight AC880 to fly from Toronto Pearson to Paris CDG on Thursday?	Flight duration, last Thursday: 06:36 h	Longest flight duration during 10-day period: 06:51 h	Shortest flight duration during 10-day period: 06:23 h	07:04 h
In USD, what will Tesla's (TSLA) stock price be at the Nasdaq Stock Market next Tuesday (July 3) at 1 PM?	TSLA's stock price last Tuesday at 1 PM: 334.34 USD	TSLA's stock price low last Tuesday: 326.00 USD	TSLA's stock price high last Tuesday: 343.55 USD	310.86 USD

that contain information regarding the topic of the question. The first hint is always shown and is identical for all users. The second hint can be unlocked with a ‘key’. Keys automatically renew themselves every four hours for all users for free. We tried to ensure our questions (and the hints) were typical of the game, to ensure they played on the app normally. The hints provided were accurate values, i.e. while we varied the type of information available to participants, all information provided was correct. The questions are not necessarily relevant to decision-making scenarios; however, they are selected in relation to the current affairs as opposed to public knowledge questions used in previous work. This, we believe, brings the experiment closer to what individuals are exposed to, for example, on social media at the current time.

The experimental treatment was administered in the second (locked) hint. Participants who only saw the first hint are considered to be in the control group (for more details see “[Methods](#)” section). Users who opted to unlock the second hint were randomly assigned to Group A or Group B, who were then provided with the lowest or highest value of a previous outcome (Table 1). We then examined how the provision of both low and high anchors (in the form of factual hints) affects the values predicted by the users. If anchoring is effective, we would expect those provided with a high figure, to predict higher values than those who were provided with low figures. In addition to investigating a general anchoring effect of information on decision making, in an additional set of experiments, we provided non-factual hints in the form of “<someone> from Play the Future Team’s prediction” to compare the size of the anchoring effect induced by “expert’s opinion” to the one by pure factual hints. Finally, we build on research on individual differences in susceptibility to anchoring bias [7] to examine how the bias varies for individual users according to their level of engagement while playing the game, their previous performance, and their gender.

Results

General anchoring effect

We asked 42 questions on six different topics over a month. Each question was answered by an average of 219 users, with an average of 58 participants in the treatment condition (i.e. Groups A and B) for each question. The full list of questions and hints is provided in Supplementary Table 1. The actual results (i.e. the correct answers) do not affect the experiment or analysis, given they were known and revealed only after the experiment was over. The answer distributions to the questions shown in Table 1 are shown in Fig. 2. Yuen’s test for independent sample means with 15% trimming was applied to all questions to test whether the predictions in the two treatment groups are different from each other. The results indicate that the answer distributions are indeed significantly different ($p < 0.01$) in all questions except one, which is marginally significant ($p = 0.05$).

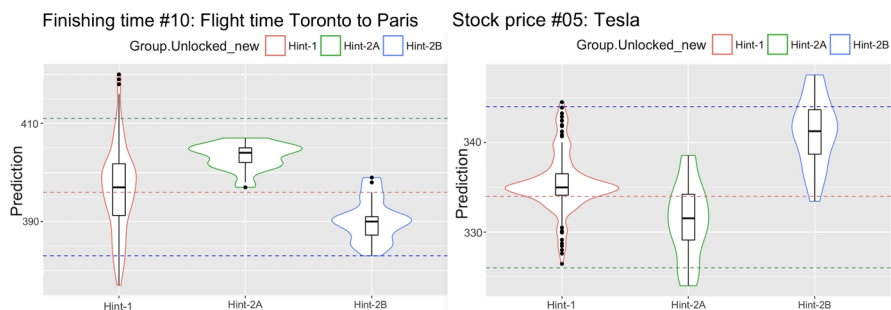


Fig. 2 Distribution of participants' answers depending on group allocation for the two example questions in Table 1. Provided anchors in each group indicated by dotted lines. The diversity of predictions in the two treatment groups is smaller in the left graph compared to the right

Eight questions providing irrelevant information in the first hint were included in the experiment as a check to determine whether control group participants are blindly following the values provided in the first hint without questioning their actual relevance for the target outcome (see the questions in Supplementary Table 1). If this were the case, it would be inferred that the cognitive effort invested by participants in this game is minimal and that the observed responses are not due to anchoring effects but to users' laziness. The detailed results of this test are provided in Supplementary Figs. 1–2, but in short, it was confirmed that the players of the game, generally pay attention to the relevance of the hint to the question and do not follow the values in the hint blindly.

