Patches and Payoffs: A Pilot Study of Information Foraging and Forecasting Accuracy on Cultivate

Thomas Y. Li
University of St. Gallen

Abstract

How people use Cultivate can predict how well they forecast. Using webcam eye and screen-tracking (7 participants, 44 sessions, approx. 1612 mins) on the online forecasting platform, Cultivate, we compared the top forecasters to other groups. Three effective forecasting behaviors emerged: 1) time-allocation, top forecasters spent a smaller proportion of time on-platform (64% vs. 71-86%) and avoided overinvesting in crowd rationales (14% vs. 23-54%); 2) execution, when they did log in, they were productive (1.5 forecast updates per session vs. <0.86); and 3) calibration, they made smaller adjustments (6% vs. 8-14%). Lower performers showed two potential traps: excessive forecast formulation and rumination, as well as endless crowd rationale engagement without further investigation. Our findings from the pilot study show that platform engagement does not necessarily translate to better forecasters; selective, integrative foraging does. We propose platform tweaks, including varied visual hierarchy, decay of scent for over-visited areas, re-entry and exit prompts, and "update assistants" to nudge advantageous behaviors.

Keywords:

Forecasting platforms, Information Foraging Theory (IFT), superforecasters, screen tracking, eye tracking, forecast calibration, strategic decision making, UI, UX, interface navigation

Introduction

While we know a great deal about what makes someone a good forecaster, we know far less about how good forecasters forecast. Research on forecasting expertise has produced a detailed inventory of psychological attributes that distinguish top performers from the rest, and investigations have also begun to examine how emotional and cognitive factors, both independently and in interaction, might influence prediction accuracy. This body of work has substantially advanced our understanding of the traits that characterize effective forecasters. But knowing the ingredients of expertise is not the same as knowing the recipe: prior work tells us what distinguishes good forecasters, but less about *how* they apply these strengths in practice. In other words, in what way do expert forecasters truly navigate the process of making a prediction?

Our question is crucial because behavioral observation often reveals aspects of expertise that performance measures may miss. Watching how chess players view the board showed that experts perceive meaningful patterns when novices see individual pieces (Chase & Simon, 1973) and tracking how radiologists scan mammograms highlighted that experts detect abnormalities in seconds through different search patterns (Kundel et al., 2007). If forecasting expertise follows similar principles, then observing how top forecasters navigate information-rich environments – where they look, how long they engage, when they move on – may reveal as much as measuring their psychological traits or performance. Accordingly, our pilot study shifts from the laboratory to the field, employing eye-tracking methodology to observe how forecasters genuinely behave on Cultivate Labs (hereafter "Cultivate"), a widely used crowd forecasting platform.

Prior work in large geopolitical tournaments identified a cohort of exceptionally accurate forecasters, often termed “superforecasters”, representing the top 2% of participants (Mellers, Stone, Murray, et al., 2015; Mellers et al., 2017). The superforecasters consistently outperform non-superforecasters in a myriad of ways. On average, superforecasters, compared to non-superforecasters, are more diligent problem solvers (i.e., break intractable problems into tractable subproblems), open-minded with greater fluid and crystallized intelligence, motivated (i.e., both in the number of questions attempted, and the number of prediction updates made), as well as have several forecasting-

specific skills (e.g., identification of accurate base rates) (Mellers, Stone, Murray, et al., 2015; see also Mellers et al., 2017).

Although all forecasters are engaged with the same platform to make predictions, prominent User Interface (UI) research (e.g., Shneiderman et al., 2016) has suggested that a user's flow (movements) on a website, as well as one's focus of attention, can be influenced by the layout of the website. As such, every website, including Cultivate, contains design choices, intentional or not, that shape the behavior of any user participating on the website.

It would be relevant to the forecasting community to qualitatively examine how top forecasters make predictions on Cultivate and whether their utilization of the platform differs from that of non-top forecasters. Understanding these behavioral distinctions can provide valuable insights into how forecasting accuracy develops and how future training and platform design might be support it, especially through structuring the UI or user experience (UX) in ways that encourage expert-like forecasting strategies.

Leveraging extant research on the behavioral characteristics of superforecasters from Mellers, Stone, Murray, et al. (2015), we first attempted to predict how our cohort of top forecasters and non-top forecasters (average and bottom forecasters) would behave on Cultivate. Using eye-tracking to monitor forecasters' interactions, we then sought to validate these predictions in a naturalistic setting. The overarching research question guiding this study is: *How do top forecasters predict differently, compared to non-top forecasters, on the online forecasting platform, Cultivate?*

Drawing on preliminary evidence from this pilot study, we offer an initial view of how behavioral differences manifest on Cultivate through insights that not only illuminate advantageous forecaster behavior and platform design but also contribute to developing a key forecasting environment that fosters more accurate, reflective, inclusive, and informed decision-making in the future.

Theory and Hypotheses Development

All forecasters on Cultivate have one fundamental goal: to make the most accurate prediction. To do so, forecasters must not only leverage external research but also utilize

the wisdom of the crowd (i.e., the other forecasters) (e.g., Atanasov et al., 2017; Atanasov et al., 2025) to maximize the chances of deriving the most accurate forecast for any given question. Beyond the obvious utility of Cultivate providing a platform for entering and updating predictions, equally important is the wealth of information available through other users' predictions and rationales. Phrased differently, a key role of a forecaster on Cultivate is to search for, identify, and integrate valuable information both from within and outside the platform.

Fundamentally, a forecaster forages for relevant information to make a prediction. The Information Foraging Theory (IFT) (Pirolli, 2007; Pirolli & Card, 1999) draws an analogy between information-seeking behavior and foraging behavior in animals. Just as animals forage for food in an environment, individuals seek information to fulfill their informational needs and make decisions. The IFT predicts that information seekers will try to maximize information gain while minimizing the cost of information seeking. As such, the behaviors of a forecaster on Cultivate can be paralleled to an information forager, whereby the forecaster is seeking information to efficiently confirm, disconfirm, and add to their existing (or new) forecast.

Before conceptualizing Cultivate within the context of IFT, it is important to highlight the existing setup of the forecasting platform. The forecasting platform contains six main area-of-interests (AOIs) (see Djamasbi, 2014), a concept often used to determine one's reaction to specific areas of a page. Area A (Figure 1) is the landing page where all the forecasting questions, crowd forecast summary, number of forecasters participating, as well as total number of forecasts are displayed. After clicking a specific forecasting question from the landing page, multiple sections can be visited.

