Title

Advice Design to Increase the Use of Advice with an Interval to Overcome Algorithm Aversion

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Kagawa, Rina
Honda, Hidehito
Nosato, Hirokazu

Publication Date

2024

Peer reviewed

Advice Design to Increase the Use of Advice with an Interval to Overcome Algorithm Aversion

Rina Kagawa[1] (sonata.skazka@gmail.com)
Institute of Medicine, University of Tsukuba, 1-1-1, Tennoudai, Tsukuba-shi, Ibaraki, 305-8575, Japan

Hidehito Honda (hitohonda.02@gmail.com)
Faculty of Psychology, Otemon Gakuin University, 2-1-15, Nishiai, Ibaraki-shi, Osaka, 567-8502, Japan

Hirokazu Nosato (h.nosato@aist.go.jp)
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology
1-1-1, Umezono, Tsukuba-shi, Ibaraki, 305-8560, Japan

Abstract

Despite computational algorithms outperforming humans in certain tasks, algorithmic advice is less used than human advice (algorithm aversion). Thus, algorithmic advice should be designed to avoid algorithm aversion. However, few studies have discussed the use of advice with an interval (e.g., 60.0 ± 2.0 %), a common format in algorithmic advice. This study confirmed in two behavioral experiments (N = 200) that differences in advice sources lead to differences in advice use, mainly by influencing the step at which the judge decides whether to ignore the advice. Therefore, we proposed to individualize the presentation of advice so that the advice would be such that decreases the rate advice being ignored. Our individualization of the advice presentation focused on the distance between the advice and the initial judgment, a significant factor in advice utilization. Another behavioral experiment (N = 100) confirmed that our proposed advice design overcomes differences among advisors.

Keywords: decision-making, algorithm aversion, algorithm appreciation, design advice

Introduction

Advice-Taking in the Era of Artificial Intelligence

Decision-making is difficult for humans. People often seek the advice of others to make better decisions. For example, physician may ask other specialists about a patient's diagnosis. However, humans do not fully utilize advice from others (egocentric discounting) (Yaniv & Kleinberger, 2000).

Today's computational algorithms (hereafter referred to as "algorithms") have outperformed humans in some tasks given the rapid development of machine learning (i.e., artificial intelligence [AI]) and currently support a wide range of decision-making scenes, such as medical diagnostic support and automated driving. Experimental findings have shown that algorithmic advice (i.e., advice calculated by an algorithm) tends to be used more often than human advice (algorithm appreciation) (Logg, Minson, & Moore, 2019); however, many experiments have shown that algorithmic advice tends to be used less often than human advice (algorithm aversion), which is a problem in the era of AI (Dietvorst, Simmons, & Massey, 2015[2]). Then, designing advice that overcomes algorithm aversion and can be applied to a wide variety of AI-supported decision-making is required. If the advice design can be described mathematically or logically, a module for individualizing advice can be integrated into AI decision support systems (Fig. 1).

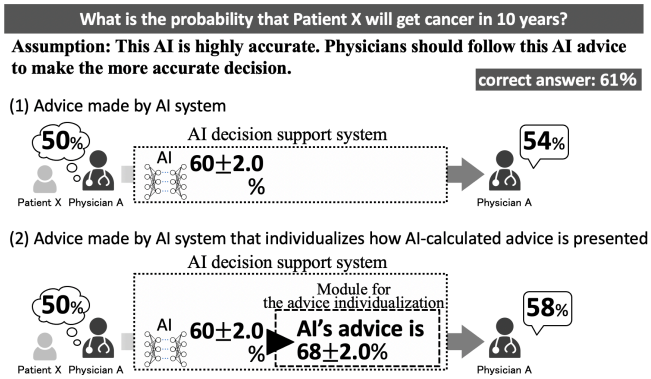

Figure 1: Conceptual image of the AI system showing individualized AI-calculated advice.

Challenges in the Design of Algorithmic Advice

Advice Taking. Advice taking is generally studied in the context of the judge-advisor system (JAS; Sniezek & Buckley, 1995), a widely used experimental protocol for examining advice-based decision-making (Fig. 2). The judge initially decides without advice, checks the advice, and then makes a final decision. Egocentric discounting is the most consistent finding in the field, regardless of the source of advice (Bailey, Leon, Ebner, Moustafa, & Weidemann, 2022). However, many challenges remain in designing algorithmic advice.