For 42 standard questions, the anchoring stimulus (Eq. (1)) and the anchoring response (Eq. (2)) were calculated and shown in Fig. 3.

$$\text{Stimulus} = \frac{|\text{Hint Group A} - \text{Hint Group B}|}{\sigma_{\text{Control Group}}} \quad (1)$$

$$\text{Response} = \frac{|\text{Median answer Group A} - \text{Median answer Group B}|}{\sigma_{\text{Control Group}}}, \quad (2)$$

where Hint Group X is the numerical value of the provided hint to group X and σ is the standard deviation of all the answers in the group.

As expected, a larger stimulus leads to a larger anchoring response, however, after an initial increase in the size of the induced anchoring effect, saturation appears.

A closer look at examples shown in Fig. 2, however, reveals that in some cases the diversity of answers in the treatment groups is very small (left panel) whereas in other cases the answers are widely dispersed (right panel). This observation suggests that for some questions the provision of two anchors lead to higher collective prediction certainty when compared to the control group, while for other questions it introduced more doubt about the true value of the likely outcome among the group members.

This observation warrants a more systematic analysis of the ratio of the treatment to the control group's standard deviation depending on the size of the provided stimulus.

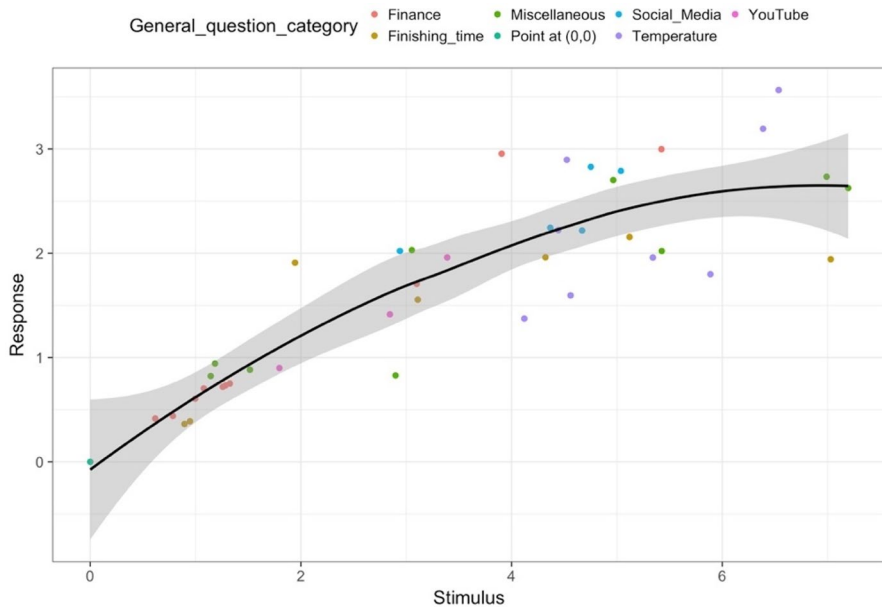


Fig. 3 Anchoring response function based on 42 standard questions. LOESS curve fitted to all data points. Points coloured according to question category

Figure 4 illustrates how the relative group diversity—that is the diversity of answers within each treatment group divided by the diversity of predictions in the control group (Eq. (3))—changes with the group stimulus (Eq. (4)).

$$\text{Relative group diversity } X = \frac{\sigma_{\text{Group } X}}{\sigma_{\text{Control Group}}}, \quad \text{where } X = A, B \quad (3)$$

$$\text{Group stimulus} = \frac{|\text{Hint Group } X - \text{Median answer Control Group}|}{\sigma_{\text{Control Group}}}, \quad \text{where } X = A, B \quad (4)$$

Smaller anchoring stimuli lead to smaller relative group diversity, which means that treatment group users were collectively more certain of their answers compared to the control group for these questions. When the size of the anchoring stimulus increases, the relative group diversity also increases.

Based on this observation, we define a new measure of the anchoring bias by normalising the difference between the median answers of the two treatment groups by the average of the standard deviations of the two treatment groups instead of the standard deviation of the control group (Eq. (5)).