The first page after clicking a specific question is area B (Figure 2), the Forecast page of a specific prediction question, which can be separated into three parts: a) B1 displays the question being viewed, as well as any additional details, b) B2 is the menu bar allowing quick access to different areas of information and c) B3 is the answer area to input and submit your written rationales and predictions.

Figure 1: *Area A, Landing Page*

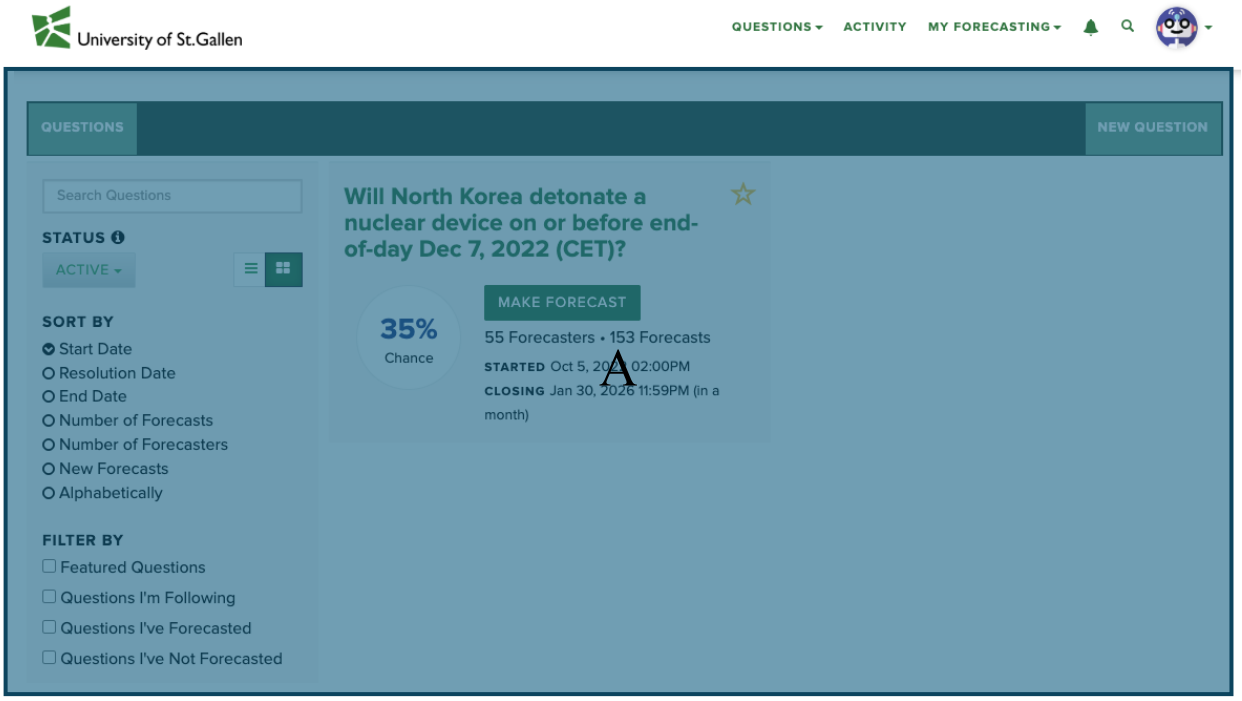


Figure 2: *Area B, Forecast Page*

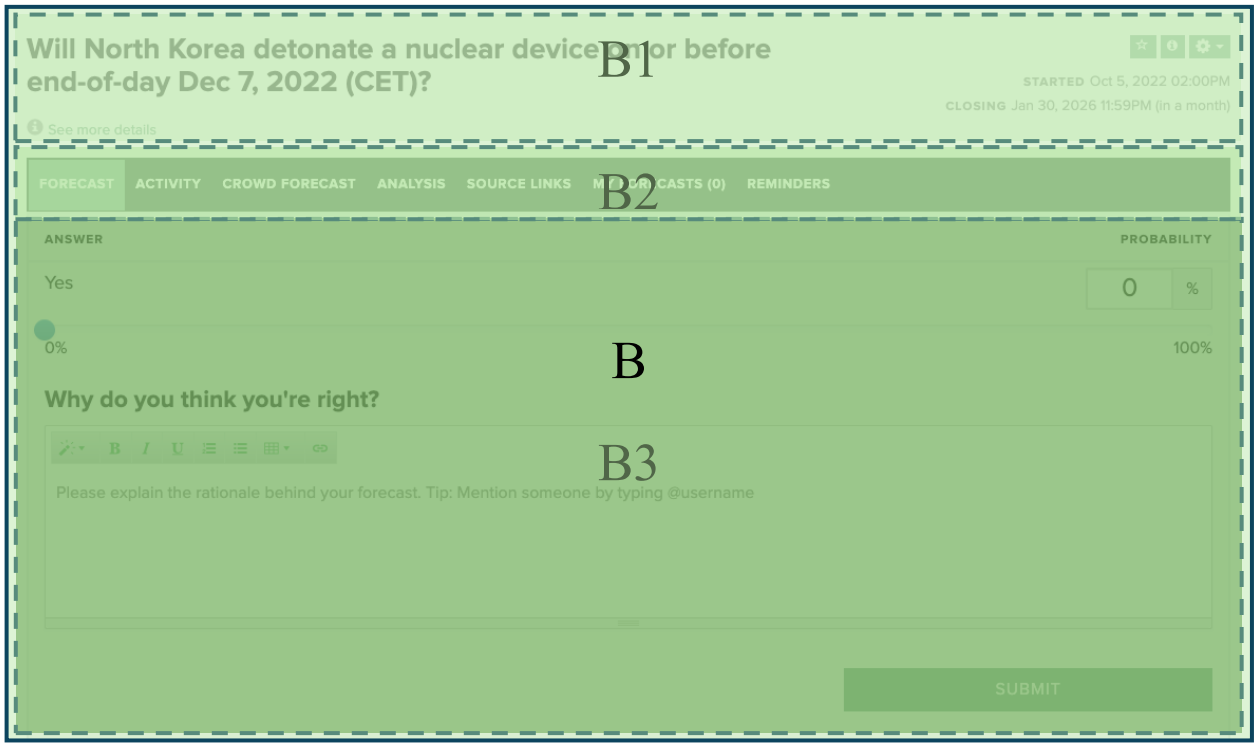


Figure 3: *Area C, Activity Page*

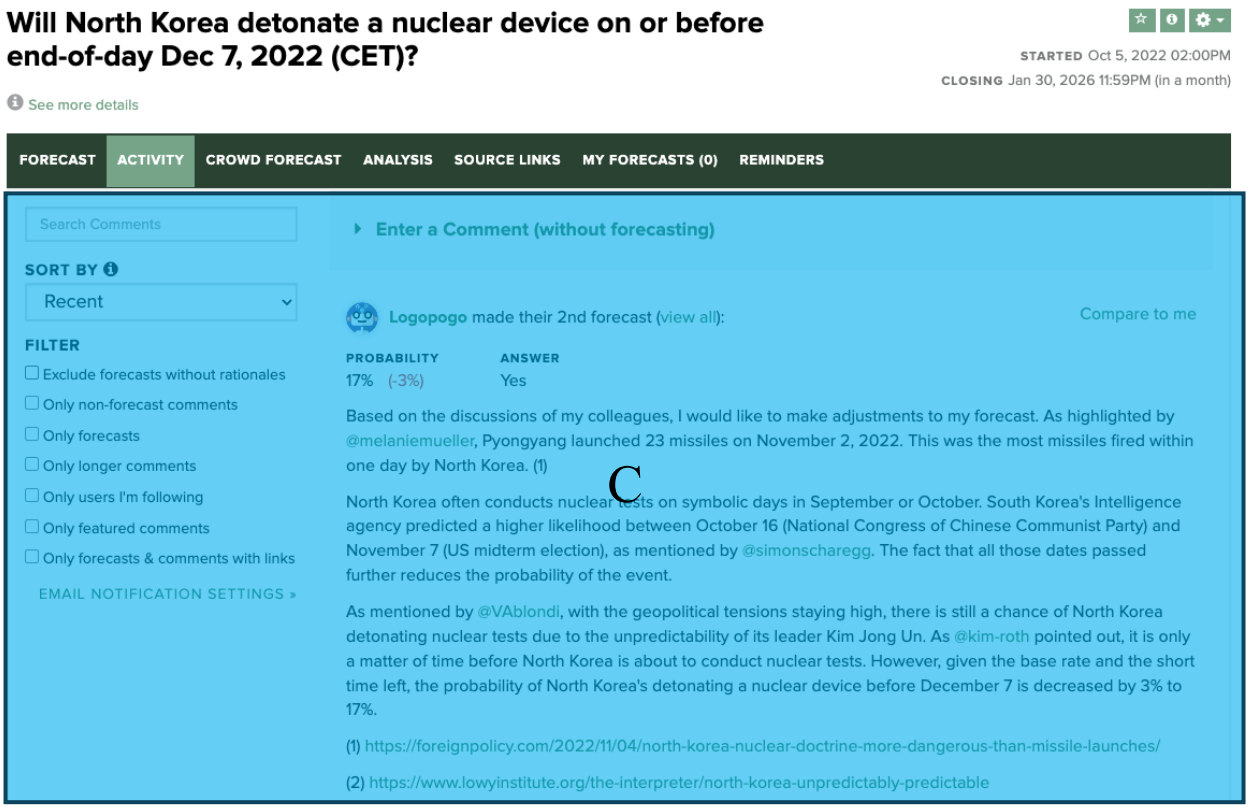


Figure 4: *Area D, Crowd Forecast Page*

Area C (Figure 3) contains the Activity page with numerical and verbal forecasts of other forecasters, and area D (Figure 4) is the Crowd Forecast page, including D1, the current forecast and forecast profile and D2, the crowd consensus trendline.

Area E (Figure 5) is the Analysis page, showing a forecast distribution of all forecasters, as well as update histories for each individual forecaster. Moreover, area F (Figure 6), Source Links, displays an overview of all cited sources from all participants, followed by area G (Figure 7), My Forecasts, listing the user's own forecast.