First, a widely accepted and unified model that accounts for advice taking is nascent. While a widely accepted and unified model that accounts for these factors is expected, research is still nascent (Himmelstein, 2022). No unified perspective, such as Bayesian updating (Robalo & Sayag,

[1] The present affiliation is Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology.
[2] While this paper discusses the effect of seeing an algorithm err has on people's likelihood of choosing the algorithm, this paper has been generally cited as explanation of how infrequently people use algorithms.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

2018) or heuristic approaches (Pescetelli, Hauperich, & Yeung, 2021; Yaniv, 1997), has been reached regarding how advice is integrated into judgments.

The factors that influence the use of advice, such as the difficulty of the task, the performance of the advice, the expertise of the advisor, the confidence of the judges, and the distance between the advice and the initial judgment (distance effect) have been reported (Bailey et al., 2022; Himmelstein and Budescu, 2023; Moussaïd, Kämmer, Analytis, & Neth, 2013). The distance effect has been reported to have a significantly greater impact on advice use than differences in advice sources (Himmelstein and Budescu, 2023) or judges' confidence (Moussaïd et al., 2013).

Second, a few studies discuss advice with an interval. Algorithmic advice for numerical estimation is often presented as estimates with intervals (e.g., 60.0 $\pm$ 2.0%) rather than based on point estimates (e.g., 60.0%). Dievorst & Bharti (2020) and Goodwin et al. (2013) discussed algorithmic advice with an interval, but their work did not follow the experimental protocol of JAS. Önkal et al. (2017) discussed two levels of trust in advice sources and their impact on the use of advice with an interval but did not examine distance effects.

Third, few studies have attempted to design advice. Cheng & Chouldechova (2023) proposed overcoming algorithm aversion through intervention in the algorithm development stage, but it does not follow the experimental protocol of JAS and lacks broad applicability.

AI Trust or Transparency and Advice Taking. While attempts to improve the transparency (explainability) of AI to increase judges' understanding and trust in AI advice have been made (Ali et al., 2023), judges' understanding of advice performance does not necessarily affect advice discounting (Goodwin, 2000; Goodwin & Fildes, 1999; Lim & O'Connor, 1995). Moreover, trust in or preferences for advisors do not necessarily affect advice use (Goodwin et al., 2013; Önkal et al., 2017; Himmelstein and Budescu, 2023). Simply increasing transparency or trust in AI is not sufficient to increase the use of algorithmic advice.

Objects and Research Questions

As a first step to realize a widely applicable advice design that overcomes algorithm aversion, this study exploratively describes the characteristics of the use of advice with an interval and proposes a method to individualize advice with an interval to improve advice use and overcome differences in advice use across advice sources.

Focusing on the distance effect, which has been reported to have a significant impact on advice use, this study individualizes the presentation of advice so that the distance between the advice and the initial judgments to reduce the rate that advice is ignored. Since the distance effect is quantifiable, it is suitable for mathematical individualizing advice and incorporating individualized advice presentation into AI decision support systems. Since there is no unified model of the advice-use process, even for advising point estimates, we design advice based on the results of behavioral experiments.

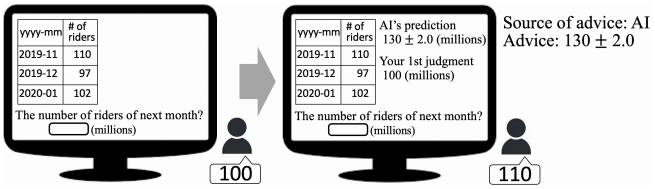

Figure 2: The experimental protocol following JAS.

Object 1: This study exploratively describes the characteristics of the use of advice with an interval.

RQ1-1. Is egocentric discounting observed for advice with an interval? Does advice distance affect advice use?

RQ1-2. Do differences in advice sources affect advice use?

Object 2: Our individualization of the advice presentation eliminates differences in the utilize of advice (with an interval) across sources of advice.

RQ2. Can individualized advice presentation based on distance effects overcome differences in advice use across sources?

Overview of Behavioral Experiments

Three experiments corresponding to each research question were conducted online. Model details, codes, and results are published in the supplemental files[3].

General Procedure. For each experiment, human judges (participant) answered 60 randomly presented tasks. All tasks followed the context of JAS due to its prevalence in AI decision support systems (Fig. 1). We focused on advice with an interval whose mean value is the correct answer (hereafter *"target"*). Human judges followed the procedure below.