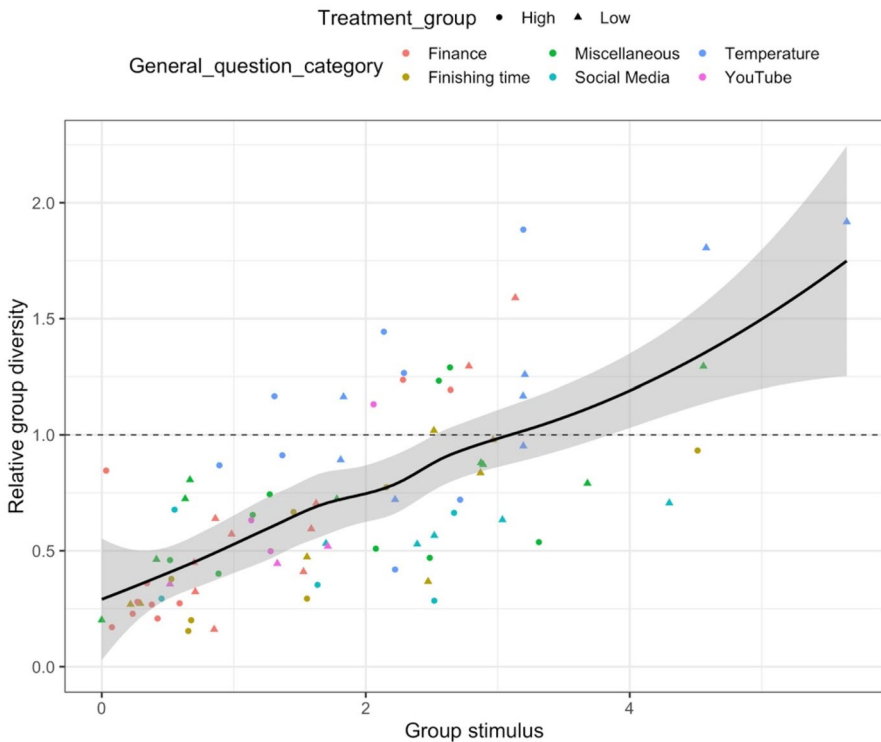


Fig. 4 Relative group diversity vs. group stimulus. Points coloured according to question category and shaped according to group allocation (high vs. low anchor). The line at $y = 1$ indicates the equal ratio of the treatment to the control group's standard deviation

$$\text{Modified response} = \frac{|\text{Median answer Group A} - \text{Median answer Group B}|}{\left(\frac{\sigma_{\text{Group A}} + \sigma_{\text{Group B}}}{2}\right)} \quad (5)$$

The resulting modified response accounts for the participants' collective confidence in their predictions. A small average standard deviation of the two treatment groups is assumed to be indicative of higher certainty of answers among participants since users appear to be in collective agreement regarding the true value of the target. The modified response function is shown in Fig. 5.

Medium-sized stimuli ($2 < x < 5$) seem to have caused the majority of participants to believe that the anchors might be plausible, resulting in a larger modified response, i.e. a larger difference in the answers of the two groups with higher collective confidence in each group. High anchors induce more uncertainty among participants: not all users follow high anchoring stimuli, instead, a considerable proportion of participants starts adjusting their predictions to less extreme values, thus increasing the diversity of answers.

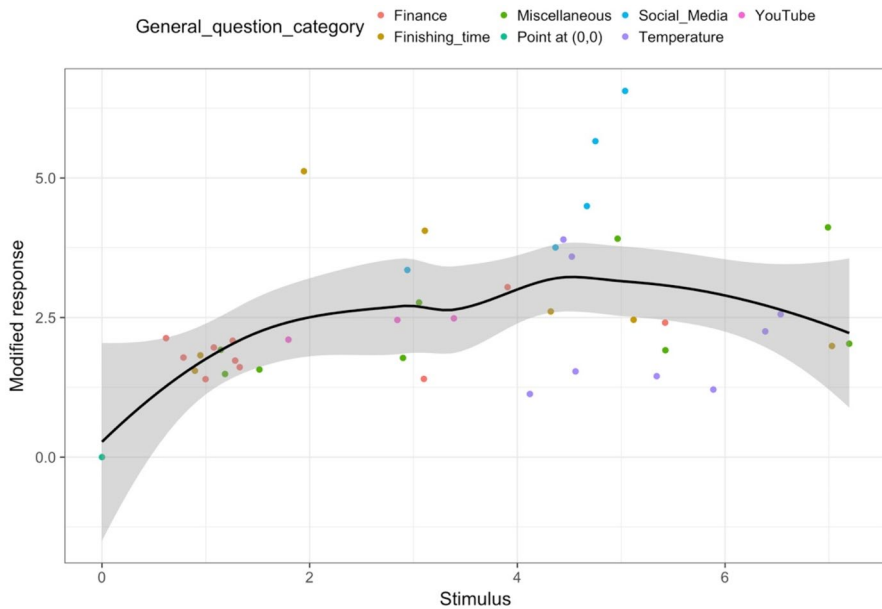


Fig. 5 Modified response vs. stimulus. Points coloured according to question category

“Expert-opinion” anchors

By substituting the factual information in the second hints by fictitious “Play The Future-prediction” values (which we carefully selected to resemble realistic predictions) in additional 12 questions, we examined how strongly these values impact participants’ predictions. The hints of these questions presented as the prediction of a hypothetical member of the Play The Future team. An example is given in Table 2 and the rest of the questions are in Supplementary Table 1.

