Data & Knowledge Engineering 162 (2026) 102530

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Debiasing judgmental decisions by providing individual error pattern feedback

Nathalie Balla, Thomas Setzer *

Catholic University of Eichstätt-Ingolstadt, Ingolstadt, Bavaria 85049, Germany

ARTICLE INFO

Keywords:
Decision support system
Debiasing
Error feedback
Self-reflection
Collaborative intelligence

ABSTRACT

We present a Decision Support System (DSS) that provides experts with feedback on their personal potential bias based on their previous error pattern. Feedback is calculated using a knowledge database containing a library of biases and typical error patterns that suggest them. An error pattern means any identifiable structure of errors. For instance, an inference engine might detect continuously too high forecasts of an expert submitted via a user interface, regularly exceeding the actual quantities observed later. The engine might then positively evaluate a rule indicating an overestimation bias and provide feedback on the detected error pattern and/or the presumed bias, potentially including further explanations. As the feedback stems from an expert's own error pattern, it intends to enhance their self-reflection and support wise consideration of the feedback. We assume that this allows experts to acquire knowledge about their own flawed judgmental heuristics, that experts are able to apply the feedback systematically and selectively to different decision tasks and to therefore reduce their potential bias and error. To test these assumptions, we conduct experiments with the DSS. Therein, subjects provide point estimations as well as certainty intervals and subsequently receive error feedback given by a machine based on his or her previous answers. After the feedback, subjects answer further questions. Results indicate that subjects reflect on their own error pattern and apply the feedback selectively to further, upcoming estimations and reduce overall bias and error.

1. Introduction

The accuracy of estimations is important for businesses as they are the foundation for crucial decisions and ultimately impact a company's success. Many of these decisions build on judgmental approaches, where decision makers typically have individual attitudes and estimation heuristics and therefore exhibit different biases [1–3]. Hence, a key research topic in Decision Support System (DSS) research is debiasing, meaning to ameliorate decision outcomes and informing decision makers about potential biases [4].

Despite using supposedly well designed DSSs that support by filtering and visualizing information to drive rational decisions, research demonstrates that decisions often still come out systematically erroneous, including biases like overconfidence, mean or regression bias, or anchoring [5–12].

As an example, Blanc and Setzer [11] identify mean and regression biases in expert corporate cash flow forecasts. The authors find that error patterns can be detected statistically and then corrected automatically, overall enhancing accuracy. Automated correction refers to the correction of experts' forecasts by a statistical model learned on previous experts' errors. While their results and the

* Corresponding author.
E-mail address: Thomas.Setzer@ku.de (T. Setzer).

https://doi.org/10.1016/j.datak.2025.102530
Received 24 February 2024; Received in revised form 7 October 2025; Accepted 13 November 2025
Available online 15 November 2025

outcomes of other studies show that machines (i.e. machine learners) can detect consistent judgmental error patterns in practical settings, auto-correction applied to expert judgments without consideration of possible differences in certainty in the original judgment may also “correct” well-made expert estimations in the wrong direction. This can lead to unnecessary higher errors of a part of the auto-corrected estimations [12].

The capability of DSSs to adapt to the individuality of humans is crucial for their effectiveness. This is due to different reasons such as humans having different biases, thinking processes, and preferences. The importance of this capability of DSSs is emphasized by de Lima Neto, Lima Martins, and Vossen [13], who propose a semiotic-inspired machine for personalized decision support and discover that it helps to find well-suited decisions. Previous research has further shown that it is beneficial to include humans and machines into interactive processes to combine their complementary strengths [14–16]. For instance, after a literature review on human-machine collaboration regarding Natural Language Processing, Deshmukh and Shahade [16] conclude that for creating content and making decisions the collaboration of humans and machines leads to better results. Humans are better at recognizing novel or transferable situations, or unseen developments based on domain knowledge and intuition and integrating contextual information. Machines are stronger in extracting regular patterns from data [17–20].

In this sense, as judgmental errors might have structure, we present a DSS that learns patterns in individual error data of an expert’s past judgments and feeds back potential systematic patterns (and potential underlying biases) found to that expert. The expert (not the machine) then decides how to incorporate the feedback in further judgments. The intention of the DSS is to increase estimation accuracy and make aware of and allow to reduce bias, particularly over- and underestimation and overconfidence. In addition, the DSS intends to mitigate the false-correction problem with auto-correction.

Following the assumption that experts can apply the feedback in a beneficial fashion, the feedback should be rejected if a decision belongs to a domain in which the expert is well versed or is confident that the feedback is not applicable to the judgment at hand. If an expert is less certain to be unbiased, the feedback should be considered.

To test whether this holds true, we conduct experiments in which subjects provide point estimations of quantities from varying categories, which are not disclosed to the participants, together with certainty intervals. The categories are meant to represent structures for the human to recognize, featuring different expert tasks where different heuristics are applied and in which experts have different levels of knowledge. The feedback consists of the personal mean bias and is given after a first sequence of questions. It is meant to stimulate awareness of potential biases and self-reflection, and contemplation on how strongly and to questions of which category to apply the feedback.

We test whether humans can recognize categories, reflect on own, potentially category-specific error patterns, and whether they can wisely and selectively apply feedback to reduce error and bias. Further, we test if humans perform better, applying feedback than machines applying auto-correction.

Overall, the contribution of this work comprises an approach to achieve collaborative intelligence by using a DSS that involves individual error pattern feedback of point estimates learned by a machine, which can then be applied selectively by the human. To our knowledge this has not been examined so far. There are studies investigating the use of DSSs with judgmental adjustment [21] or other concepts of combining judgments and statistics [22], and studies on the impact of feedback on human judgments such as in Kim et al. [23]. However, our work differs from existing work as we aim at avoiding false auto-corrections through selective adjustments, reducing bias and allowing for reflection using individual error pattern feedback of point estimations.

The paper is structured as follows. In Section 2, previous research on auto-correction, feedback and self-reflection is reviewed. In Section 3, the infrastructure and procedure of our proposed DSS is described. In Section 4, the research design of our experiment is presented. In Section 5 the results are shown, which are then discussed in Section 6. Finally, in Section 7 the work is summarized and an outlook for future research is provided.

2. Prior work on auto-correction, feedback, and self-reflection

In this section, we review literature on auto-correction, feedback that is related to humans’ biases, and self-reflection.

2.1. Auto-Correction

Although the intention of DSSs is to enhance decision making and reduce biases, often DSS-based decisions still exhibit systematic error. However, DSSs can also be used to detect such systematic errors, whether originally supported by DSSs or not, which is considered for instance, in Goodwin [24], who discovers mean and regression biases in judgmental sales forecasts and applies statistical correction leading to cost savings.

This is in line with the findings by Blanc and Setzer in [11], where the authors study accuracy gains of corporate cash-flow forecasts when applying auto-debiasing. They observe overall accuracy gains with auto-correction, i.e., when replacing the expert predictions with the corrected predictions.

However, their results in [12] also show that on their empirical data set the variance of error differences between original and corrected forecasts increase with the magnitude of corrections: the decile bins with the highest discrepancy have the strongest accuracy improvements and deteriorations. The authors reason that auto-correcting estimates without considering how sure the experts are, can consequently also lead to correcting initially accurate estimations in a detrimental way, resulting in large errors especially in outer decile bins.

Based on these findings, the authors propose a DSS presenting the expert, after the submission of their judgment, the prediction of a statistical correction model including a specification of the bias possibly pushing the gap to the statistical estimation. The suggestion is to derive this kind of benchmark prediction by correcting error patterns learned from previous estimations that are consistent over time to a current estimation task and inviting the expert to either adopt or to edit the model prediction. In the ideal case, the expert would overwrite those model predictions that lead to large false corrections.

2.2. Feedback

The form of feedback plays a primary role in whether and how experts are willing to accept or reject the feedback and choosing the right feedback type is therefore key to such feedback systems. A known differentiation of feedback types is outcome feedback (OFB) and cognitive feedback (CFB). OFB refers to “information that describes the accuracy or correctness of the response” [25], and constitutes often only the correct answer. CFB is “information regarding the how and why that underlies this accuracy” [25].

Some researchers [26,27] find evidence that OFB, when simply giving correct answers, is fairly useless, as also found in many studies generally considering OFB rather effectless and even obstructive [27]. However, OFB in other forms can be beneficial – for example, as individual performance feedback as shown in Benson and Önkal [28] who investigate the latter in probability estimation. The subjects in their experiment make four weekly predictions of football games for the following weekend for the winning probability of a certain team. In the treatment group performance feedback is provided, whereas this is not done for the control group. Their results show that the performance feedback leads to forecasting accuracy improvement.

Examining OFB and CFB, Sengupta [29] performs experiments in which participants are prompted to make decisions on personnel screening. OFB together with CFB is provided to the treatment group and only OFB to the control group. In this case, OFB is represented by rating decisions made by the expert committee and CFB by the committee’s decision strategy referring to similar jobs, consistency scores, and information regarding a participant’s own strategy. Confirming findings of similar research, results show that combining OFB and CFB supports participants in exceeding the performance of those participants only receiving OFB.