First, a judge makes an initial judgment J_1 (100 in Fig. 2). Next, the advice "$target \pm Err_{advice}$" is presented to the judge (130 $\pm$ 2.0 in Fig. 2). The judge then makes the final judgment J_2 (110 in Fig. 2).

To prevent the judges from consistently answering the value of $target$ as J_2, they were not informed that the median advice value was *target*, nor was the value of *target* for each task revealed to them.

Details of the Task. This study employed tasks to forecast the number of New York subway riders (in millions) for the following month using values from the past nine months,[4] referring to Himmelstein & Budescu (2023). The mean of the *targets* was 521.2 $\pm$ 198.2. The three experiments in this study vary in terms of advice source and conditions, with specifics detailed in their respective chapters.

Advice with an Interval. Advice is a combination of Err_{advice} and a source of advice. For example, "AI's advice is $target \pm Err_{advice}$." Err_{advice} was set at level 60 (e.g., 0.0, 0.95, 1.9, …, 56.3) based on the differences between

[3] https://osf.io/58mkp/?view_only=410b1a7e8f2c4119a1f8f355a3ac1ab1

[4] https://catalog.data.gov/dataset/mta-subway-customer-journey-focused-metrics-beginning-2015 (last accessed 2023/08/30)

judgments and the $targets$ in a pilot study in which the same 60 tasks were judged without advice so that judges did not notice that the mean value of the advice was $target$[5] .

Overview of the Analysis

The common points of the three experiments are described.

Evaluation metrics. The Weight of Advice (WoA; Harvey & Fischer, 1997), a widely used metric for advice use[6], was adopted. When the final judgment J_2 equals $target$, a judge is assumed to have perfectly followed the advice. A WoA of 0 means no judgment updating, while a WoA of 1 means that J_2 equals $target$. A WoA of 0.5 means that judgment was updated by half of the distance between J_1 and $target$.

$$WoA = (J_1 - J_2)/(J_1 - target) \quad (Eq.1)$$

Statistics Analysis. We conducted a statistical analysis primarily using multilevel model linear regression in accordance to previous studies that discussed distance effects for advice based on point estimates (Pálfi, Arora, & Kostopoulou, 2022; Schultze, Rakotoarisoa, & Schulz-Hardt, 2015). Parameter fitting is performed using Markov chain Monte Carlo methods, and a parameter is considered significantly positive or negative if the 95% credibility interval (CI) of the estimated value of each parameter does not contain zero. R-hats were under 1.1 for all analyses and parameters. All multilevel models, judges, tasks, and Err_{advice}, which was noted to be related to the advice use in the previous work, were set as random intercepts and slopes.

R 4.1.0, brms 2.15.0, and RStan 2.21.2 were used for the subsequent analysis.

Recruitment of the Participants and Ethics

Participants (judges) were recruited for each of the three experiments through Lancers, one of the largest commercial crowdsourcing platforms in Japan. The judges were native Japanese speakers, and all the experiments were conducted in Japanese. Since power analyses for random effects multi-level models are rather complex (and beyond our current understanding), we followed the previous study that unveiled the distance effects for advice based on point estimates (Schultze et al., 2015), and 100 judges were recruited for each experiment. None of the judges worked for a railway company; therefore, we judged that none of the judges were task experts. Each judge was paid approximately $10. All experiments were approved by the Ethics Committee of Institute of Medicine, University of Tsukuba (Approval No. 1734-1).

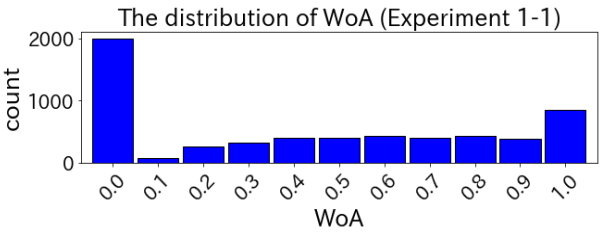

Figure 3: The distribution of the WoA of Experiment 1-1.

Experiment 1-1 (RQ1-1)

Experiment 1-1 is an exploratory description of the use of advice with an interval when there is one source of advice.