The upper panel of Fig. 6 shows the size of the anchoring effect versus the size of the anchoring stimulus for 42 standard and 12 PTF-prediction style questions (note the data points belonging to the standard questions in Fig. 6 are identical to ones of the Figs. 3, 4, 5 and are repeated here for the purpose of comparison). It becomes immediately apparent that the questions containing PTF-prediction values in the treatment hints result in consistently larger responses compared to the standard questions with factual information. The middle panel of Fig. 6 shows that not only the medians of treatment groups distributions are moved further with the fictitious hints, but also the relative group diversity tends to be lower for PTF-prediction questions, which is the result of less variation in the treatment groups’ answers compared to the control group’s predictions. Hence, in the modified response curve is shown in the lower panel of Fig. 6 we see an even larger amplification of the anchoring effect emerged from the fictional hints.

Individual analysis

To analyse the effect of the anchoring stimulus provided in the experiment on each participant's predictions, the median bias for each user during the experiment was calculated. Firstly, the normalised difference between the user's prediction and the control group's median prediction was computed for each question (Eq. (6)).

$$\text{Bias per participant } i \text{ per question} = \frac{|\text{Prediction}_i - \text{Median answer Control Group}|}{\sigma_{\text{Control Group}}} \quad (6)$$

Next, the median bias was calculated per individual user for all questions answered in the control condition and for all questions answered in the treatment condition. A higher individual bias indicates that a certain user's prediction values are further away from the median of the control group's predictions, which may be the result of the influence of the anchoring stimuli.

User engagement

It could be hypothesised that the anchoring effect is stronger in less experienced users. We, therefore, tested for a difference between participants who answered less than half of all experimental questions (casual users) and those who made predictions for more than half of the questions ("loyal users").

Results are shown in Supplementary Fig. 3. Focussing on control groups only, we observe that loyal users in the control group seem to make predictions that resemble the median control group's predictions more closely than casual users in the control group (Welch's *t*-test significant $t=6.986$, $p<0.001$). The highly engaged users may have concluded that making moderate rather than extreme predictions (if no further information in the form of a second hint is provided) constitutes a relatively successful strategy in this game.

However, among the treatment groups, barely any difference between casual and loyal participants can be detected (Welch's unequal variances *t*-test insignificant $t=1.448$, $p=0.149$). This implies that all users regardless of their level of engagement on the app are roughly equally susceptible to the provided anchoring stimuli. Thus, it is concluded that even among high-frequency players no 'learning effect' regarding the true purpose of this experiment occurred.

User prior accuracy

Many of the experiment participants had already played the game and the records of their predictions were available to us. However with the caveat that there were not enough prior data for each category of questions and we had to calculate users' prior accuracy based on their average performance at the aggregate level rather than a question specific level. This "coarse-graining" could be rectified in future work. We divided the players into high accuracy and low accuracy groups based on their

Table 2 Example of experiment question with fictitious hints

Question	Control group	Treatment groups		Result
		Hint 1 (identical for all users)	Hint 2	
In USD, what will Sotheby's (BID) closing stock price be at the NYSE on Tuesday?	Hint 1 (identical for all users) Closing stock price last Tues.: 59,06 USD	Group A	Hint 2 Group B	55.84 USD
		Sally from PTF's prediction: 60.31 USD		Sally from PTF's prediction: 57.81 USD

Fig. 6 Upper panel: response vs. stimulus. Middle panel: relative group diversity vs. stimulus. Lower panel: modified response vs. stimulus for 42 standard and 12 PTF-prediction questions. LOESS curve only fitted to standard anchoring questions for reference

accuracy score in all the games they had played before our experiment and compared the induced bias for the two groups (see “[Methods](#)” section for details).

Considering only control groups, participants who made less accurate predictions before the start of the experiment provided answers that were relatively close to the overall median answer during the experiment (Supplementary Fig. 4). Previously better-performing users made more distinct predictions, potentially because they put less trust in the information provided in the first hint. The results of Welch’s *t*-test indicate that the individual biases in the two control groups are slightly different from each other ($t(139.11)=1.933$, $p=0.055$), however, this result is statistically significant only at the 0.1-level.