The aim of addressing and reducing cognitive bias by combining OFB with CFB through a DSS has also been pursued by other researchers [30–32]. Nussbaumer et al. [30] present a framework to automatically detect the confirmation bias and apply different feedback methods to mitigate the confirmation bias. A questionnaire must be answered before and after the subject completes a task, based on which a detection algorithm identifies a bias. The different feedback methods range from changing the perspective of data by presenting it in a different form, to computer-aided questions that call the current decision into question.

Dunbar et al. [31] investigate different feedback designs in a serious game to decrease cognitive biases with 411 participants from US universities. They compare the effect of game-based learning to a professional training video. The subjects received feedback based on their actions in the game either just in time or delayed. The results show that the digital game was substantially more effective compared to the video training and the just in time feedback seems not to be more effective than the delayed feedback.

Król and Król [32] examine giving feedback with an automated feedback mechanism in the field of financial education with two groups of 100 students. They conducted two experiments, where the first experiment serves to train an algorithm to predict whether a subject will make a good decision, based on eye movements. The second experiment is designed on the outcome of the first experiment in that a different group of subjects is given “eye feedback” after every decision, which is produced by the algorithm of the first experiment. The “eye feedback” is however only a color (green or red) depending on a “good” or “bad” decision. The findings of these experiments indicate that the algorithm can evaluate the decision of a subject based on examples of other subjects and that the eye feedback creates a change in attentional patterns, which leads to enhanced decisions.

Overall, these research contributions explore feedback mechanisms in DSS, while, to our knowledge, there has been no research conducted on automated and personalized feedback on point estimations based on error patterns of the respective subject aimed at differentiated, selective feedback consideration.

As of its particular importance to our work and experiment, we will now also review feedback approaches aimed at reducing bias in interval estimation, more specifically, overprecision, which besides overestimation and overplacement is one of the three types of overconfidence. Overprecision refers to being too certain that one’s estimate is more accurate than it actually is [33], and awareness of overprecision is therefore key to accepting the feedback.

Amongst others, Klayman, Soll, Gonzalez-Vallejo, and Barlas [34] introduce a common method for interval estimation. Participants must give a numerical estimate as well as a 90 % confidence interval, referring to an upper and lower bound where the probability is 90 % that the true answer lies inside the interval.

While normatively the correct answer should lie in arond 90 % of estimates within the 90 % confidence range indicated by a participant, in their study, the correct answer fell inside the participants’ confidence ranges in less than 50 % of the time. In their paper, the authors also refer to previous experimental studies that have shown significant overprecision when participants were asked to provide confidence ranges.

Soll and Klayman [35] also show that, although unsystematic judgmental error may also contribute to overconfidence, subjective confidence intervals are systematically much too narrow. These and related studies signify that a person is regularly overly self-assured of their judgmental accuracy.

A decline of overprecision might be accomplished by feedback on previous judgments provided after the first judgments made [36].

Despite not finding research focusing on the selective application of feedback as it is the subject in our setting, the results concerning feedback on performance and corrective capability demonstrate their potentials to improve judgments, which motivates the usage of feedback based on error patterns for our purposes.

2.3. Self-reflection

A prerequisite for the effectiveness of any feedback-DSS is the willingness of acceptance and therefore adoption of feedback by experts [37], which are often overconfident and neglect recommendations and feedback. This frequently holds true despite being told the opposite by a software, which may be due to the fact that it is external feedback [8]. To overcome this challenge, facilitation of a self-reflective process, that is the interpretation and assessment of own thoughts, emotions, and actions, is advised [37,38].

Research related to the reflection on feedback by experts is conducted by Goodwin [39] with an experiment in which forecasters are asked to review their judgmental predictions. Requiring forecasters to self-reflect by giving a statement for why they adjust the prediction the way they do, leads to higher performance and stronger accuracy improvement. Moreover, Sargeant, Mann, van der Vleuten, and Metsemakers [40] conduct interviews with physicians who evaluate assessment feedback they receive, showing that reflection is valuable referring to the manner of feedback application. This is in line with the research by Haddara and Rahnev [41], who find that feedback does not affect behavior through automatic reinforcement mechanisms but by giving users the opportunity to adjust their strategy.

Overall, prior research suggests counteracting the false-correction issue by feeding back error patterns and biases learned from previous own error patterns. Regarding the type of feedback, the literature suggests the combination of personalized and performance related OFB with CFB, where it is advisable to also use this type of feedback to reduce overprecision and therewith promote a wise and selective feedback acceptance and incorporation. Furthermore, mirroring an individual bias boosts self-reflection and learning.

Overall, various endeavors regarding bias reduction with the support of feedback exist. However, despite its significance for several areas of business, so far it has not been investigated whether a DSS integrating statistical error pattern feedback to reduce bias and enhance accuracy of point estimation judgments by a wise and selective application of the feedback can improve accuracy, and it is unclear how such a system shall be designed.

We now propose the anatomy of a DSS aimed at collaborative intelligence operationalizing these conclusions and encouraging an expert's differential confidence in decision or judgmental tasks. Then, we present the design of an experiment to study the efficacy of this DSS type.

3. DSS infrastructure and procedure

We now describe the infrastructure and procedure of the DSS we propose. The steps followed with the DSS are depicted in Fig. 1.

First, an expert's judgments, which might relate to different tasks, or question categories, are gathered via a user interface and stored in the database. That database also contains a library with potential biases and typical error patterns that suggest them. Based on judgments given by an expert and the observed errors, an inference engine might detect such a typical error pattern, positively evaluate a rule indicating a potential bias at play, and feedback the detected error pattern and/or the presumed bias to the expert, potentially including further explanations.

This feedback can be of any type, depending on the kind of bias that is aimed to be detected and mitigated, but must be based on the expert's personal prior judgment errors. As an example, in our first experiment, meant to be a basic implementation for a proof of concept, the personal error pattern of the subject is represented by the mean percentage error (MPE) computed over their past judgments. We note that this exemplary statistical model can be replaced by other models later on to detect other and more complex biases.

After the feedback, the experts make new judgments, i.e., they answer novel questions that may relate to the same tasks or categories of questions. The feedback intends to make aware and confront the expert with their own personal potential error pattern to stimulate self-reflection on previously given judgments and drive thoughtfulness on whether and how to apply the feedback on the next judgments.

After potentially recognizing certain categories or domains the questions might relate to, the expert may, for example, also apply

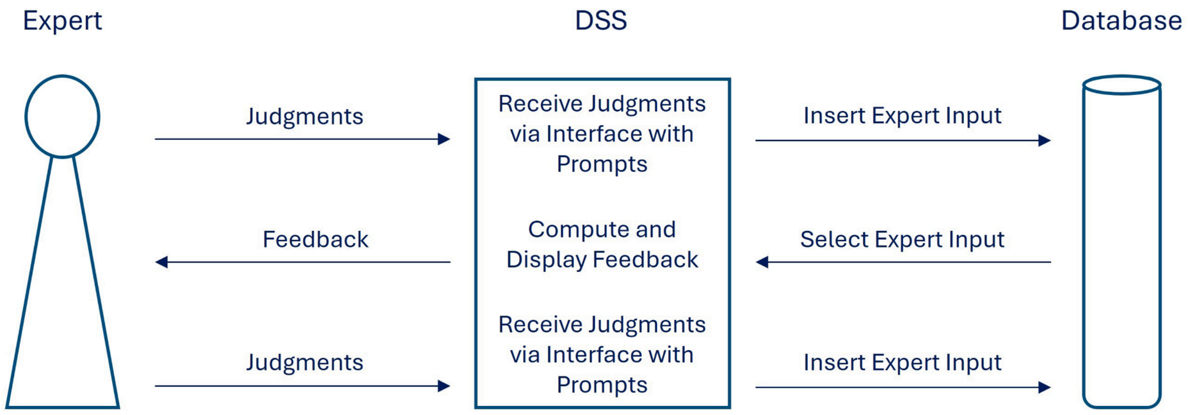
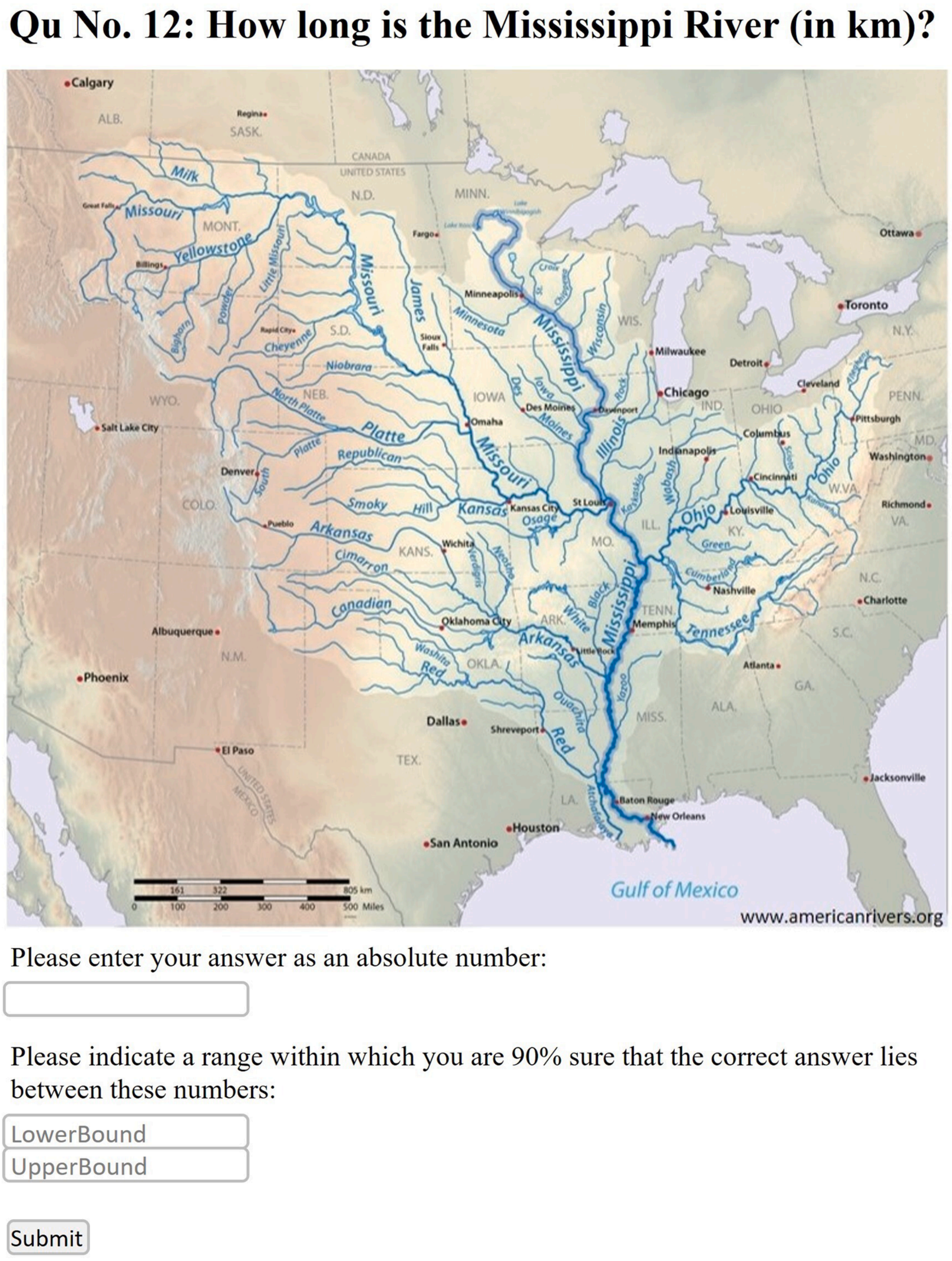