Figure 5: *Area E, Analysis*

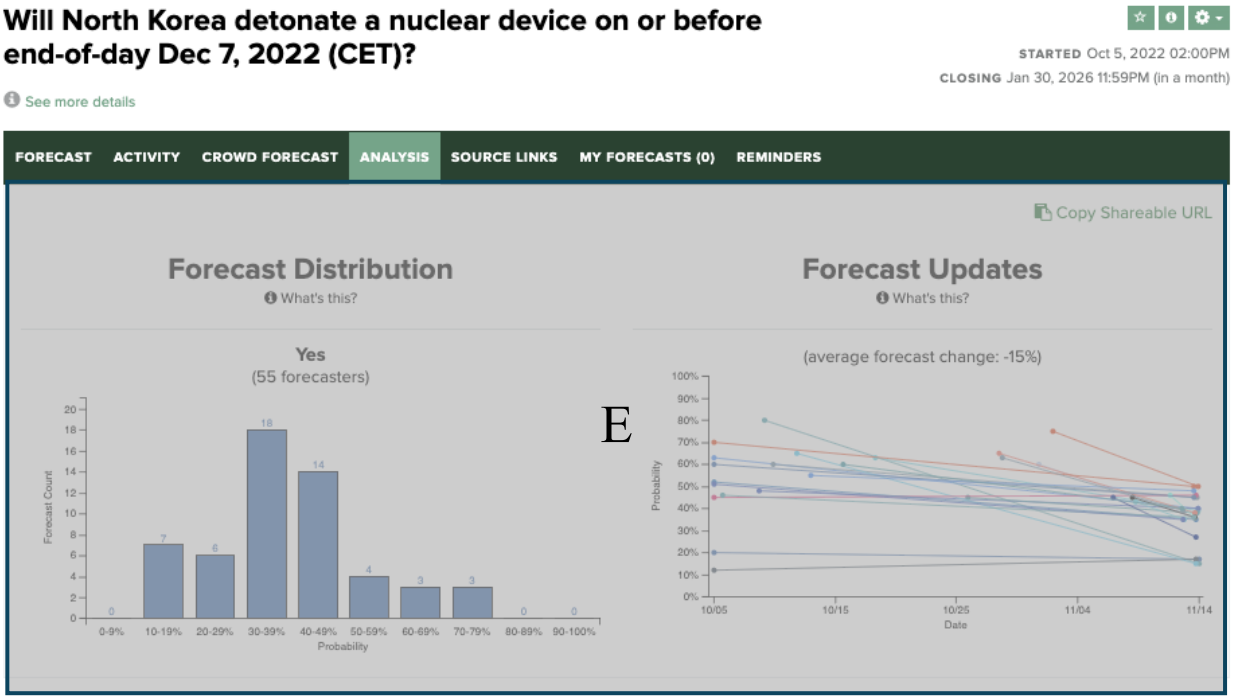


Figure 6: *Area F, Source Links Page*

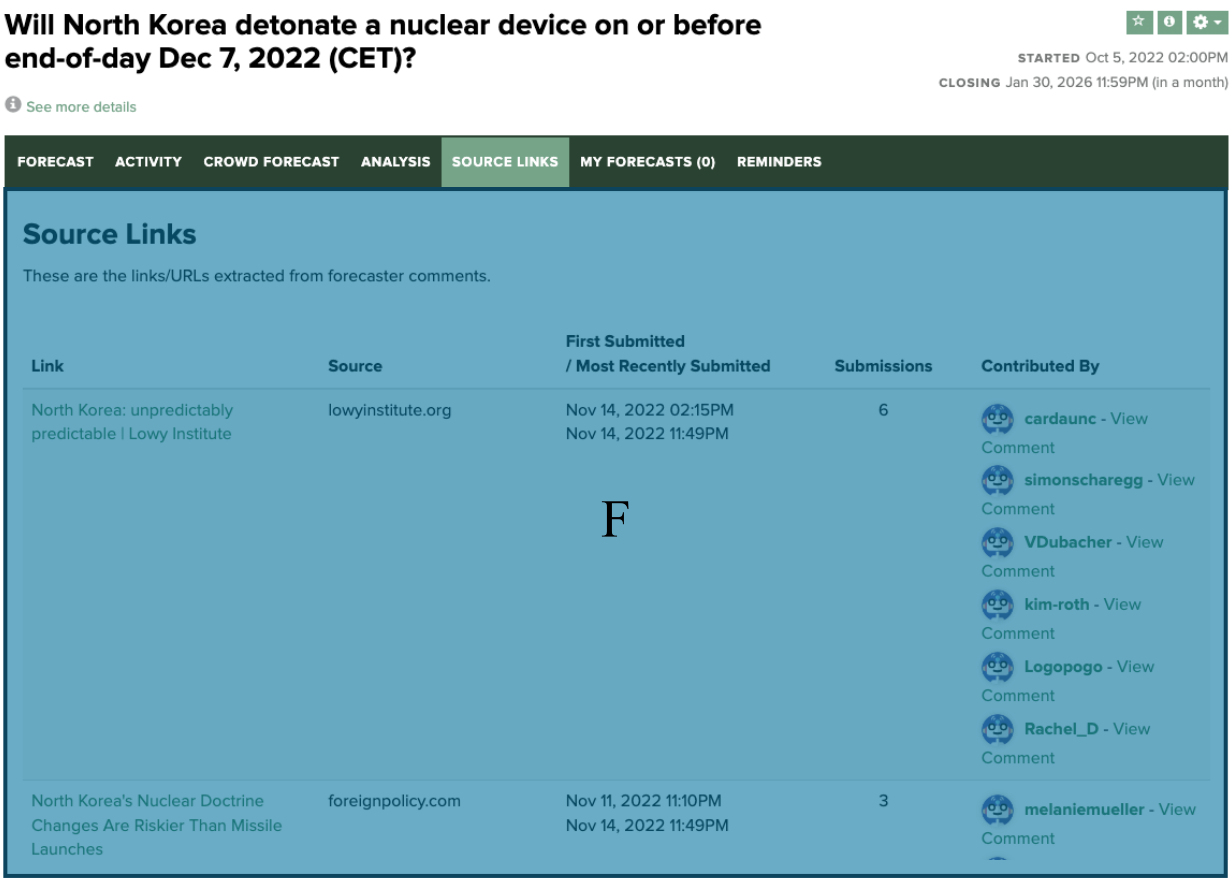
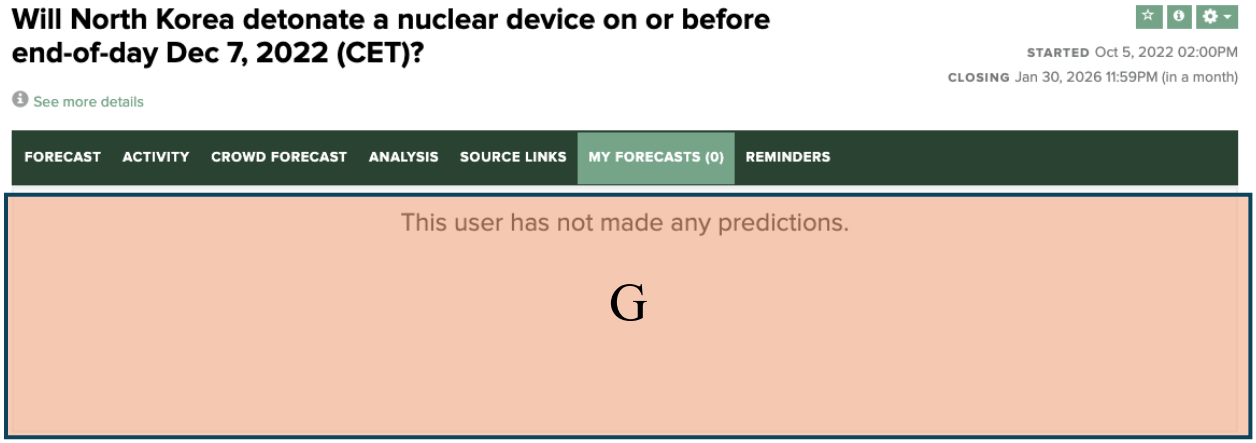


Figure 7: *Area G, My Forecasts Page*



When exploring web interfaces, including those of Cultivate, elements of the IFT are readily noticeable. Information scent (Pirolli & Card, 1999) refers to the perceived value or relevance of information based on cues in the environment. Users follow these cues (scents) to find information, and the strength of the scents influences whether users continue or abandon this search path. In the context of a website, the information scent refers to any cues that users can perceive with potential to help them assess the relevance and potential value of the information available. These can include headings/buttons (e.g., "Crowd Forecast" in the mid-upper menu bar), links (e.g., "Activity" on top-right corner), graphics (e.g., the circular crowd forecast percentage), and colors (e.g., green "make forecast" button), among many others. Researchers (e.g., Chi et al., 2001; Huberman et al., 1998; Pirolli & Fu, 2003) showed evidence that the difficulty of foraging on the web is related to the quality of the information scents available to the user. With strong scents, users move quickly toward the target information, leading to a much shorter information scent trail. Conversely, if the scent is weak, visitors take more of a meandering path through a page.

Information diet (Pirolli & Card, 1999) refers to a user's strategy for selecting and consuming information. Users make decisions about the type and amount of information to pursue based on the expected value and cost associated with obtaining the information. For a website, information diet refers to the distribution of attention and a pattern of browsing. For example, forecasters may focus mainly on browsing the Activity area and complement it briefly with Crowd Forecast area.