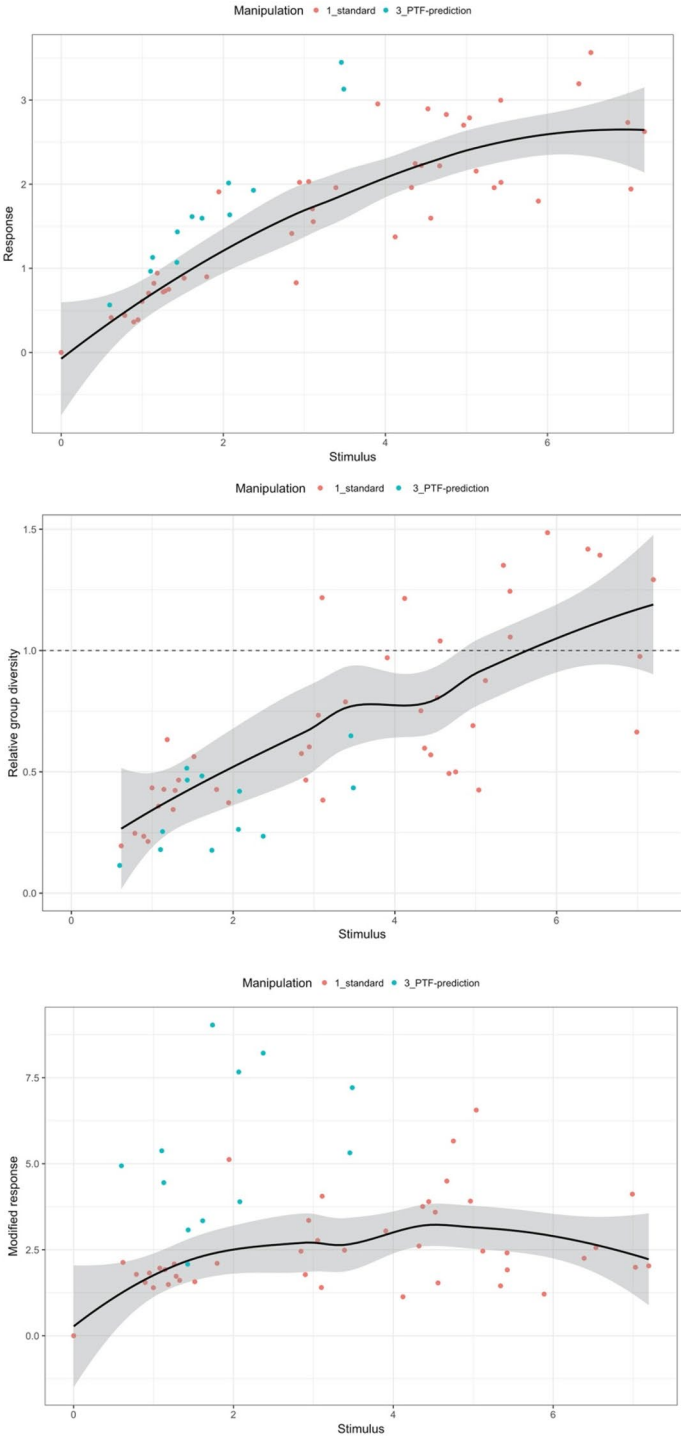
For previously low-performing users in the treatment group, the individual bias seems to be slightly larger compared to the bias observed for previously well-performing users. This would imply that their answers were slightly more influenced by the anchoring stimuli compared to better-performing users. However, Welch’s *t*-test reveals that the visually observed difference is not statistically significant ($t(97.99)=-1.443$, $p=0.15$).

Gender

Finally, we analysed the users based on their gender and compared their cross-group errors (see Supplementary Fig. 5). Even though Welch’s *t*-test with unequal variances confirms that this difference between males and females in the control condition is indeed significant ($t(349.86)=-2.412$, $p=0.016$) meaning that female users tend to make predictions that are closer to the overall median answer compared to male users in the control conditions, both the visual analysis and Welch’s *t*-test for the same comparison in the treatment condition show that there is no difference among male and female individual biases ($t(194.92)=-0.929$, $p=0.354$). Thus, both sexes appear to be equally susceptible to the anchoring stimuli provided in the treatment condition.

Discussion

Over one month, 549 participants made numeric predictions to over 62 questions about the outcome of a variety of future events after having been exposed to different numeric anchors. Overall, 13,000 predictions were recorded during the experiment. After defining and applying a consistent normalisation method to the distributions of answers to all questions, the various anchoring stimuli and responses were compiled to form the anchoring response function.



The calculation of the anchoring index (AI) proposed by Jacowitz and Kahneman [12]¹ for individual hints in all groups results in a mean AI of 0.61 for all questions. This means that participants did not fully accept the presented anchoring stimulus in their predictions (which would be the case if $AI=1$) but instead adjusted their predictions to a level of roughly 60% of the size of the anchor on average. This effect is in agreement with the one observed by Jacowitz and Kahneman [12], who found a mean AI of 0.49 (with slightly different effects for low and high anchors). This large size of the anchoring index here is likely attributable to the specific nature of the provided anchors: firstly, the numeric hints usually contained information about the size of the target outcome in the past, making the anchors appear somewhat relevant to the question. Secondly, the anchor value and target outcome were typically provided in the same unit, such as °C, USD, etc. This provides support for Mochon and Frederick's [16] scale distortion theory, which proposes that anchors are effective when the stimulus and response share the same scale. Here, further research should be conducted by reducing the relevance of the information provided in the hints even further and varying the units.