Expert
DSS
Database
Judgments
Receive Judgments via Interface with Prompts
Insert Expert Input
Feedback
Compute and Display Feedback
Select Expert Input
Judgments
Receive Judgments via Interface with Prompts
Insert Expert Input

Fig. 1. Infrastructure and Processes of DSS.

the feedback differently to different tasks or categories although no tasks or categories are communicated. Thus, an expert may make use of rather generic feedback, computed over all of their past judgments, in a wise manner to make novel judgments. Specifically, an expert may have tasks or a categorization in mind, which she or he would group the questions into, and may have category-specific confidence in her or his judgmental ability, leading to incorporating the feedback for certain questions and neglecting the feedback for others. It is important to note that an expert may have a personal categorization scheme in mind, which may be very different to the



Qu No. 12: How long is the Mississippi River (in km)?

Please enter your answer as an absolute number:

Please indicate a range within which you are 90% sure that the correct answer lies between these numbers:

LowerBound

UpperBound

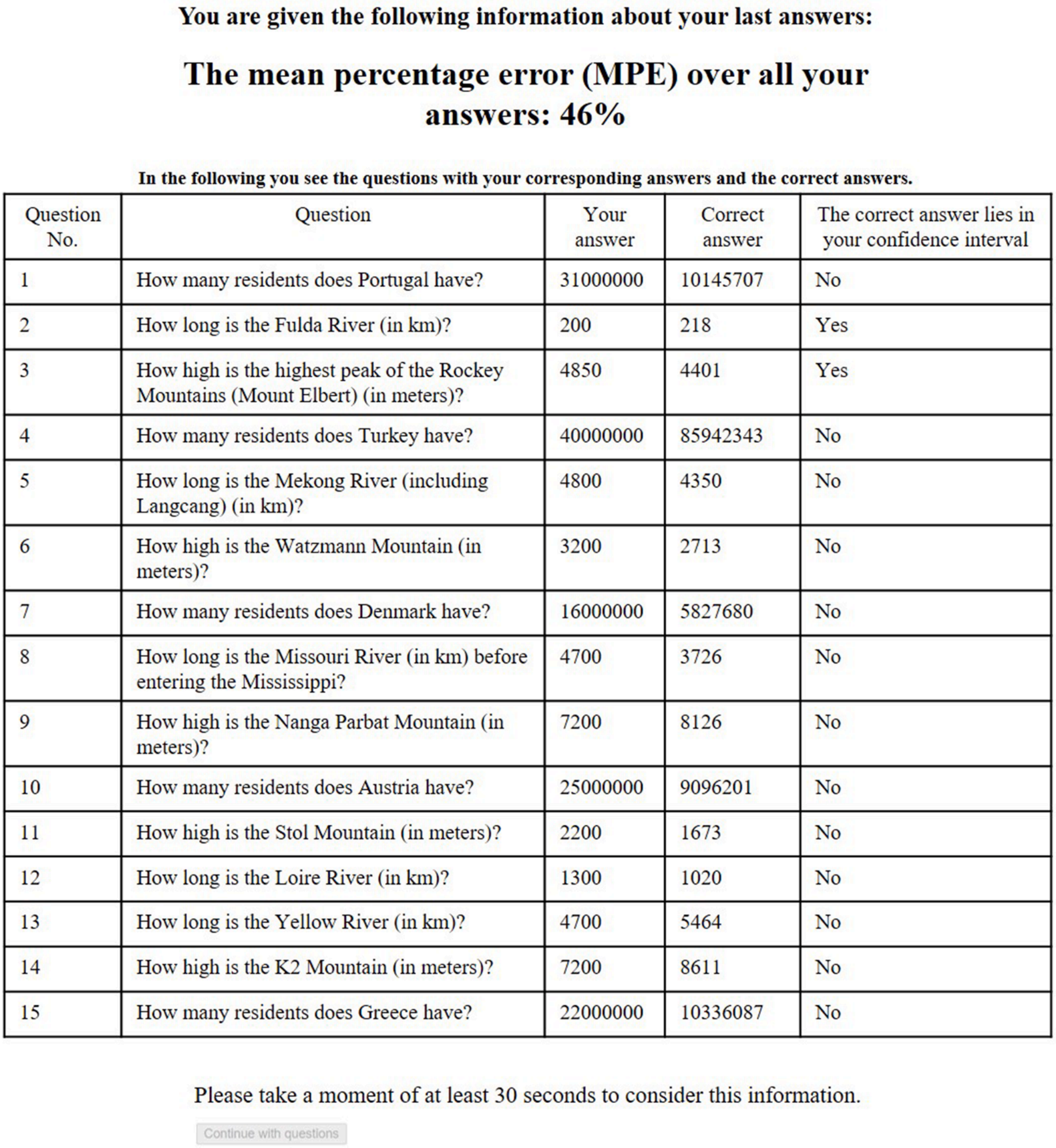
Fig. 2. DSS interface – User Prompt Example.

one of a statistical model, motivating the mirroring of more aggregated feedback. An example is provided in the next section that presents the experimental design.

After another set of judgments is captured, feedback might again be computed and mirrored to allow for adjusting debiasing over time.

Regarding the technical part, the DSS has been developed with Dynamic HTML (PHP) as frontend, a Relational Database Management Server (MySQL) as backend storing parameters for the DSS as well as questions and answers given by experts. The DSS is designed as Web-Application so it can be used on every computer. The tools used for determining error patterns, displaying them as feedback, and computing loss functions are developed in PHP and R.

In this section, we introduced the procedure, infrastructure and intention of the DSS, and we will now describe its implementation for our experiment.



You are given the following information about your last answers:

The mean percentage error (MPE) over all your answers: 46%

In the following you see the questions with your corresponding answers and the correct answers.

Question No.	Question	Your answer	Correct answer	The correct answer lies in your confidence interval
1	How many residents does Portugal have?	31000000	10145707	No
2	How long is the Fulda River (in km)?	200	218	Yes
3	How high is the highest peak of the Rockey Mountains (Mount Elbert) (in meters)?	4850	4401	Yes
4	How many residents does Turkey have?	40000000	85942343	No
5	How long is the Mekong River (including Langcang) (in km)?	4800	4350	No
6	How high is the Watzmann Mountain (in meters)?	3200	2713	No
7	How many residents does Denmark have?	16000000	5827680	No
8	How long is the Missouri River (in km) before entering the Mississippi?	4700	3726	No
9	How high is the Nanga Parbat Mountain (in meters)?	7200	8126	No
10	How many residents does Austria have?	25000000	9096201	No
11	How high is the Stol Mountain (in meters)?	2200	1673	No
12	How long is the Loire River (in km)?	1300	1020	No
13	How long is the Yellow River (in km)?	4700	5464	No
14	How high is the K2 Mountain (in meters)?	7200	8611	No
15	How many residents does Greece have?	22000000	10336087	No

Please take a moment of at least 30 seconds to consider this information.