Patch foraging (Pirolli & Card, 1999) refers to the idea of searching for information in a specific area (or "patches", p. 646) rather than searching the entire information space. Users tend to focus on locations where they believe have a higher likelihood of finding valuable information. Throughout a website, the main information patches would be the subsections of the website that are often accessed through a menu bar. Adopting the IFT framework, each area on Cultivate (areas A through G) represents a distinct information patch. Marginal Value Theorem (Charnov, 1976), which IFT adapts to information patches, predicts that optimal foragers will allocate time to patches based on their expected

profitability (expected informational value relative to expected extraction cost) with the goal of maximizing profitability during information foraging.

On Cultivate, profitability is expected to vary across patches. Area C (Activity), area A (Landing Page) and area D (Crowd Forecast) contain concentrated decision-relevant information that varies in profitability as new forecasts are tallied. Area B (Forecast) contains mostly static information after the initial impression and provides an area to formulate forecasts. We expect top forecasters will demonstrate more optimal foraging behaviour by spending time proportional to their perception of each patch's expected value for making accurate predictions.

Notably, limited optimizations have been made to the visual hierarchy of Cultivate, as there are no obvious areas that are designed to quickly guide the attention of the user, which may provide relatively weak differential cues about patch profitability[1] (Chi et al., 2001; Faraday, 2000) and information scent. For example, the size of the text is relatively similar across the website, except for a few instances, such as the forecast question. Similarly, Cultivate appears to use a relatively uniform and constrained color palette, which may create a flatter visual hierarchy. Visual hierarchy research predicts that stronger contrast in size, color and/or position can guide more strongly towards the intended areas (Djamasbi et al., 2010; Faraday, 2000). Currently, little to no optimizations – whether by accident or on purpose – have been made to the layout of the forecasting platform by the developer.

Forecasting Behaviors on Cultivate

When making a forecast on Cultivate, two possible scenarios exist. The first scenario is where a participant is the first to forecast for a specific question on the platform, and there are no pre-existing numerical forecasts or rationales. The second scenario is that one is forecasting a question that already has pre-existing rationales and numerical forecasts. Behavioral differences are expected to arise depending on the scenario that a top forecaster is faced with. As our naturalistic experiment setting only provides observation opportunities for the examination of the second scenario, whereby prediction questions

[1] Patch profitability is the expected information gain (e.g., unique arguments encountered) per unit of time in a patch (area); when its marginal return falls below alternatives, efficient forecasters switch patches (Pirolli, 1999).

already have pre-existing rationales and numerical forecasts, the expected behaviors of the second scenario are tested. Nevertheless, the initial forecast scenarios are briefly described below to help highlight expected behavioral differences.

Top Forecaster Behavior Predictions: Initial Forecast (No Crowd and Existing Crowd[2])

Upon entering Cultivate, the first action we would expect top forecasters to take is to glance over the entire page of area A (i.e., the landing page when entering Cultivate). After selecting and entering a question-specific page, we would expect them to read both the question and question description (under "see more details") (area B1). In fact, we expect them to carefully parse the question and details – revisiting them as needed (Tetlock & Gardner, 2015). Next, we would expect the top forecasters to examine the rest of the page (area B2 and B3), where they further evaluate the decisions that need to be made. We would expect top forecasters to examine the choice architecture (Thaler & Sunstein, 2008; Thaler et al., 2013), particularly the type of choice environment (i.e., yes or no answer prediction, or multiple options) (area B3), as it would shape the type of answer that the top forecaster will compute. Similarly, the top forecasters are expected to glance over the menu bar (area B2), to grasp what sources of information are available to them.

Existing Crowd Scenario. In the existing crowd initial forecast scenario, top forecasters are expected to explore the Activity section (area C), the Crowd Forecast sections (area D1 and D2), Analysis section (area E), as well as the Source Links section (area F). In particular, the Crowd Forecast and the Analysis sections may intuitively – given their nomenclature – emit stronger information scent and likely attract initial attention from top forecasters. Given the abundance of information available on Cultivate, the top forecasters are expected to spend significantly more time in areas C through F, compared to the non-top forecasters. Top forecasters may go back and forth between the areas to isolate key pieces of information that are helpful for constructing their forecasts, and thereby, focusing on specific information patches. The main difference expected between top forecasters and non-top forecasters is the total amount of time spent forecasting during the initial visit.

[2] There is no consensus on what the minimum number of individuals should be in a crowd; required size appears task-dependent; however, crowds can sometimes be quite small (i.e., "just a handful") (Lyon & Pacuit, 2013, p. 1)

No Crowd Scenario. In the no crowd initial forecast scenario, there is no additional information value for the forecasters to check the other pages; however, if the top forecasters are extra diligent, they may still wish to browse through areas C through F to see whether there is any input, as the menu bar (area B2) serve as an information scent. Given the limited amount of information available on Cultivate, the top forecasters can be expected to quickly leave the platform early on. The main difference expected between top forecasters and non-top forecasters, for the no crowd scenario, is the total amount of time spent on the platform, as the top forecasters are expected to be more detail-oriented and to spend more time browsing through the platform.

The setup for the forecasting platform for our study is unique in that the forecasters were first asked to make private forecasts individually (i.e., before being introduced to Cultivate). Then, the individual forecasts were published on Cultivate by the forecasters simultaneously. As such, the "initial" forecast in this context was the first time a participant visited Cultivate after publishing their individual forecasts on the platform. As the forecasters would have already spent time in area A and B, it is expected that most of the forecasters' time during their forecast updates will be spent in other areas. The hypotheses are then indicated for the subsequent visits to Cultivate.

Top Forecaster Behavior Predictions: Subsequent Visits

We expect systematic behavioral differences between top forecasters and non-top forecasters during subsequent visits to Cultivate. Given top forecasters' tendency to collect and synthesize evidence, we expect them to devote a larger share of their on- and off-platform time to reading, examining, and questioning the forecasts and rationales of the crowd. Specifically, we would expect top forecasters to search for existing base rate estimates that represent outcome frequencies from similar prior cases (e.g., Chang et al., 2016; Kahneman & Lovallo, 1993; Tversky & Kahneman, 1982), whether calculations or verbal explanations, from existing forecasts. Top forecasters are skilled at identifying base rates and making calibrated adjustments (Mellers, Stone, Atanasov, et al., 2015; Tetlock et al., 2014).