The relationship between the anchoring stimuli and responses we find advances the result found by Chapman and Johnson [2], who suggested that a limit of the anchoring effect exists when stimuli become too extreme. We see a monotonic, yet decelerating increase in the response curve, and when we modify the measured response by considering the standard deviation of the predictions, we see a decline in the modified response curve. This result is confirmed in the present work by the modified anchoring response function, which resembles the shape of an inverted U. By incorporating the standard deviation of predictions in the treatment groups as a normalisation factor it was observed that larger anchoring stimuli not only translate to larger responses but also lead to higher collective uncertainty among participants.

This finding suggests that the limit of the anchoring bias finds expression in a higher variation of participants' answers rather than in a smaller response. The greater diversity of predictions may be explained by both the original anchoring-and-adjustment, and the selective accessibility theory: participants may have started an insufficient adjustment process from the implausible anchor, or searched-for information confirming the value presented in the anchor. The lack of clear evidence for the presence of one of these mechanisms may suggest that both theories are applicable in this case. Further lab experiments could shed more light on this.

On the level of the individual user, it was shown that all participants in the treatment groups were equally affected by the anchoring stimuli in their responses. This holds regardless of participants' level of engagement during the experiment, their previous performance on the app, or their gender.

One of the implications of this study can be in marketing, where the anchoring effect has already been discussed in the context of pricing [20]. Following our results, an optimal pricing strategy might be to first produce a distribution based on prices suggested by several different individuals. After cleaning the distribution and

¹ Jacowitz and Kahneman [12] compute the AI as follows: $AI = \frac{\text{Median (high anchor)} - \text{Median (low anchor)}}{\text{High anchor} - \text{Low anchor}}$.

removing the outliers, calculate the median and standard deviation of the population, and then opt for a price that is 2–5 standard deviations higher than the median of the distribution. This proposition however needs proper testing through further experimentation.

We must note that our work has a limitation that is mostly due to the experimental design: the anchors provided to the participants in the form of hints had to include relevant information to keep users engaged on the app, whereas, Chapman and Johnson [3] argue that proof for the presence of the anchoring bias is strongest when the information influencing participants' answers is uninformative because subjects had no rational reason to follow that value. The present experiment uses two consecutive anchors that both contain relevant information, however, users have to "pay" to unlock the second hint. Hence, participants' predictions might be biased towards the treatment hint partly because they implicitly (and irrationally) associate higher importance with it (endowment effect). To counteract this effect as much as possible, the information provided in the second hint in many questions was even less pertinent for answering the searched-for target outcome than the first hint, or equally uninformative. Thus, the anchoring bias observed in this experiment is due to *overly influential* factual information [2].

A major contribution of this experiment lies in the ecological validity of this research. Not only does the overall setting of the experiment in the form of a mobile application closely resemble a real-world situation in which information is provided by various sources and subsequently processed by the individual, but the different questions also reflect realistic forecasting decisions.

The fact that the anchoring effect was especially strong for questions in which the anchoring stimuli were labelled as 'PTF-prediction' values constitutes a very important finding of this experiment. This suggests that people's perception of the usefulness or veracity of information can be influenced by the provision of an "authority" label. In an era in which clickbait and fake news are frequently used tools to compete for people's attention and influence their opinions, individuals must understand the strong impact of purposefully (or even unintentionally) set anchors on their perception of certain events: the slogan "We send the EU £350 million a week—let's fund our NHS instead" that was used by the *Leave*-campaign in Brexit vote in 2016, is a prime example. It goes without saying that social media provide people with power and influence a great platform for providing "advice" to the members of the public and considering the effects reported in our work, it becomes even more important to regulate and verify such information before their spread on social media.

Further research should investigate whether anchors with the same numerical value that are associated with different sources produce distinct anchoring effects. Moreover, given the specific experimental platform used in this work, it would also be interesting to determine how information about predictions made by other users on the app influence participants' predictions as part of a collective intelligence experiment in future work. Considering the design of social media platforms which "bombard" the users with information of other users' actions and decisions (social information) and previously reported influence of social information on decision making [17], it becomes ever more relevant to ask how cognitive biases in general

and anchoring bias, in particular, can be amplified and consequently hinder collective intelligence in this new informational environment.

Methods

Experimental design

The platform we used for the online field experiment is an application (app) for smartphones called “Play the Future” (PTF), which has been in operation since the beginning of 2016. It is freely available for both Android and iOS devices.² Over 300,000 users are currently registered on the app, with around 82,000 monthly active users making numerical predictions regarding the outcomes of various economic, social, sports, entertainment, and other events. For each question, two hints are provided that contain relevant information. The first hint is always shown, whereas the second hint can be unlocked with a ‘key’. Keys automatically renew themselves for free every four hours so that up to a maximum of three keys are available at any given time.