Fig. 3. DSS Interface – Feedback Page Example.

4. Experimental research design and hypotheses

First, the research design of the experiment is described, whereby information thereon has partly been provided in the previous section to make the idea and concept of the DSS more tangible. Second, the hypotheses and corresponding measures for analysis are presented.

4.1. Research design

Our gender-balanced sample of subjects consists of 97 university students from different fields between 18 and 31 years old. They are randomly assigned to treatment (51 subjects) and control group (46 subjects).

The configuration of the experiment is stored in a database and includes the estimation questions, the correct answers and visual cues, the form and timing of feedback, the pages for briefing and debriefing, comprehension questions, rules when an experiment terminates, and a final questionnaire.

After comprehension questions are answered, subjects are required to answer point estimate questions from general knowledge categories. In addition, subjects are asked to indicate a 90 % certainty interval for every question. The categories contain questions about number of residents of a country, river length, and mountain height. Example questions are: "How many residents does France have?", "How long is the Hudson River (in km)?", "How high is the Mount Everest (in meters)?". Categories are not communicated to subjects but arguably easy to anticipate by humans.

This scenario intends to simulate expert judgment by supposing experts have expertise and basic confidence in all domains they are responsible for, like the subjects in the experiment who most likely have basic knowledge of the general knowledge questions. The categories in the experiment are meant to mimic different domains, where experts as well as the subjects may apply different heuristics and perform better in some than in others. Due to these categories or types of questions, wherein proneness to biases is assumed, humans may recognize patterns that a machine would not be able to detect. Subjects may also have a different categorization in mind such as regions or continents of the world, in case a subject realizes she or he is more adept at questions regarding, for example, Europe compared to Asia.

Having the categories in mind, a mountaineer will have great knowledge of mountains and will probably make quite accurate estimations in this category even before the feedback and being self-aware of this knowledge, may not apply the feedback or may apply it less pronounced to subsequent mountain height questions compared to questions in other categories.

With every question a visual cue for estimation support and to reduce error variance is displayed. This again mimics the environment of experts, which are also supported by orientational data, figures or graphs when making decisions.

For estimations of residents in a country, a map of the country including the ten largest cities with an indication of a range of their size is presented. For river lengths, a map of the respective river with a scale in the legend and for mountain heights, a topographical map with a reference mountain height is shown. An exemplary question is depicted in Fig. 2. Here, the subject is prompted to estimate the number of kilometers of the Mississippi River as well as a range within which he or she is 90 % sure that the correct answer lies inside by indicating an upper and lower bound. For estimation support, a map of the Mississippi including a scale is displayed with the question.

In total, there are 30 questions divided into two sequences with an interruption after 15 questions, where the treatment group receives feedback and the control group a blank page inviting to take a break. After the interruption, subjects receive another 15 new and unseen questions. Fig. 3 shows an exemplary feedback page with the mean bias of the subject.

The feedback comprises a subject's own MPE computed across its first 15 point estimates as well as information per question on its given answer, the actually correct answer, and if the interval includes the correct answer.

The MPE is computed by taking the difference between the provided and correct answer per question, dividing this difference by the respective correct answer, multiplying it by 100, and taking the mean of this over all the previous answers and thus over all categories.

We choose the MPE for the first experiment as it is a comprehensible statistic and theoretically simple in application even though it is not trivial for subjects to apply the feedback correctly to unseen questions. For instance, an MPE of 50 % means that given answers exceed correct answers by 50 % on average and application thereof would mean to take 2/3 of an upcoming estimate. The information on the individual answers indicates which categories push the MPE, or where over- or underestimation is discovered to encourage reflection on the manner of adapting future estimations. As the feedback on intervals is meant to give guidance if their intervals were set too narrow, possibly due to being overconfident in their answers, this may lead to decreasing overconfidence and better self-reflection. We note that feedback is strictly related to patterns in a subject's own error history.

For performance determination we calculate the mean absolute percentage error (MAPE) per subject, which is similar to the MPE, with the difference that absolute values of the percentage errors are averaged so that errors cannot balance each other out. Therefore, we can identify if a subject improves or deteriorates in performance after the feedback or the blank page, which is also important for payouts. In a briefing prior to the experiment, subjects receive guidance on how the MPE can be interpreted (without telling them they will receive feedback) including information on the MAPE performance measure and its impact on payouts.

After the second sequence of questions, all subjects receive the above-described feedback. The experiment ends with feedback from subjects and demographic questions.

Every subject receives a payout for participation and can win an additional prize money per group depending on their MAPE performance. The lower the MAPE, the higher is the probability to win a prize, which is meant to incentivize subjects. The experimental procedure is illustrated in Fig. 4.

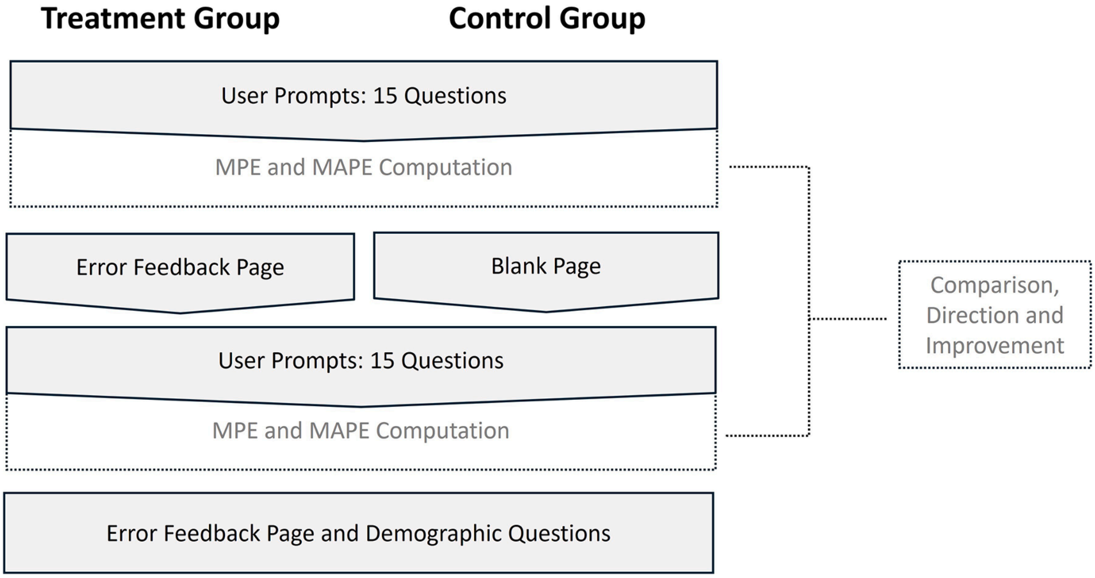

Treatment Group
Control Group
User Prompts: 15 Questions
MPE and MAPE Computation
Error Feedback Page
Blank Page
User Prompts: 15 Questions
MPE and MAPE Computation
Error Feedback Page and Demographic Questions
Comparison, Direction and Improvement

Fig. 4. Experimental Design and Procedure.

As described above, the intuition of using question categories is to imitate experts' working environment. This includes: first, the assumption of a sound general knowledge in their field of expertise with emphasized skills in some subfields; second, different estimation heuristics in different subfields; and third, subjects having category-specific knowledge, judgmental ability and error levels, promoting a category-specific consideration of error-feedback.

Subsequent to the conduct of the experiment, we undertake a Winsorization on the APE values before further analysis of results, to prevent potentially strong outliers distorting the results. We will set the low border as the 5 %- quantile and the high border as the 95 %-quantile of the APE values.

We mention that this research design/ experiment is meant to be the first out of a series of basic scenarios that will be described in Section 7.

4.2. Hypotheses

We now formulate seven hypotheses (H1-H7) for the key assumption that humans are capable of recognizing different structures and reflecting on own error patterns to selectively apply feedback to reduce (point estimation) error and bias.

H1 and H2 relate to adjustment behavior in the right direction after the feedback. This is examined overall as well as category-specific, meaning if right-direction MPE change is emphasized in categories where MAPE is higher before the feedback. H3 and H4 refer to accuracy enhancement after the feedback. This is again analyzed in total and category specific. H5 concerns auto-correction versus human-correction with feedback. H6 and H7 relate to the certainty interval estimation and the reduction of overprecision by examining if and how subjects broaden their certainty interval after the feedback, again once in general and once category specific.

H1. *MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction.*

For H1, the MPE is determined per subject as described in Section 4.1 for the first and second sequence, meaning before and after the feedback or blank page. An example for the meaning of right direction is, if a subject has an MPE of −30 % in the first sequence, suggesting underestimation, and an MPE above −30 % in the second sequence, this hints to an acceptance and incorporation of the feedback to counteract underestimation.

The expectation is that the ratio of right-direction MPE changes in the treatment group exceeds the ratio in the control group and the control group reaches a ratio around 50 % as no feedback is given to possibly cause systematic MPE change. To find significance for H1, a Fisher's exact test of independence between the ratios of right-direction MPE changes in treatment and control group is conducted.