Experiment Procedures and Participants

The source of advice is AI[7] for all tasks. Each of the 100 judges (n_{female} = 34, n_{male} = 65, n_{other_gender} = 1, M_{age} = 43.3, SD_{age} = 9.08) answered the 60 tasks. 5,996 responses were included in the analysis; four were excluded due to missing data records caused by communication errors.

Results and Discussion

Egocentric discounting of the advice with an interval. First, we calculated WoA for all responses (Fig. 3). The mean value was 0.44 (SD = 0.38, Median = 0.43). We ran an empty multilevel model with WoA as the dependent variable and random intercept and slope by judges, tasks, and Err_{advice}. The model confirmed that WoA was significantly larger than 0, and 95% credibility interval were under 0.50 (b = 0.44 [0.38, 0.49]). This means that, on average, judgments were updated by under half the distance between J_1 and $target$. Thus, egocentric discounting was observed for advice with an interval.

Advice Distance and Advice Use. For advice using point estimates, distance effects have discussed the distance between J_1 and advice. In the case of advice with an interval, first, it is necessary to discuss which point of advice should be used to determine the advice distance. This study discussed the three types of points; the point closer to J_1 of the upper or lower limit of the advice (ADV_{near}; 128.0 in Fig. 2), the median of the advice ($target$; 130.0 in Fig. 2), and the point far to J_1 of the upper or lower limit of the advice (ADV_{far}; 132.0 in Fig. 2). The distance between these three types of points and J_1 are called $Dist_{near}$, $Dist_{target}$, and $Dist_{far}$, respectively. All types of the distances were normalized by $target$. All distances were the absolute value,

[5] For more details, please refer to supplemental file.

[6] WoA is typically clipped to have a bounded magnitude (0: no judgment updating–1: $J_2 = target$). We replaced all scores larger than 1 with 1 and all negative scores with 0 based on previous studies (Soll & Larrick, 2009; Schultze et al., 2015; Logg et al., 2019). The number of responses in which WoA was replaced is shown in supplement file for each experiment. No response was found in this study for which $J_1 = target$, i.e., WoA was infinite.

[7] In Japanese, the word "AI" semantically includes algorithms and is more popular than the word "algorithm," so this study uses AI instead of algorithm. For all the experiments, each judge answered the free description question "What is artificial intelligence (AI)?" The first author checked all the responses and found no obvious misunderstandings. Then, we judged that the judges knew about AI. "AI" as source of advice is just an experimental setting; the "AI" did not actually perform the estimation.

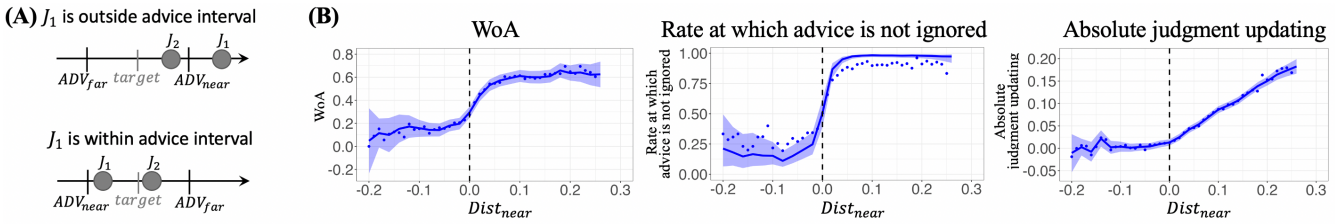

Figure 4: (A) The concept image of each variable of advice and judgement. (B) Distance effects for Experiment 1-1. The blue dots were observed data. The bold blue line represents the mean value, while the shaded area represents the 95% credible interval, which were estimated using a state–space model. A negative $Dist_{near}$ means that J_1 is within the advice interval. The black dashed means that $Dist_{near} = 0.0$.

except that $Dist_{near}$ was set to a negative value when J_1 is within the advice interval (128.0 – 132.0 in Fig. 2) (Fig.4 (A)).

As methods, this study referred to a previous work that discussed the distance effects of advice using point estimates (Schultze et al., 2015) and WoA, the rate at which advice is not ignored, and absolute judgment updating were used as metrics. The rate at which advice is not ignored represents the rate in which judgment is changed based on the advice (i.e., $J_1 \neq J_2$). Absolute judgment updating was defined as the distance approaching the *target* by updating from J_1 to J_2 (0 when $J_1 = J_2$) to avoid situations in which differences cancel each other out. For each of these three metrics, the model was expressed as a linear sum of the linear and squared distances with random intercepts and slopes by judges, tasks, and Err_{advice}. These models are constructed for each of the three types of distances. Among the three types of distances, a type of distance with the best model fit was selected to be used for discussion of distance effect. Model fit was evaluated using the widely applicable information criterion (WAIC)—a smaller WAIC means a better model fit.