Moreover, we expect the top forecasters to pay more attention to the forecasters that disagree with them compared to those who agree with their assessment (Karvetski et al.,

2022; Lord et al., 1984; Tetlock & Gardner, 2015). Non-top forecasters are expected to examine and interpret fewer opposing rationales (Hart et al., 2009). Because crowd rationales live in the Activity Page (area C), we test whether top forecasters allocate more (or less) time there:

Hypothesis 1. Top forecasters spend more time in area C compared to non-top forecasters.

Upon finishing the examination of crowd rationales, or perhaps even prior to examining the crowd rationales, we expect top forecasters to check the aggregate crowd forecast analytics (area D), especially the trendline that highlights the longitudinal change of forecasts (area D2). This information patch provides a holistic view of the opinions of the crowd and is expected to provide a quick, high-level summary for the participant. We expect top forecasters to scrutinize area D.

Hypothesis 2. Top forecasters spend more time in area D compared to non-top forecasters.

Top forecasters are then expected to explore the Analysis Page (area E), as well as the Source Links Page (area F), as these information patches that contain targeted and relevant information that complements the crowd rationales. Again, top forecasters are expected to be more diligent and spend more time in these sections, compared to non-top forecasters. The top forecasters may also revisit My Forecasts Page (area G), as they want to reflect on the prior forecasts, synthesis learnings, and adjustments they have made. Thus, we would posit that:

Hypothesis 3. Top forecasters spend more time in area E compared to non-top forecasters.

Hypothesis 4. Top forecasters spend more time in area F compared to non-top forecasters.

Hypothesis 5. Top forecasters spend more time in area G compared to non-top forecasters.

Top forecasters, more than non-top forecasters, are expected to conduct external research on the internet (e.g., base rate datasets, historical event analogs, official statistics, domain-expert analysis, etc.) to validate whether the crowd's arguments and their own premises are warranted (Karvetski et al., 2022). More importantly, top forecasters are expected to engage in a cycle of seeking evidence, fine-tuning their beliefs, and refining their forecasts. Depending on the complexity of the question, we would expect that top forecasters spend a significant portion of time outside of the platform for external research.

Hypothesis 6. Compared to non-top forecasters, top forecasters spend more time on external information sources outside the platform

H1 to H6 addressed predictions for the allocation of attention across on- and off-platform areas. We also predict execution dynamics to impact forecasting performance. We expect that, throughout the entire duration of the forecast session, top forecasters will revisit parts of the platform, potentially the crowd rationales and aggregate analytics (areas C and D), as they try to isolate and synthesize relevant information before updating forecasts. This is expected especially if there are an abundance of crowd rationales. As noted above, top forecasters are also expected to engage in meaningful external research. Non-top forecasters, in contrast, are expected to browse more quickly through areas A to G once, with limited or no revisits to any of the sections and conduct minimal external research. The differences in behavior should be reflected in the total time spent per forecasting session (Atanasov & Himmelstein, 2023; see also Mellers, Stone, Atanasov, et al., 2015).

Hypothesis 7. Top forecasters, on average, spend more time per forecasting session compared to non-top forecasters

Top forecasters are also expected to visit the forecasting platform and/or update their prior forecasts more frequently than non-top forecasters. Prior work has highlighted that "belief updating" is essential to forecasting effectiveness (Csaszar & Laureiro-Martinez, 2018; Kapoor & Wilde, 2023).

Hypothesis 8a. Top forecasters update their forecasts (either numerical or verbal) more frequently than non-top forecasters

Hypothesis 8b. Top forecasters visit the forecasting platform more frequently than non-top forecasters

Moreover, top forecasters are expected to make measured adjustments, striking the right balance between under and overreacting to emerging evidence. We expect they can carefully separate mere noise from signals worth incorporating. Behaviorally, top forecasters' probability adjustment should be smaller than that of non-top forecasters (e.g., 5% vs 15%), consistent with evidence that incremental updates are associated with greater accuracy (Atanasov et al., 2020).

Hypothesis 9. The absolute value of change in forecast is lower, on average, for top forecasters, compared to non-top forecasters

Prior work suggests that top forecasters exhibit both consistency and discrimination – they make similar judgments when cases are alike and different judgments when cases are unlike (Mellers et al., 2017). Achieving such consistency is nontrivial; for instance, physicians often render divergent diagnoses for identical cases (Skåner et al., 1998). Extending this insight to forecasting on Cultivate, we expect top forecasters to show more stable, repeatable patterns of attention across their behaviors both on and off-platform when making forecasts, relative to non-top forecasters.

Hypothesis 10. Top forecasters follow a more consistent forecasting pattern across sessions than non-top forecasters

Methodology

The objective of this pilot study is to provide an initial analysis of how top forecasters and non-top forecasters differ in forecasting behavior on Cultivate. Specifically, we investigated whether the predicted forecasting behaviors were supported by real-time eye tracking and screen tracking. Participants' eye movements were recorded, and we collected a) webcam-based approximate gaze traces (i.e., where participants looked), b) fixation/dwell through time-on-area derived from screen recording (i.e., how long each area was on-screen while participants were engaged), and c) on-screen content (which on or off-platform area was viewed), all time-aligned to the session video and mapped to areas

of attention[3] (see Table 1 for an overview of every AOA). Our goal for the study is to provide descriptive and exploratory research to motivate larger-scale investigations.

Table 1: *Definition of Each Area of Attention*

Area of attention	Definition	Corresponding AOI
Landing Page	A page that lists all active forecasting questions and their current status, providing quick entry into each question; the layout reflects the forecaster's display settings.	Area A
Forecast Page	A page that lets the forecaster draft, edit, and submit a probability estimate and rationale for a specific question. Details of the respective question are also provided.	Area B
Activity Page	A page that shows recent crowd forecasts and participants' rationales for the question, enabling review of peer reasoning and discussion.	Area C
Crowd Forecast Page	A page that visualizes the aggregate crowd estimate, including its distribution and a time-series trendline of changes over time.	Area D
Analysis Page	A page that summarizes diagnostics for the question, such as forecast distributions and participant-level update histories.	Area E
Source Links Page	A page that lists curated external sources and references from Activity Page relevant to the question	Area F
My Forecasts	A page that archives the forecaster's prior estimates and comments for the question, enabling review and comparison across updates.	Area G
External Research	An activity bucket for off-platform work that directly contributes to the forecast, such as external research (reading, calculations, etc.).	
Review Closed/ Pending Questions	A series of pages that lists resolved and awaiting-resolution questions outside the current active question.	
Miscellaneous	An activity bucket for auxiliary tools and tasks (e.g., translation) support the formulation of a forecast.	
Switch Questions	A session marker that records when the forecaster changes the active question during a session.	
Set Reminders	A session marker that records when the forecaster schedules a follow-up reminder to revisit or update a question after a chosen interval.	