H2. *MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE, resulting in larger MPE change in this category.*

To test H2, the MPE per subject and category for the first and second sequence and the MAPE for the first sequence is computed. Then, the category with the highest MAPE in the first sequence and the category with the highest right direction MPE change in the second sequence is identified, again per subject.

If these categories between first and second sequence match, it indicates that the subject adjusts their estimations in the right direction the most in that category where it is most necessary. This would indicate that subjects do not blindly adopt the feedback and apply it to all questions in the second sequence but use it selectively.

The ratios of category matches are compared between treatment and control group, expecting the treatment group to achieve a higher ratio. If the categories match in more than 1/3 of cases (the baseline in case of randomness) in the treatment group, we assume category specific feedback application. For H2 we conduct a Fisher's exact test between the ratios of category matches in the treatment and control group to detect the significance of ratio differences.

H3. *MPE-feedback leads to more MAPE reductions compared to no feedback.*

Per subject the MAPE is computed for the first and second sequence of estimations. Then, it is analyzed per subject if an increase or a decrease in MAPE between the sequences can be observed. The ratio of MAPE reductions between treatment and control group are compared, where we expect the treatment group to exhibit higher MAPE reduction than the control group. Due to absent feedback and hence random MAPE increases or decreases, we expect a ratio around 50 % in the control group. A Fisher's exact test is conducted to verify a significant difference between treatment and control group results.

We note that the difference between H3 and H1 is that H1 observes the changes of MPE direction, whereas H3 deals with changes of MAPE as accuracy measure. It is possible that the MPE of a subject changes in the right direction without the MAPE being improved when adjusting the estimation too strong. In this case the feedback might have been adopted but applied too intensely.

H4. *MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category.*

Per subject and category, we compute the MAPE for the first and second sequence. Subsequently, we determine the category with the highest MAPE in the first sequence and the category with the highest MAPE decrease between the first and second sequence. Then, we count the number of these category matches for both groups. If the treatment group achieves a higher ratio of category matches compared to the control group and the former reveals category matches in more than 1/3 of cases, we assume a general ability of beneficial selective feedback application. This is because there are three categories, for which reason the probability of one category match is 1/3 when assuming randomness. Again, we use the Fisher's exact test to find significance in the difference of results between treatment and control group.

H5. *MPE-feedback leads to more MAPE reductions compared to coarse-grained auto-correction.* For H5 we compare the MAPE improvement between the auto-corrected answers in the second sequence and the answers given by subjects in the second sequence in the treatment group. The hypothetical answers the auto-correction would provide in the second sequence are computed from answers of the control group of the second sequence. This hypothetical answer per question is determined by taking the subject's answer and dividing it by the corresponding MPE+1 observed in the first sequence of questions (the assumed error pattern a.k.a. mean bias of a subject).

This simulates a statistical method that applies the error pattern across all future judgments the subject makes without differentiating between questions. Then, we compute the MAPE of these answers in the second sequence and evaluate per subject if there is an improvement from the first to the second sequence. This ratio of MAPE reductions of auto-corrected judgments in the control group is compared to the human corrected judgments in the treatment group (from H3). As we assume the humans to apply the feedback wisely, and category-specific, while the statistical model applies the correction in a broad, category-agnostic fashion, we expect the ratio of MAPE reductions in the treatment group to exceed the ratio of MAPE reductions by auto-correction in the control group and conduct a Fisher's exact test to find significance between human-corrected results and auto-correction.

As discussed, in the scenario considered in our experiment, it is intentionally rather straight-forward to determine the categories and apply the feedback by adjusting judgment heuristics category-wise (although, as mentioned before, participants might have different categories than mountain heights, population sizes and river lengths in mind, e.g. questions related to Europe vs. Asia vs. America etc.).

As, arguably, in this scenario a machine learning or statistical model might also be capable to detect the categories – or, more precisely, might figure out that the participants are likely to have these categories in mind and exhibit category-specific biases – we additionally analyze how auto-correction would perform if the machine would correct the answers category-specifically. Therefore, we compute the MPE per category per subject in the control group and correct the answers in the second sequence with the respective category-specific MPE (we note that category-specific MPEs are then computed over only around five answers (instead of 15 answers), which might impact the reliability of the MPE estimates). Of these auto-corrected answers we calculate the MAPE per subject and compare it to the MAPE of the first sequence.

H6. *The certainty intervals become broader after the feedback, if they were too narrow to include the correct answer before the feedback, more often in the treatment than in the control group.*

Regarding H6, the assumption is that if subjects set the upper and lower bound of their 90 % certainty interval too narrow so that the correct answer does not lie in-between them in approximately 90 % of cases, they are overprecise.

To test H6, per subject the relative frequency of correct answers being inside the subject's interval in the first sequence is determined. As from results of previous experiments it can be expected that for most participants the percentage of correct answers lying inside the interval will be much below 90 % (as humans tend to be overprecise), and to also account for randomness, we define a threshold below 90 % for the percentage the correct answer must lie inside the interval to distinguish probably non or moderately overprecise participants from participants apparently more prone to overprecision. Concretely, we define a conservative threshold of 50 % (and repeat the analysis with an even more conservative threshold of 35 % for reasons of robustness). If the threshold percentage is surpassed by a subject, we consider the average interval of the subject as clearly set too narrow and the subject overprecise.

Per overprecise subject, his or her average relative size difference of the intervals between first and second sequence is computed. We normalize interval size by dividing each difference between upper and lower bound by the provided point estimate to make the interval sizes comparable. We anticipate the percentage of normalized interval size increases from the first to the second sequence be higher in the treatment group than in the control group.

To test H6, we again conduct a Fisher's exact test to quantify the significance of the results.

H7. *The certainty intervals become broader more often especially in those categories, in which the intervals were too narrow to include the correct answer before the feedback compared to no feedback given.*

For H7, per subject and per category the relative frequency of correct answers lying outside the interval as well as the average relative size difference of the intervals between first and second sequence including normalization of interval size as above is computed. Then, the category with the maximum relative frequency of correct answers lying outside the interval and the category with the maximum average normalized size increase of intervals is identified. We expect this relative frequency to be higher for the treatment compared to the control group. We conduct a Fisher's exact test for significance detection.

In total, with the hypotheses we investigate the (selective) change in judgment through personal error feedback and whether this results in presumable bias reduction and accuracy improvement. Additionally, we examine whether humans can reduce the MAPE stronger than coarse-grained auto-correction.

5. Results

In this section, results per hypothesis are summarized.[1]

H1. *MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction.*

In 82.4 % of cases the MPE values of subjects in the treatment group changed in the right direction after the feedback. For subjects in the control group, we observe MPE changes in the right direction in 54.3 % of cases. The p-value of the Fisher's exact test is 0.0027, which demonstrates significance of the difference between treatment and control group at a 5 % significance level.

H2. *MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE, resulting in larger MPE change in this category.* In the treatment group, 62.7 % of the subjects had the greatest MPE change in the right direction in that category where the MAPE was the highest in the first sequence. This was the case for 43.5 % in the control group. The p-value for the Fisher's exact test is 0.0447, which means the difference in results is significant at a 5 % significance level.

H3. *MPE-feedback leads to more MAPE reductions compared to no feedback.* For 64.7 % of subjects in the treatment group there was a MAPE reduction after the feedback compared to 52.2 % in the control group after the blank page. For H3 we could not determine a significance at a 5 % significance level with a p-value of 0.1479 of the Fisher's exact test.

H4. *MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category.*

64.7 % of subjects in the treatment group made the largest MAPE improvement in the second sequence after the feedback in that category with the highest MAPE in the first sequence. This was the case for 52.2 % in the control group after the blank page. The p-value for the Fisher's exact test is 0.1479, for which reason we cannot state significance at a level of 5 %.

We note that the results of H3 and H4 look alike as the majority of subjects who reduced their MAPE in general also reduced their MAPE selectively per category. This seems reasonable as the category-specific application of feedback leads to a reduction of MAPE in total.

H5. *MPE-feedback leads to more MAPE reductions compared to coarse-grained auto-correction.*

Compared to the first sequence of the control group the auto-correction yielded a MAPE improvement in 26.1 % of cases in the second sequence. In the treatment group, 64.7 % of subjects improved their MAPE after the feedback. A p-value of 0.000132 of the Fisher's exact test indicates significance at a 5 % level for the stronger MAPE decline through human correction with feedback than auto-correction.

Regarding the category-specific auto-correction, there was a MAPE improvement compared to the first sequence in 50.0 % of cases. This result improved heavily compared to non-category-specific auto-correction (26.1 %), while it is still inferior to the human-corrected results with feedback (64.7 %).