As a result, as shown in Table 1, the model using $Dist_{near}$ had the smallest WAIC for two of the three metrics. Fig. 4(B) shows the relationship between $Dist_{near}$ and the values of each metric based on the observed results and the estimated values using the state space model. WoA and the rate at which advice is not ignored did not increase as $Dist_{near}$ increased and peaked when $Dist_{near}$ was 0.18 and 0.10, respectively. Absolute judgment updating tended to increase monotonically with an increase of $Dist_{near}$. These are equivalent to the results reported for advice using point estimates (Schultze et al., 2015). The stimulus-response model (Stevens, 1957) has been proposed as one of the models of theoretical background for distance effects. It was assumed that the theoretical background could be applied to advice with an interval.

Based on these results, we concluded that there exists a distance effect when advice is used with an interval, and that the distance should be discussed using $Dist_{near}$. In addition, Fig. 4(B) shows that when J_1 is within the advice interval (i.e., $Dist_{near} \leq 0$), advice tends to be unused.

Table 1: WoA for each parameter and distance.

	$Dist_{near}$	$Dist_{target}$	$Dist_{far}$
WoA	**-24.5**	337.2	927.5
Rate at which advice is not ignored	**2825.4**	3200.2	4033.7
Absolute judgment updating	-23191.0	**-23418.7**	-22463.8

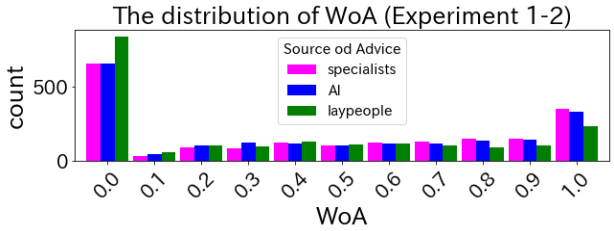

Figure 5: The distribution of the WoA score of the three types of sources of advice of Experiment 1-2.

Experiment 1-2 (RQ1-2)

Experiment 1-2 was an exploratory description of the differences of the use of advice with an interval by different source of advice.

Experiment Procedures and Participants

Experiment 1-2 set up three sources of advice: AI, 10 railway company employees (hereafter "specialists"), and 100 laypeople who had answered the corresponding task in the past (hereafter "laypeople")[8]. Source of advice and Err_{advice} was randomly changed for each task. In summary, Experiment 1-2 is the within-judges comparison.

Each of the 100 judges (n_{female} = 36, n_{male} = 64, M_{age} = 44.0, SD_{age} = 9.07) answered the 60 tasks. 5,993 responses were included in the analysis, and seven were excluded due to communication errors.

Results and Discussion

Source of Advice and Advice Use. The mean WoA scores for each source of advice were as follows: 0.46 (SD = 0.40, median = 0.45) for specialists, 0.44 (SD = 0.39, median = 0.40) for AI, and 0.35 (SD = 0.38, median = 0.25) for laypeople (Fig. 5). We constructed a multilevel regression model with random intercepts and slopes by judges, tasks, and Err_{advice} in which we regressed WoA on the source of

[8] As in Experiment 1-1, the three sources of advice are just experimental settings; the presentation to the judges about the source of advice was automatically changed. Nor did the "10 railway company employees" or the "100 laypeople" actually estimate for each task.

advice. In this model, the value of WoA when the advice source is laypeople is estimated as the intercept, and the difference in WoA when the advice source is AI or specialists is estimated as the coefficient value. The results were as follows: intercept (laypeople): 0.35 [0.30, 0.41], b_{AI}: 0.08 [0.05, 0.12], and $b_{specialists}$: 0.10 [0.06, 0.14]. The findings that b_{AI} and $b_{specialists}$ were significantly positive indicate that WoA was significantly higher when the source of advice was AI and specialists compared to when the source of advice was laypeople. This result showed that the use of advice with an interval was affected by the source of advice.