Table 1 presents the complete behavioral coding taxonomy used to classify all on- and off-platform activity. Hypothesis 1 through 10 address attention allocation across areas where prior forecasting literature motivates behavioral expectations. The remaining behavioral categories (i.e., Review Closed/Pending Questions, Switch Questions, Set Reminders, and Miscellaneous) were included as part of the analysis to ensure a complete record of forecasting behavior. Although no a priori hypotheses were formulated for these additional categories, they are reported in the results, as they emerged as informative signals that differentiated behavior across performance groups.

[3] Area of attention (AOA) is used as an umbrella term to describe all AOIs (areas A to G) on Cultivate, as well as other categories of behaviours (e.g., external research, setting reminders, switching questions, engaging in miscellaneous tasks) that are monitored using screen recording and/or eye tracking

Participants

Ten students from a highly selective graduate-level study program (out of 55 students) in Switzerland were randomly selected to participate in the study. Out of ten randomly selected participants, one was unable to participate; as such, one replacement was randomly selected. All participants had no prior forecasting experience and received CHF 400 (approximately USD 420 at the time of data collection) each as compensation for their time, and a consent form (see Appendix A) was signed before the study. For this study, we used the *two* highest-accuracy participants as the top forecasters subgroup. To contrast the differences between top forecasters and non-top forecasters, we categorized remaining participants as average and bottom forecasters to support our analysis.

Design and Procedure

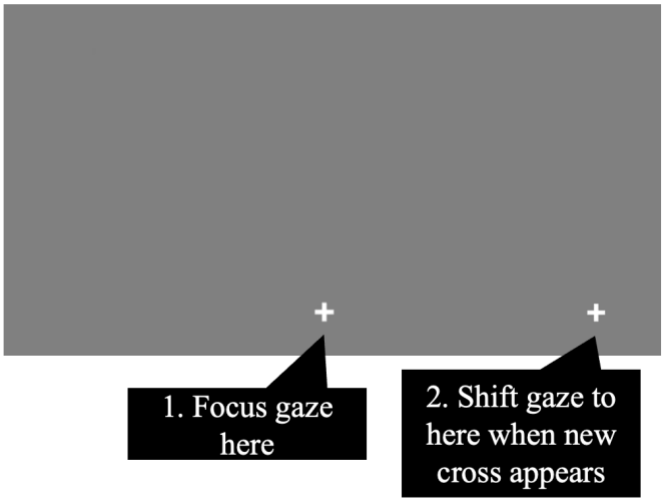
The forecasting tournament was part of every participant's course requirement and required relatively little additional effort to participate in the eye-tracking study. The study period was between October 7, 2022, and November 28, 2022. Prior to entering the Cultivate platform, all students, including the participants, were given the six forecasting questions (see Appendix B), and asked to pre-populate a response independently, without any crowd influence.

Whenever the participants chose to access the forecasting platform or conduct activities related to forecasting, the participants were asked to use a special link that activates the eye-tracking function of the study. Upon clicking the special link provided, participants were directed to a webpage provided by iMotions to calibrate their individual webcam to be used as an eye tracker. The calibration was done both before the participant's session on the forecasting platform, as well as after the participant's session, with the goal of calibrating the eye tracker to the participant's viewer position prior to and after visiting the platform to ensure accurate eye position data.

The calibration was a brief procedure in which the participant watched and followed a dot move to four corners and a center area of the screen. Each five-point dot calibration took less than 60 seconds. The participant was then directed to the forecasting platform

after the initial calibration and asked to return to the calibration page once again after completing their session on the forecasting platform (see Figure 8).

Figure 8: *Illustration of Eye Tracking Calibration*



Note: Text bubbles are for illustration purposes only

Beyond a pre- and post- calibration through iMotions, there was no difference in the workflow (e.g., access to computer trackpad to scroll through the website, ability to browse other web pages simultaneously) between participants in the study and other non-study students, indicating a high level of external validity. Basic demographic information, including age and the participant's field of study, were collected on a voluntary basis[4]. The participants were paid and thanked at the end of the data collection period.

Eye-Tracking Device

The external or internal webcam of the participant's computer or laptop was the unobtrusive eye-tracking device used. The eye tracker functions remotely from the user and does not require the removal of glasses or contacts or wearing additional equipment; thereby, allowing for observation in a natural environment. The webcam-based eye-tracking software from iMotions was used for the experiment, given its relative ease of

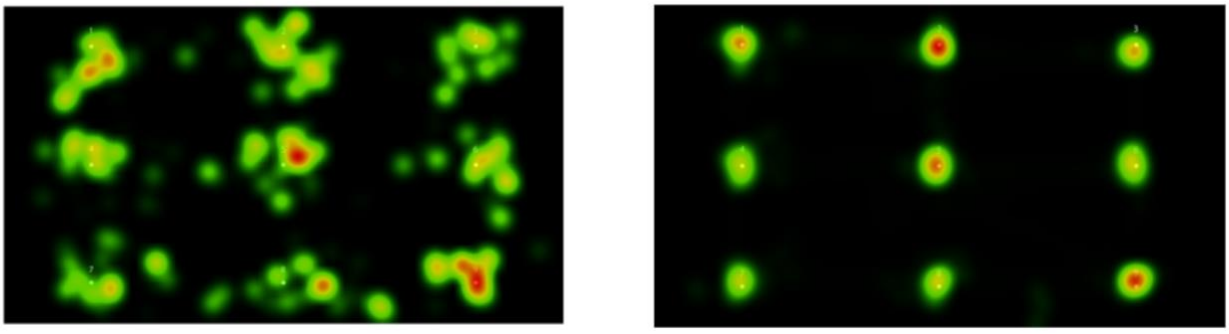
[4] Out of the 10 participants, only two students provided voluntary demographic information. The analysis of demographics was not included in this study.

access and ability to record multiple participants over long periods of time in a naturalistic setting.

The underlying software algorithm of the webcam-based eye tracking is expected to pinpoint the location of a participant's gaze within an average of 2-3 dva (degrees of visual angle) under ideal conditions (Figure 9 left), compared to 0.6 dva for a screen-based unit[5] (Figure 9 right) (iMotions, 2022). As these claims are tested under laboratory conditions, the real-world accuracy was believed to be lower than expected. In comparison to a screen-based eye tracker where the location of the gaze is tracked with finer precision, a webcam-based eye tracking device can only pinpoint a general area of fixation. Nevertheless, webcam-based eye tracking was selected for its ecological validity despite reduced accuracy compared to laboratory-grade eye-tracking gear. Given our goal of conveniently capturing authentic user behavior in a naturalistic environment, this compromise was acceptable, though limitations in tracking precision are fully acknowledged.