Participants were not specifically recruited for this experiment; all users who were active on the PTF application during the period of the experiment (May–June 2018) were treated as participants. The experimental questions were both thematically and structurally very similar to regular PTF questions (apart from the distinction between the two user groups), thus, users were not explicitly notified about taking part in an experiment, which adds to the real-world character of the research. Users were not paid for their participation (the game is free to use); they used the app out of their own volition and could discontinue playing at any point.

The experimental treatment is administered in the second (locked) hint: Its description and content are different for two user groups A and B. By recording each unlocking of the second hint it can precisely be determined which participant was exposed to the treatment and who belongs to the control group on a per-question basis. Participants who have only seen the first hint are considered to be in the control group. The information provided in the first hint is identical across all users. The second hint reveals either a low or a high value (depending on the group) of the target outcome during a specific time frame. We devised the questions and hints in a way that no experimental group has an advantage over the other group.

The two second hints contain values that are above and below the value shown in the first hint. Ideally, the distance to the medium value is roughly the same for both groups, however, due to the hints containing real-world factual information this was not possible in all cases. While generally, the hints are factually accurate, some questions include a fictitious “PTF’s prediction” value in the second hint to explore whether users are equally anchored by facts and predictions from unverified sources. Additionally, some questions contain irrelevant information in the hint for the

² The app is currently unavailable in Europe, due to adjustments being made to accommodate the General Data Protection Regulations which took effect on May 25, 2018.

control group to test whether that information also serves as an anchor. The current events and factual information for all questions and hints were manually researched by the authors. Please refer to Supplementary Table 1 for a list of all questions.

Out of the six to ten questions that were asked on the app every day, two to four questions were part of the experiment and the remaining questions were regular, non-A/B testing questions designed by the PTF team. Each question was available for a limited period of approximately 24–48 h. Within that period users could freely decide whether and when to submit a response, thus, time pressure was minimal. We did not control for whether users gathered additional information (e.g. consulting weather statistics) before making a prediction. Users received an accuracy score and point proportional to that two to four days after the results were in.

At the beginning of the experiment, all the registered users were randomly allocated to two equally sized groups A and B and new users were subsequently assigned to one of the two groups. The second hints were designed in a way that each participant would receive a roughly equal number of high and low second hints in alternating order. Participants decided on a per-question basis whether to unlock the information contained in the second hint or not given their limited number of keys. This self-selection into the treatment or non-treatment condition is characteristic for online field experiments [21]. As users need to log in to play the game, we could keep track of their actions and the various treatments, as well as their gender and country.

Ethical approval

Ethical approval for this experiment was obtained from the Social Sciences and Humanities Interdivisional Research Ethics Committee (IDREC) of the University of Oxford (CUREC 2 reference number: R52154/RE001). The users had agreed to the terms and conditions laid down by PTF upon signing up for the app. Moreover, since the questions in the experiment were in line with previous questions asked on the app, participants could not notice the difference between the experimental and normal settings. At no time were participants exposed to any risks; no deception took place and the questions only pertained to events for which the results will be publicly known and that were appropriate for all ages. As a further precaution, the data collection was kept to an absolute minimum, and any personal information was anonymised by the PTF team before it was released to the researchers for analysis, meaning that no user was individually identifiable.

Participants

Overall, 549 users took part in the experiment by answering at least one of our questions. One-third of participants only answered five questions or less. Roughly one-quarter of users answered 80% of all experimental questions or more. 56% of users identified as female, 41% as male, and 3% did not disclose their gender. Users are predominantly from North America (60% from Canada and 36% from the US).

Data preparation and analysis

Outlier removal and key statistics for each question

The visual inspection of the three answer distributions (control group, group A, group B) showed that all groups contained outlier values. Extreme values can bias a distribution's key statistics so that the calculated and standard deviation are not representative of the data's actual distribution, thus necessitating their exclusion. The outlier answers could be by players who decide to type in a number and submit without thinking about the question or misunderstood the question altogether. This is evident from rare answers which are orders of magnitude off or physically impossible (e.g. a three-digit prediction for weather temperature). The existence of such outliers is inevitable in any "field" experiment and it is common practice to remove them from the dataset before analysis. However, the removal should be done according to a pre-defined systematic protocol to avoid "cherry-picking". We developed such a protocol based on our preliminary observations and prior to data collection and cleaning: data points were classified as outliers and consequently removed if they lay outside the range of median ± 2.5 times the median absolute deviation (MAD).