The Influence of Different Sources of Advice on the Advice-Using Process. This analysis aimed to identify which stages of the advice-use process are affected by differences in the source of advice. For this purpose, we referred to the steps of process of advice use proposed theoretically by Himmelstein (2022): (1) the process of deciding whether to use the advice (i.e., whether to update the judgment), and (2) the process of deciding where to update the judgment if the judge use (i.e., does not ignore) the advice. Then, this study analyzed which of the rate at which advice is not ignored and absolute judgment updating when the advice was not ignored was affected by differences in the source of advice.

For each of the rate at which advice is not ignored or absolute judgment updating, we constructed the multilevel regression model with random intercepts and slopes by judges, tasks, and Err_{advice} in which we regressed each of the two metrics on the source of advice. In this model, the value of each metric when the advice source is laypeople is estimated as the intercept, and the difference of each score when the advice source is AI or specialists is estimated as the coefficient value. The results for the rate at which advice is not ignored, intercept (laypeople): 0.60 [0.53, 0.67], b_{AI}: 0.09 [0.04, 0.13], $b_{specialists}$: 0.09 [0.05, 0.14]. For absolute judgment updating, intercept (laypeople): 0.07 [0.06, 0.08], b_{AI}: 0.00 [-0.00, 0.01], and $b_{specialists}$: 0.01 [0.00, 0.01]. The findings that b_{AI} and $b_{specialists}$ were significantly positive for the rate at which advice is not ignored imply that differences in advice sources lead to differences in advice use, mainly by influencing the step in deciding whether to use the advice.

Experiment 2 (RQ2)

This study proposed to individualize the presentation of the advice referring to the results of Experiments 1-1 and 1-2. A behavioral experiment tested whether our proposed advice design would improve the use of advice with an interval when specific source of advice is used.

Advice Design

Experiment 1-2 confirmed that the rate at which advice is not ignored was significantly affected by differences in the source of advice. Moreover, the distance effect has been reported to have a main impact on advice use. Therefore, this study proposed an advice design method to individualize the presentation of advice corresponding to J_1 so that the rate at which advice is not ignored is higher than when advice presentation is not individualized. For individualizing the advice presentation, the value of *target* would be changed and the value of Err_{advice} would be not changed. The value of the changed target is called $target_{tmp}$ and the corresponding $Dist_{near}$ is called $Dist_{near\_tmp}$.

original advice "$1{,}010 \pm 50$", $J_1 = 900$

target	1,010	$target_{tmp}$	1,020.1	1,030.3	1,040.6	1,051.0
ADV_{near}	960	ADV_{near_tmp}	970.1	980.3	990.6	1,001.0
$Dist_{near}$	0.059	$Dist_{near\_tmp}$	0.069	0.078	0.087	0.096
Expected value of WoA based on $Dist_{near}$	0.56	*Expected value of WoA based on* $Dist_{near\_tmp}$	0.56	0.58	0.58	0.61
Distance between *target* and expected J_2	47.98	Distance between *target* and expected J_2	42.28	34.22	28.23	**17.26**

individualized advice "$1{,}051 \pm 50$"

Figure 6: The example of the advice individualization.

For design of advice, this study focused the following two points. First, in Experiment 1-1, the $Dist_{near}$ at which the rate at which advice is not ignored peaked was 0.10 and the rate at which advice is not ignored did not decrease significantly even when $Dist_{near}$ became larger than 0.10 (Fig. 4(B), middle figure). Second, we cannot deny the possibility that J_2 may move away from the *target* when judge updates the judgement based on $target_{tmp}$ (i.e., individualized advice), compared to when judge updates the judgement based on the original *target* (i.e., non-individualized advice). This study aimed to design $target_{tmp}$ so that the judgment move toward *target*. Here, based on the Eq. 1, J_2 can be calculated from the values of WoA, J_1, and *target*. Moreover, if we also use the results of the relationship between $Dist_{near}$ and WoA (Fig. 4 (B), left figure), which was estimated using the state space model in Experiment 1-1, we can calculate the expected value of J_2 corresponding to $Dist_{near}$. Then, $Dist_{near\_tmp}$ would be obtained from $target_{tmp}$, Err_{advice} and J_1. And based on these values, the expected J_2 value for each $target_{tmp}$ would be calculated.