H6. *The certainty intervals become broader after the feedback, if they were too narrow to include the correct answer before the feedback, more often in the treatment than in the control group.*

Applying a 50 % threshold for a participant to be considered as (strongly) overprecise, meaning that the correct answer must lie

[1] Of the 97 subjects in the experiment, seven subjects missed one question (due to technical problems), for which reason one answer is missing for each of these seven subjects. The remaining answers of these affected are included in the analyses.

inside the given interval in under 50 %, involves 41 subjects in the treatment group and 35 subjects in the control group. Of these 41 subjects in the treatment group, 68.3 % broadened their normalized certainty interval after the feedback in the second sequence. Of the 35 subjects in the control group 34.3 % broadened their normalized certainty interval after the blank page in the second sequence. This result is significant at a 5 % level with a p-value of 0.0030 for the Fisher's exact test.

The 35 % threshold includes 34 subjects in the treatment and 27 in the control group, where 70.6 % of the treatment group broadened their certainty interval vs. 29.6 % of the control group. This result is also significant as shown by the p-value of 0.0016 of the Fisher's exact test.

H7. *The certainty intervals become broader more often especially in those categories, in which the intervals were too narrow to include the correct answer before the feedback compared to no feedback given.*

In the treatment group 60.8 % of the subjects broadened their interval after the feedback in the second sequence in that category in which the correct answers lay outside the interval the most before the feedback in the first sequence. This was the case for 50.0 % in the control group. The p-value for the Fisher's exact test is 0.1941 for which reason the difference in results between the groups is not significant at a 5 % level.

Summarizing, subjects that received feedback achieved higher error reductions than subjects without feedback. This holds true in general as well as for the category specific application. Moreover, subjects in the treatment group were able to achieve higher accuracy after the feedback compared to a statistical method using the MPE for auto-correction (with or without category-awareness), indicating potential to mitigate the false-correction problem inherent with auto-correction. In addition, subjects receiving feedback exhibit larger decreases in overprecision than subjects not receiving feedback, overall and category specific.

6. Discussion

In this section, first we discuss the results in general and, second, the subjects' answers to the usage of feedback during the experiment.

6.1. Discussion of results

The results presented above show that four out of seven (sub-) hypotheses are significant at a 5 % level. For the other three of the seven (sub-) hypotheses, where the results are not significant to a 5 % level, the obtained results point into directions supporting the underlying, general assumption that prospects are capable to utilize the feedback wisely and selectively to improve their judgmental performance compared to no feedback or auto-correction.

Overall, we observe support for the key hypothesis of wise and selective consideration and application of feedback based on one's own error pattern, leading to bias reduction and accuracy improvement.

More specifically, we detect a higher proportion of subjects in the treatment compared to the control group matching the categories between highest MAPE in the first sequence and strongest MPE change in the right direction as well as the strongest MAPE reduction in the second sequence. This underpins the ability of humans to reflect on the own error feedback and use it selectively for further judgments. Therefore, errors are reduced the most where they are the largest.

The findings of H5 indicate benefits and human skills to use feedback wisely in a way that can lead to lower error than auto-correction. Even when the machine would know the categories, the subjects still performed better, i.e. improved their MAPE more. Overall, the combination of the machine providing feedback and the human applying it appears to be a promising approach of collaborative intelligence for accuracy improvement.

The results of H6 and H7 then indicate that subjects receiving feedback show a lower degree of overprecision in the second sequence than subjects not receiving feedback. This finding also indicates that subjects reflect on their own errors and are willing and able to adapt judgments or judgmental heuristics to reduce overprecision.

An interesting observation is that more subjects in the treatment group were able to adjust their MPE in the right direction (H1: 82.4 %) than reduce their MAPE (H3: 64.7 %) after the feedback. That means, 21.6 % of subjects in the treatment group adjusted their MPE in the right direction without improving their MAPE. One likely explanation is that these subjects adjusted their estimations with the right intention, but too excessively such that their MAPE increased after the feedback. This is confirmed by inspecting the individual answers of the subjects. In 45.5 % out of the 21.6 % observations, the subjects' errors indicate a strong over- or underestimation in the category of resident numbers before the feedback and then surpass the optimal level of adjustment after the feedback so that other categories where the MAPE has decreased could not balance this out. This raises the question how the feedback could be modified to mitigate the problem of over-adjustment beyond beneficial levels, for instance by sensitizing subjects for the magnitudes of adjustment and the risk of potential over-steering.

To provide additional support for our findings, we conducted an additional experiment with different question categories. The results of this experiment are comparable to the results obtained in the experiment described in this paper and therefore underline our hypotheses. The details thereof are described in the appendix.

6.2. Subjects' answers to the reception and utilization of the feedback

The outcomes of the experiment generally mirror the subjects' intentions and thoughts during the experiment. At the end of an experiment, each subject in the treatment group was asked the following question: "During the experiment, how did you use the information of the feedback? If you did not use it, why not?". 25.1 % of subjects explicitly stated to have adjusted their estimation in a certain direction due to their recognition of over- or underestimation after the feedback. 30.8 % even claim to have adjusted their estimation in a selective, category-specific manner. Furthermore, 34.4 % of subjects indicate to have broadened their certainty intervals after the feedback.

These declarations show that and how subjects reflected on the feedback and how they intended to apply it for debiasing. In the following, selected quotes of subjects that were given to the question above are presented.

> *"Calculating the 'measured' length of rivers times 2 instead of rounding up (because they are not straight), setting the confidence intervall wider"*

> *"I saw that especially concerning the resident number, I guessed too high, so I tried to reduce it. Also, I saw that is is better to have a higher answer confidence range so that the correct answer ist in the range."*

> *"I looked how much in general I was off with my numbers. For example I doubled most of my numbers after the feedback in the county category and I added a 0 on the length of all the following river numbers I thought were right."*

Hence, we assume that the feedback stimulates reflection on own error patterns and wise and systematic application on further judgments.

7. Conclusion and managerial implications

This article aims at demonstrating that humans can recognize structures (i.e. categories), reflect on feedback based on own errors, and use it systematically and selectively to achieve bias and error reduction. It also shows that humans can achieve stronger error reduction using the feedback compared to a machine applying (straight-forward) auto-correction of assumed error patterns.

Our results are in line with the findings in previous work. Blanc and Setzer [13] detect systematic error patterns, which we do as well, and which we transform into feedback. We show that performance feedback leads to accuracy improvement as Benson and Önkal [28] found out, however, they do not consider whether feedback is applied selectively. In addition, our results are in line with those of Russo and Schoemaker [36], where overprecision declines with feedback on previous judgments. Finally, the idea that stimulating self-reflection leads to better performance is also shown by Goodwin [39].

Nevertheless, to our knowledge, this work is the first to investigate whether and how a DSS integrating statistical error feedback can reduce bias and enhance accuracy of point estimations by wise and selective consideration of feedback, and to test a respective system experimentally.

Overall, this work demonstrates that the proposed debiasing approach has potential, while primary limitations of the current stage of this research are related to the transferability to other, more complex biases and the transferability to practice.

To address transferability, we plan to explore two more scenarios in upcoming experiments. In the first scenario considered in this article, latent topics (categories) can be considered easily identifiable by humans, while the machine is assumed to be unaware of the categories and can therefore only provide aggregated feedback (although we also tested a machine that is aware of the categories).

In a second scenario, we will consider situations where the latent topics are communicated to subjects and the machine that can then search for category-specific error patterns, give category-specific feedback and can also apply category-aware auto-correction. In case human biases are indeed category-specific, the subjects, when applying the feedback, as well as the auto-correction performance of the machine, may benefit from this information. However, this requires a larger set of provided answers, such that the machine has sufficient training data to reliably learn category-specific error patterns, ideally allowing for cross-validation or other techniques to regularize correction estimates. Another option might be that a machine considers also, to some extent, errors in other categories.

The third scenario will cover situations with high complexity for both human and machine, containing questions that cannot be clearly assigned to a category and will have a rather vague reference to each other so that categories are not obvious. While the machine might nevertheless be able to identify latent categories (topics) based on textual analysis, specifically topic modeling, and use this for category-specific auto-correction, a subject may have no or very different categories in mind.

On the one hand, this can limit the applicability of the differentiated feedback such that aggregated feedback would be the better suited option, where a subject may still be able to apply the general feedback selectively based on domain knowledge or (implicit) categorization. On the other hand, this might support reflection of the machine-learned categories and potential biases used in one of the (previously unknown) categories that may exist. Overall, the intention of the scenarios is to understand the situations and conditions in which feedback of which type can be expected to be beneficial.

Managerial considerations and implications are manyfold. To allow practitioners to use such a DSS for self-reflection, bias-awareness and adjustment of judgment heuristics, the feedback must be comprehensible and actionable but also have merit, i.e. not be systematically misleading. The scenario described in this article is, as a proof of concept, designed such that the feedback given can be considered highly comprehensible, intuitive and actionable. Results indicate that the feedback systematically helped the subjects to reflect and improve their judgmental decisions.

However, estimating error patterns from relatively small sets of observations (15 answers and errors in this case) generally entails high estimation uncertainty such that the reliability of the feedback is limited. For instance, one or two extreme errors in one direction might flip the sign of the MPE, leading to the opposite recommendation derived from MPE feedback. For reasons of feedback reliability, a larger set of training error observations is desirable. In addition, more robust measures such as the median percentage error instead of the mean percentage error might be explored as a metric to derive more stable feedback.

For the estimation of more complex error patterns and biases, more parameters must be estimated, requiring an even larger set of available error observations (as mentioned above, more training error observations are also required when category-specific feedback should be given, as the number of training observations per category decreases with the number of categories).