For subsequent analyses (see below), the median prediction value was calculated for each of the three groups based on the cleaned dataset. Even after removing the most extreme values from the dataset, the distributions may still be skewed. Thus, the median rather than the mean prediction serves as a better basis when comparing the various groups because it is more robust to skewness in distributions and therefore representative of participant's collective prediction value in each group [12].

Curve fitting

The anchoring response curve is the result of plotting a smooth curve through the resulting data points. A non-parametric model relying on local least squares regression (LOESS) was used to create the curve, thus eliminating the need to determine the specific form of the model before fitting [4, 8]. This method is particularly suitable for data exploration. The shaded area around the LOESS curve indicates the 95% confidence interval.

Error index

Users' previous performance on the app is indicated by an error index. The error index is based on the accuracy of the last 10 predictions (or fewer depending on how many times a given user had previously played on the app) before the experiment started. For each of those questions, the absolute difference between the user's prediction and the actual result was calculated and normalised by the standard deviation of all predictions for that question according to Eq. (7) (after having removed extreme answers as described above). Subsequently, the median of all calculated

error scores per participant was computed. A lower error score is indicative of a better previous performance.

$$\text{Error score per participant } i \text{ per question} = \frac{|\text{Prediction}_i - \text{Actual result}|}{\sigma_{\text{All predictions (cleaned)}}} \quad (7)$$

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42001-021-00158-0>.

Acknowledgements We thank the Canadian company Play the Future, Inc. that has kindly provided the app for research purposes. We also thank Myrto Pantazi for comments on the experimental design, David Sutcliffe for valuable comments on the manuscript, and Mahdi Nasiri for help with analysis. TY was partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. The funder had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author contributions TY outlined the study and designed the experiment. JR generated the experiment questions and completed data collection and analysis under the supervision of TY. TY and JR wrote the manuscript.

Funding Open Access funding provided by the IReL Consortium. TY was partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. The funder had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Data availability Anonymised data including the raw participants' predictions and the protocol for analysis is available for sharing.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Ethical approval for this experiment was obtained from the Social Sciences and Humanities Interdivisional Research Ethics Committee (IDREC) of the University of Oxford (CUREC 2 reference number: R52154/RE001).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Bahník, S., Englich, B., & Strack, F. (2017). Anchoring effect. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (2nd ed., pp. 223–241). Routledge.
2. Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, 7(4), 223–242.

3. Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). Cambridge University Press.
4. Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596–610.
5. Collins, J., De Bondt, W., & Wärneryd, K. (2018). Looking into the future. In A. Lewis (Ed.), *The Cambridge handbook of psychology and economic behaviour* (pp. 71–100). Cambridge University Press. <https://doi.org/10.1017/9781316676349.004>
6. Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32(2), 188–200.
7. Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26, 116–133.
8. Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). SAGE Publications Inc.
9. Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40, 35–42.
10. Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4), 657–669. <https://doi.org/10.1037/0022-3514.81.4.657>
11. Haselton, M. G., Nettle, D., & Murray, D. R. (2016). The evolution of cognitive bias. In D. M. Buss (Ed.), *Handbook of evolutionary psychology* (pp. 968–987). Wiley.
12. Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11), 1161–1166.
13. Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.
14. LeBoeuf, R. A., & Shafir, E. (2006). The long and short of it: Physical anchoring effects. *Journal of Behavioral Decision Making*, 19, 393–406.
15. Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin and Review*, 25, 322–349.
16. Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgements. *Organizational Behavior and Human Decision Processes*, 122, 69–79.
17. Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, 341(6146), 647–651.
18. Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, 31(6), 678–701.
19. Nikolov, D., Lalmas, M., Flammini, A., & Menczer, F. (2018). Quantifying Biases in Online Information Exposure. *Journal of the Association for Information Science and Technology (JASIST)*. Advance online publication. Preprint at <http://arxiv.org/abs/1807.06958v1>
20. Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, 39, 84–97.
21. Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: Studying social interactions in context. *Social Psychology Quarterly*, 80(1), 1–19. <https://doi.org/10.1177/0190272516680842>
22. Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
23. Plous, S. (1989). Thinking the unthinkable: The effects of anchoring on likelihood estimates of nuclear war. *Journal of Applied Social Psychology*, 19(1), 67–91.
24. Schroeder, R. (2014). Does Google shape what we know? *Prometheus*, 32(2), 145–160.
25. Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134, 207–222.
26. Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). North-Holland Pub. Co.

27. Smith, A. R., Windschitl, P. D., & Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level. *European Journal of Social Psychology*, 43, 97–108.
28. Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73(3), 437–446.
29. Thorsteinson, T. J. (2011). Initiating salary discussions with an extreme request: Anchoring effects on initial salary offers. *Journal of Applied Social Psychology*, 41(7), 1774–1792.
30. Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
31. Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293–315.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.