Specific Protocol. Based on above discussion, our proposed advice design is as follows: When $Dist_{near}$ is less than 0.1 for J_1, the $target_{tmp}$ was changed by 0.01 times of *target* so that $Dist_{near\_tmp}$ becomes larger while $Dist_{near\_tmp}$ is less than 0.1. The expected J_2 value corresponding to each $Dist_{near\_tmp}$ and each $target_{tmp}$ was calculated. Among the various values of $target_{tmp}$, the $target_{tmp}$ value with the smallest error (absolute value) between the expected J_2 and *target* was adopted as the median value of the individualized advice (Fig. 6). If the distance (absolute value) between the expected J_2 and *target* was larger than the value calculated from the original *target* in any of the $target_{tmp}$ values, the advice was not changed.

Experiment Procedures and Participants

As in Experiment 1-2, we set up three different sources of advice (specialists, AI, and laypeople). We tested the effect of our proposed individualization of advice presentation by setting the individualization for the cases in which the laypeople were the source of advice because significantly

lower WoA values were shown for laypeople assigned as the advice source in Experiments 1-2. Of the 20 tasks in which laypeople were assigned as the advice source, 13 or 14 randomly assigned tasks were set up with individualization and the others were set up without individualization. The judges were not informed of this setup. Experiment 2 is a within-judges comparison.

Each of the 100 judges ($n_{female} = 47$, $n_{male} = 53$, $M_{age} = 45.3$, $SD_{age} = 10.13$) answered the 60 tasks, and 5,990 responses were included in the analysis and 10 responses were excluded due to communication errors. In Experiment 2, the judge's confidence for each judgment (J_1 and J_2) was obtained on a 101 Likert scale (0: *not confident at all* – 1: *very confident*) after the judgment.

Results and Discussion

For the sake of brevity, we present the results as if they were from four different sources of advice: specialists, AI, laypeople with advice-individualization, and laypeople without advice-individualization. We must note that the original *target* was used for the calculation of WoA.

The Effects of Our Advice Design. The mean value of WoA corresponding to each source of advice is as follows; 0.49 (SD = 0.39, median = 0.51) for specialists, 0.47 (SD = 0.39, median = 0.50) for AI, 0.41 (SD = 0.40, median = 0.36) for laypeople without individualization, and 0.57 (SD = 0.42, median = 0.69) for laypeople with individualization.

We constructed a multilevel regression model with random intercepts and slopes by judges, tasks, Err_{advice}, and confidence scores for J_1, where we regressed WoA on the source of advice. In this model, the value of WoA when the advice source is laypeople without individualization is estimated as the intercept, and the difference in each score when the advice source is AI, specialists, or laypeople with individualization is estimated as the coefficient value. The results were intercept (laypeople without individualization): 0.42 [0.35, 0.48], $b_{laypeople\_w/\_individualization}$: 0.17 [0.12, 0.22], b_{AI}: 0.06 [0.01, 0.10], $b_{specialists}$: 0.08 [0.04, 0.12]. The result that $b_{laypeople\_w/\_individualization}$ was significantly positive confirmed that our advice design significantly improved advice use. Furthermore, since the 95% CI for $b_{laypeople\_w/\_individualization}$ was not overlapped and greater than the 95% CI for the b_{AI}, we can interpret this as a significant increase in advice use by laypeople with individualization over the use of advice by AI. These results are confirmed by the mean values and 95% intervals of the results of the sampling with replacement by 10,000 times of the results for each source of advice: 0.493 [0.448, 0.540], 0.474 [0.429, 0.519], 0.410 [0.358, 0.463], and 0.575 [0.520, 0.628] for specialists, AI, laypeople without individualization, and laypeople with individualization, respectively (Fig. 7). These results indicate that our proposed advice design overcomes poor advice utilization due to differences in advice sources.

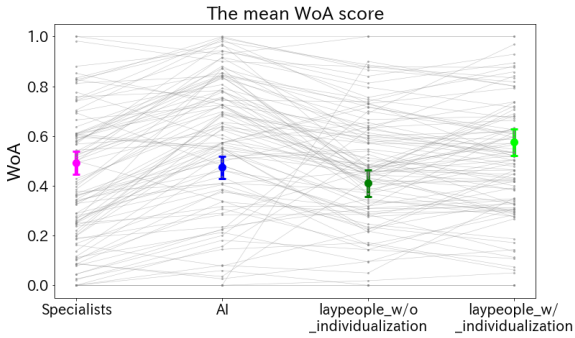

Figure 7: For Experiment 2, the mean WoA for each source of advice. The color points and 95% CIs were the results of the sampling with replacement. The gray dots are the mean value of each judge, and the values of the same judge are connected by gray lines.