Therefore, insufficient numbers of estimates provided by an expert limit the applicability of such an error feedback-based system. An approach to solve this problem is to ask experts to make additional estimates, which may not be mandatory in business but help to estimate stable error structures.

An appealing use case of such a DSS is also the provisioning of periodic feedback given immediately after novel judgments are provided and prompting an expert for a guided correction of the current estimate before its submission. For this to be successful, it must be considered that error patterns may change dynamically due to learning and heuristics adjustment, as feedback intends to mitigate the strength of error patterns. Hence, the machine must check whether the error observations prior to the last feedback is still valid. If so, the feedback has not been effective and it must be analyzed whether this stems from a lack of feedback acceptance, a wrong understanding or consideration of the feedback, or whether the feedback may be misleading in different respects, permitting a stronger correction of judgmental heuristics.

Over time, different inference engines (rules) to determine error pattern and presumable biases causing these pattern as well as different types of feedback (purely statistical and formal information, visualizations, bias explanations etc.) can be tested to gain knowledge on the most promising variations and in which directions those might be further improvable.

Prompting experts for comments how they understood and used the feedback to adjust heuristics might provide additional, highly useful information here. In case error patterns changed, sufficient data to re-learn error patterns must be collected to derive further meaningful feedback.

Aside from our described research setting, there are other opportunities to make use of and test the proposed DSS concept. For example, using probability estimation instead of point estimation, asking subjects to predict the probability of events. Here, the feedback for instance could contain the Brier Score to show the subjects how well their estimations are calibrated.

Also, additional measures can be deployed, such as sensors for eye tracking, pulse measurement, or voice when subjects enunciate their thoughts while making decisions. This can help to attain more insight into how humans think about and use machine feedback based on own errors.

CRediT authorship contribution statement

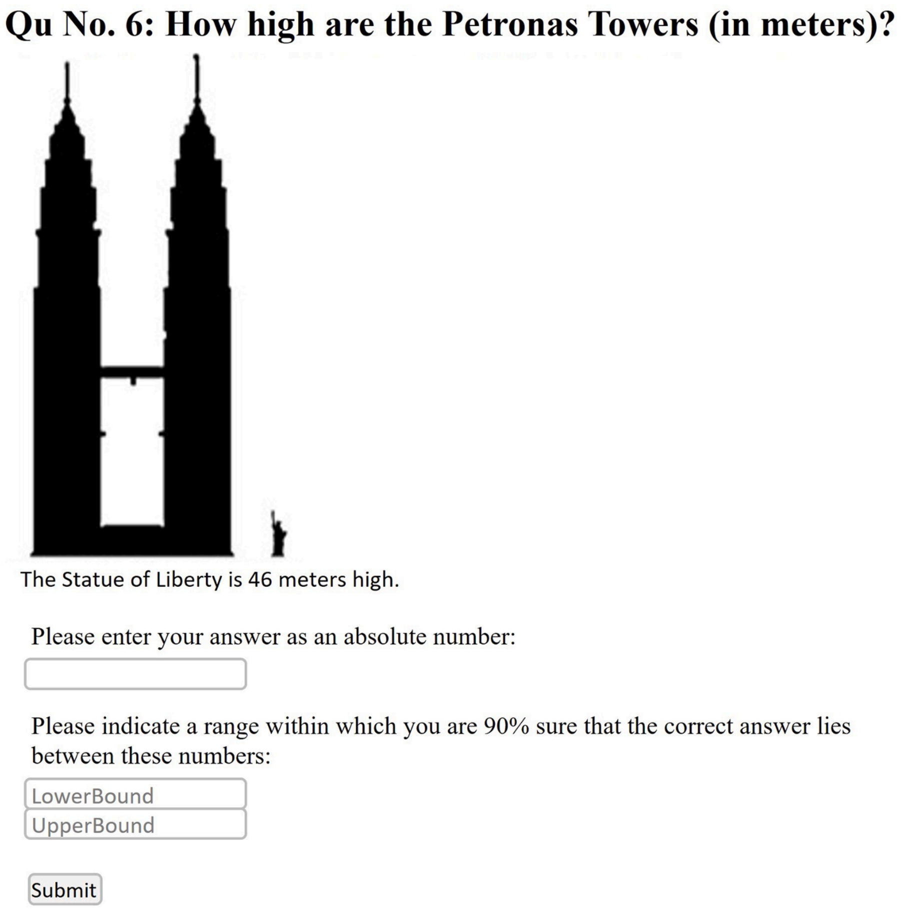
Nathalie Balla: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thomas Setzer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional Experiment

We conducted an additional experiment with different categories, where 32 subjects are randomly assigned to the treatment and 29 to the control group. The categories for this experiment are *beeline distances between cities world-wide, number of calories in a certain food*, and *heights of famous buildings*. The visual cues for these categories are respectively a map showing the distance between the two cities without a scale but a hint of the beeline distance between Berlin and Paris, a nutrition table excluding the calories, and a picture of the respective building next to the statue of liberty or a one family house with its height as a reference. An example of the third category is pictured in Fig. A.4.



Qu No. 6: How high are the Petronas Towers (in meters)?

The Statue of Liberty is 46 meters high.

Please enter your answer as an absolute number:

Please indicate a range within which you are 90% sure that the correct answer lies between these numbers:

Fig. A.4. DSS Interface – User Prompt Example with New Category.

We undertook the same analysis for this data and found support for the first five hypotheses.

For H1 90.6 % of subjects in the treatment group changed their MPE in the right direction versus 62.1 % in the control group with a p-value for the Fisher's exact test of 0.0089, showing significance.

Regarding H2, 71.9 % of the treatment group made the highest MPE adjustment in the right direction in the second sequence in that category where their MAPE was the highest in the first sequence compared to 48.3 % in the control group with a p-value of 0.05215, almost significant at a 5 % level. We found significant results for H3 with 75.0 % of subjects in the treatment group decreasing their MAPE after the feedback compared to 44.8 % in the control group with a p-value of 0.0156.

Referring to H4, 78.1 % of the treatment group reduced their MAPE the most in the second sequence in that category where their MAPE was the highest in the first sequence, whereas this was the case for 55.2 % of the control group with the difference in results almost being significant with a p-value of 0.0508.

For H5 the results were highly significant, where compared to the first sequence of the control group the auto-correction yielded a MAPE improvement in 48.3 %. This is compared to 75.0 % of the treatment group showing improvement of MAPE after the feedback. These results reveal a high difference with a p-value of 0.02928 for the Fisher's exact test.

The results for H6 and H7 are not very supportive and not significant.

In total, the results of this second experiment with the new categories underpin specifically H1-H5 of the first experiment and strengthen its generalizability.

Data availability

The authors do not have permission to share data.

References

[1] R.D. Klassen, B.E. Flores, Forecasting practices of Canadian firms: survey results and comparisons, Int. J. Prod. Econ. 70 (2001) 163–174, https://doi.org/10.1016/S0925-5273(00)00063-3.

[2] T.M. McCarthy, S.L. Golicic, J.T. Mentzer, The Evolution of Sales Forecasting Management: a 20-Year Longitudinal Study of Forecasting Practices, J. Forecast. 25 (2006) 303–324, https://doi.org/10.1002/for.989.

[3] N.R. Sanders, K.B. Manrodt, The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice, Omega (Westport) 31 (2003) 511–522, https://doi.org/10.1016/j.omega.2003.08.007.

[4] D. Arnott, S. Gao, Behavioral economics for decision support systems researchers, Decis. Support. Syst. 122 (2019) 113063, https://doi.org/10.1016/j.dss.2019.05.003.

[5] M. Lawrence, P. Goodwin, M. O'Connor, D. Önkal, Judgmental forecasting: a review of progress over the last 25years, Int. J. Forecast. 22 (2006) 493–518, https://doi.org/10.1016/j.ijfore- cast.2006.03.007.

[6] M. Lawrence, M. O'Connor, Scale, Variability, and the Calibration of Judgmental Prediction Intervals, Organ. Behav. Hum. Decis. Process. 56 (1993) 441–458, https://doi.org/10.1006/obhd.1993.1063.

[7] M. Lawrence, M. O'Connor, B. Edmundson, A field study of sales forecasting accuracy and processes, Eur. J. Oper. Res. 122 (2000) 151–160, https://doi.org/10.1016/S0377-2217(99)00085-5.

[8] J. Leitner, U. Leopold-Wildburger, Experiments on forecasting behavior with several sources of information – A review of the literature, Eur. J. Oper. Res. 213 (2011) 459–469, https://doi.org/10.1016/j.ejor.2011.01.006.

[9] J.S. Lim, M. O'Connor, Judgmental forecasting with interactive forecasting support systems, Decis. Support. Syst. 16 (1996) 339–357, https://doi.org/10.1016/0167-9236(95)00009-7.

[10] J.F. George, K. Duffy, M. Ahuja, Countering the anchoring and adjustment bias with decision support systems, Decis. Support. Syst. 29 (2000) 195–206, https://doi.org/10.1016/S0167-9236(00)00074-9.