General Discussion

This study showed that the differences in advice use due to differences in advisors were overcome by presenting advice in which the distance between the advice and the initial judgment was individualized corresponding to the initial judgment. Our proposed individualization of advice focused on the distance effect, which reportedly has a particularly large impact on advice use. We intend to explore the effects of different strategies of advice design in the future, such as promoting judges' motivation to use advice.

Our proposed advice design is widely applicable independent of the tasks. We will eventually test the effectiveness of our advice design by applying it to actual algorithmic advice used in tasks such as medical diagnosis by a physician (Pálfi et al., 2022). Conversely, our advice design specifically exaggerates advice, and it may not be socially acceptable in situations in which the user of the advice needs detailed information about the algorithm calculating the advice. Exploration of task-specific advice is a future work.

Trust in an advisor consists of presumed trust based on their position and empirical trust based on advice accuracy (Önkal et al., 2017). Judges in this study were not informed the correct answer for each task. Therefore, the differences in advice use due to advice sources in Experiment 1-2 can be attributed to differences in presumed trust in advice sources. Our proposed advice design in Experiment 2 mitigated the impact of source-based differences in advice use, suggesting that the distance effect is more influential than presumed trust for advice with an interval.

In the future, we would like to conduct a cross-domain study using a variety of tasks to prevent potential bias due to the limited language and tasks. Research on sequential advice use (Rebholz, Hütter, & Voss, 2023) and group advice use (Larson, Tindale, & Yoon, 2020) is also future work.

Conclusion

This study proposed a method to individualize the presentation of advice to change the distance between the advice and the initial judgment such that it reduces the rate advice being ignored. The method mitigated the impact of source-based differences in advice use. This study is the first step toward realizing a widely applicable advice design that overcomes algorithm aversion.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805.

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2022). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 1-26.

Cheng, L., & Chouldechova, A. (2023, April). Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. *In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-27).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science*, 31(10), 1302-1314.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting, 16(1)*, 85-99.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy?. *Journal of Behavioral Decision Making*, 12(1), 37-53.

Goodwin, P., Gönül, M. S., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. International Journal of Forecasting, 29(2), 354–366.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2), 117-133.

Himmelstein, M. (2022). Decline, adopt or compromise? A dual hurdle model for advice utilization. *Journal of Mathematical Psychology*, 110, 102695.

Himmelstein, M., & Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, 36(1), e2285.

Larson Jr, J. R., Tindale, R. S., & Yoon, Y. J. (2020). Advice taking by groups: The effects of consensus seeking and member opinion differences. *Group Processes & Intergroup Relations,* 23(7), 921-942.

Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, 8(3), 149-168.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103.

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLoS ONE*, 8 (11), 1–8.

Önkal, D., Gönül, M. S., Goodwin, P., Thomson, M., & Öz, E. (2017). Evaluating expert advice in forecasting: Users' reactions to presumed vs. experienced credibility. *International Journal of Forecasting*, 33(1), 280–297.

Pálfi, B., Arora, K., & Kostopoulou, O. (2022). Algorithm-based advice taking and clinical judgment: impact of advice distance and algorithm information. *Cognitive Research: Principles and Implications*, 7(1), 1-17.

Pescetelli, N., Hauperich, A. K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. Cognition, 215, 104810.

Rebholz, T. R., Hütter, M., & Voss, A. (2023). Bayesian Advice Taking: Adaptive Strategy Selection in Sequential Advice Seeking. *Psyarxiv*, 1-60.

Robalo, P., & Sayag, R. (2018). Paying is believing: The effect of costly information on Bayesian updating. *Journal of Economic Behavior & Organization*, 156, 114-125.

Schultze, T., Rakotoarisoa, A. F., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment & Decision-making*, 10(2).

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision-making. *Organizational behavior and human decision processes*, 62(2), 159-174.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. Journal of experimental psychology: Learning, memory, and cognition, 35(3), 780.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64(3), 153–181.

Yaniv, I. (1997). Weighting and trimming: heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69, 237–249.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93, 1–13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, 83(2), 260-281.