[11] S. Blanc, T. Setzer, Analytical Debiasing of Corporate Cash Flow Forecasts, Eur. J. Oper. Res. 243 (2015) 1004–1015, https://doi.org/10.1016/j.ejor.2014.12.035.

[12] S. Blanc, T. Setzer, Improving Forecast Accuracy By Guided Manual Over- write in Forecast Debiasing, in: Twenty-Third European Conference on Information Systems (ECIS) 66, 2015.

[13] F.B. de Lima Neto, D.M. Lima Martins, G. Vossen, A semiotic-inspired machine for personalized multi-criteria intelligent decision support, Data Knowl. Eng. 117 (2018) 225–238, https://doi.org/10.1016/j.datak.2018.07.012.

[14] T. Haesevoets, D. De Cremer, K. Dierckx, A. Van Hiel, Human-machine collaboration in managerial decision making, Comput. Human. Behav. 119 (2021) 106730, https://doi.org/10.1016/j.chb.2021.106730.

[15] R. Pinto, T. Mettler, M. Taisch, Managing supplier delivery reliability risk under limited information: foundations for a human-in-the-loop DSS, Decis. Support. Syst. 54 (2013) 1076–1084, https://doi.org/10.1016/j.dss.2012.10.033.

[16] P.V. Deshmukh, A.K. Shahade, Elevating human-machine collaboration in NLP for enhanced content creation and decision support, Data Knowl. Eng. 161 (2026), https://doi.org/10.1016/j.datak.2025.102505.

[17] R.C. Blattberg, S.J. Hoch, Database Models and Managerial Intuition: 50% Model + 50% Manager, Manage Sci. 36 (1990) 887–899, https://doi.org/10.1287/mnsc.36.8.887.

[18] Y. Nagar, T. Malone, Making Business Predictions by Combining Human and Machine Intelligence in Prediction Markets, in: ICIS 2011 Proceedings 20, 2011.

[19] M. Arvan, B. Fahimnia, M. Reisi, E. Siemsen, Integrating Human Judgement into Quantitative Forecasting Methods: a Review, Omega (Westport) 86 (2019) 237–252, https://doi.org/10.1016/j.omega.2018.07.012.

[20] M. Zellner, A.E. Abbas, D.V. Budescu, A. Galstyan, A survey of human judgement and quantitative forecasting methods, R. Soc. Open. Sci. (2021), https://doi.org/10.1098/rsos.201187.

[21] M. Van den Broeke, S. De Baets, A. Vereecke, P. Baecke, K. Vanderheyden, Judgmental forecast adjustments over different time horizons, Omega (Westport) 87 (2019) 34–45, https://doi.org/10.1016/j.omega.2018.09.008.

[22] B.S. Wibowo, Y.J. Prakoso, N.A. Masruroh, Performance of judgmental-statistical forecast combinations under product-market configurations, Int. J. Manage. Sci. Eng. Manage. 18 (2023) 104–117, https://doi.org/10.1080/17509653.2021.2015472.

[23] H.Y. Kim, Y.S. Lee, D.B. Jun, The effect of relative performance feedback on judgmental forecasting accuracy, Manage. Decis. (2019) 1695–1711, https://doi.org/10.1108/md-06-2017-0549.

[24] P. Goodwin, Statistical Correction of Judgmental Point Forecasts and Decisions, Omega (Westport) 24 (1996) 551–559, https://doi.org/10.1016/0305-0483(96)00028-X.

[25] J. Jacoby, D. Mazursky, T. Troutman, A. Kuss, When Feedback is Ignored: disutility of Outcome Feedback, J. Appl. Psychol. 69 (1984) 531–545, https://doi.org/10.1037/0021-9010.69.3.531.

[26] W. Remus, M. O'Connor, K. Griggs, Does Feedback Improve the Accuracy of Recurrent Judgmental Forecasts? Organ. Behav. Hum. Decis. Process. 66 (1996) 22–30, https://doi.org/10.1006/obhd.1996.0035.

[27] W.K. Balzer, M.E. Doherty, R.Jr. O'Connor, Effects of Cognitive Feedback on Performance, Psychol. Bull. 106 (1989) 410–433, https://doi.org/10.1037/0033-2909.106.3.410.

[28] P.G. Benson, D. Önkal, The effects of feedback and training on the performance of probability forecasters, Int. J. Forecast. 8 (1992) 559–573, https://doi.org/10.1016/0169-2070(92)90066-I.

[29] K. Sengupta, Cognitive Feedback in Environments Characterized by Irrelevant Information, Omega (Westport) 23 (1995) 125–143, https://doi.org/10.1016/0305-0483(94)00061-E.

[30] A. Nussbaumer, K. Verbert, E. Hillemann, M.A. Bedek, D. Albert, A Framework for Cognitive Bias Detection and Feedback in a Visual Analytics Environment, in: 2016 European Intelligence and Security Informatics Conference (EISIC), 2016, pp. 148–151, https://doi.org/10.1109/EISIC.2016.038.

[31] N.E. Dunbar, M.L. Jensen, C.H. Miller, E. Bessarabova, Y. Lee, S.N. Wilson, J. Elizondo, B.J. Adame, J. Valacich, S. Straub, J.K. Burgoon, B. Lane, C.W. Piercy, D. Wilson, S. King, C. Vincent, R.M. Schuetzler, Mitigation of Cognitive Bias with a Serious Game: two Experiments Testing Feedback Timing and Source, Int. J. Game Based. Learn. 7 (2017) 86–100, https://doi.org/10.4018/IJGBL.2017100105.

[32] M. Król, M. Król, Learning From Peers' Eye Movements in the Absence of Expert Guidance: a Proof of Concept Using Laboratory Stock Trading, Eye Tracking, and Machine Learning, Cogn. Sci. 43 (2019) 1–32, https://doi.org/10.1111/cogs.12716.

[33] D.A. Moore, P.J. Healy, The Trouble With Overconfidence, Psychol. Rev. 115 (2008) 502–517, https://doi.org/10.1037/0033-295X.115.2.502.

[34] J. Klayman, J.B. Soll, C. Gonzalez-Vallejo, S. Barlas, Overconfidence: it Depends on How, What, and Whom You Ask, Organizational Behavior and Human Decision Processes. 79 (1999) 216–247. https://doi.org/10.1006/obhd.1999.2847.

[35] J.B. Soll, J. Klayman, Overconfidence in Interval Estimates, J. Experim. Psychol.: Learn., Memory, Cogn. 30 (2004) 299–314, https://doi.org/10.1037/0278-7393.30.2.299.

[36] J.E. Russo, P.H.J. Schoemaker, Managing Overconfidence, Sloan. Manage Rev. 33 (1992) 7–17.

[37] A. Grant, J. Franklin, P. Langford, The Self-Reflection and Insight Scale: a new Measure of Private Self-Consciousness, Soc. Behav. Pers. 30 (2002) 821–836, https://doi.org/10.2224/sbp.2002.30.8.821.

[38] L.F. Sasse-Werhahn, C. Bachmann, A. Habisch, Managing Tensions in Cor- porate Sustainability Through a Practical Wisdom Lens, J. Bus. Ethics 163 (2020) 53–66, https://doi.org/10.1007/s10551-018-3994-z.

[39] P. Goodwin, Improving the voluntary integration of statistical forecasts and judgment, Int. J. Forecast. 16 (2000) 85–99, https://doi.org/10.1016/S0169-2070(99)00026-6.

[40] J.M. Sargeant, K.V. Mann, C.P. van der Vleuten, J.F. Metsemakers, Reflection: a link between receiving and using assessment feedback, Adv. Health Sci. Educ. 16 (2009) 399–410, https://doi.org/10.1007/s10459-008-9124-4.

[41] N. Haddara, D. Rahnev, The Impact of Feedback on Perceptual Decision-Making and Metacognition: reduction in Bias but No Change in Sensitivity, Psychol. Sci. 33 (2022) 179–338, https://doi.org/10.1177/09567976211032887.

Dr. Nathalie Balla is a Business Intelligence Consultant at Oliva Advisory since October 2024. In her work she analyses data and makes it informative for business users that must make many different decisions based on this data. These decisions are the ones that are relevant to the experimental research conducted for this paper, respectively experts deciding in their domain. She was a research assistant and doctoral student at the chair of business informatics at the Catholic University of Eichstätt-Ingolstadt (KU) from October 2019 to October 2023. In her research, she considers decision support systems, cognitive biases, different forms of feedback to optimize decisions and the collaboration of human and machine, involving laboratory experiments at different universities. She has presented her research at high-ranked international conferences.

Prof. Dr. Thomas Setzer is, since 2018, a full professor of business informatics and information systems at the Catholic University of Eichstätt-Ingolstadt (KU), and since 2022 senior professor at the Mathematical Institute for Machine Learning and Data Science (MIDS). Until 2018 he was heading the Research Group “Corporate Services and Systems” at Karlsruhe Institute of Technology (KIT). In his research, he develops new models, methods, and systems to combine and support forecasts and judgments and published his results in several top-tier outlets. He is one of the winners of the INFORMS ISS design science award and headed, amongst several other transfer projects on decision support systems, a multi-year research project with Bayer AG on novel types of forecast support systems and debiasing techniques in corporate financial controlling.