Because the webcam-based setup required no worn equipment, it was relatively unobtrusive and participants were able to move freely during the study. Eye tracking provided continuous measures of the participants' reactions as it records participants' eye movements throughout the entirety of every session.

Figure 9:
Accuracy Difference: Webcam vs. Screen-Based Eye Tracker



Note. Webcam (left), Screen-based (Right)

[5] A dva of 3 is approximately 4cm of deviation from real gaze location on a 15" monitor.

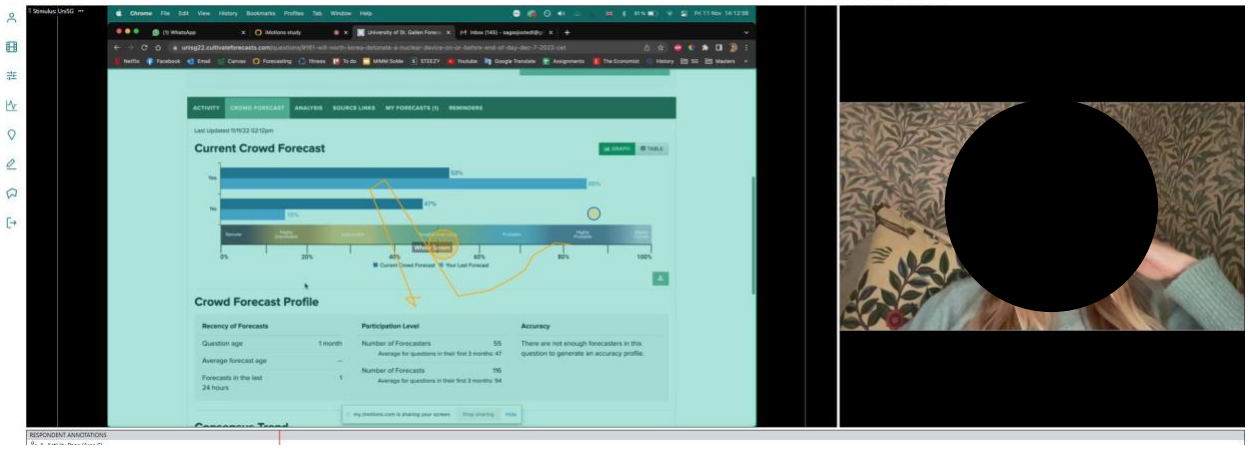
Measurements

We collected three main data points. First, for each session, we recorded webcam-based approximate gaze location synchronized to screen recordings. In practice, the webcam data did not provide sufficient precision to pinpoint fixations on specific on-screen elements. As such, we utilized gaze location primarily as a sanity check alongside the time-stamped screen recordings (e.g., to confirm that the participant was visually engaged during a page visit and to detect when transitions occurred between pages, rather than to infer detailed gaze patterns).

Second, to quantify attention, we paired the time-stamped screen recording with page tracking and computed dwell time for each AOA (e.g., Activity, external research, etc.). In other words, we measured how long each area was visible on-screen during each session, rather than word or sentence-level gaze. At this granularity, we operationalized attention as time spent in each AOA. Consistent with prior work, long dwell time serves as a coarse proxy of attention allocation and cognitive processing (Just & Carpenter, 1980; Orquin & Holmqvist, 2018; Pan et al., 2004). Using these measures, we characterize how participants navigated the platform and where attention settled during each session.

In total, approximately 35 hours of recorded data were collected for our study, of which 26 hours and 53 minutes were used for analysis. Processing, annotations and analysis were performed in iMotions 10, a commercial multimodal research platform for behavioral data collection and analysis (e.g., Hollander et al., 2020) and supplemented with forecast logs exported from Cultivate. Because webcam gaze calibration quality was consistently poor across all sessions (i.e., poor gaze calibration warning in iMotions) in this naturalistic setting (see Wisiecka et al., 2022 for comparison between in-lab vs. webcam), the different AOAs were labelled manually and derived from the screen recordings. To increase reliability, each recording was manually coded twice with random spot checks across all sessions. Please see Figure 10 for an example of a recording for one session, and Appendix C for a detailed overview of the coding process.

Figure 10: *Example of a Session Recording on iMotions 10*



Third, prediction accuracy was computed as a time-weighted Brier score[6] across six questions, whereby lower scores indicate high accuracy (Schuler et al., 2025). For group contrasts, the two highest-accuracy participants formed the top-performer subgroup; remaining participants were used as comparison groups (average forecasters and bottom forecasters).

Analysis

Comparison Between Forecaster Groups

The data for the AOAs, encompassing on-platform AOIs identified (i.e., A, B, C, D, E, F, G) and additional measurements (e.g., external research), were computed using screen recording. Specifically, whenever a participant was browsing one of the areas, clearly shown in the screen recording, the area was coded correspondingly. As the precision of eye tracking was deemed poor across all recordings and much worse (i.e., error much higher) than the expected 3 dva, areas B1, B2, and B3, as well as areas D1 and D2 were grouped together as area B and area D, respectively, for analysis. Areas B1, B2, and B3 were combined due to spatial proximity and the conceptual overlap in users' cognitive tasks (i.e., understanding the forecasting question [B1] naturally integrates with planning how to submit an answer [B3], facilitated by the ability to access different information patches [B2]). Similarly, areas D1 and D2 were combined because both sections offer complementary aggregate information about crowd forecasts. While none of the

[6] The time-weighted Brier score for our participants was computed by a colleague and not a topic of discussion for this paper.

hypotheses were framed at this sub-area level, within-area distinctions are not identifiable from this dataset, and certain nuances can be lost. All data were labelled twice manually to reduce human errors. Spot checks were randomly conducted for each session.

By analyzing gaze duration data, we calculated both total time and proportional time allocation (as a % of total time) for all participants, enabling behavioral comparisons between groups for each AOA. We utilized proportional analysis as it controls for individual differences in total platform engagement. Participants were expected to naturally vary in session length due to personal factors (e.g., reading speed, technical proficiency, etc.); these individual baselines (i.e., total amount of time) are less informative than how participants strategically distribute their time. For example, two participants might spend different absolute durations in Activity area (30 mins vs 60 mins), yet both allocate 30% of their session there, revealing similar strategic priorities. This method of analysis supported the examination of H1, H2, H3, H4, H5, H6, and H7.

As the required eye-tracking calibration through iMotions prior to every forecasting session is not an innate habit for the participants, participants may forget to set up the eye-tracking and screen sharing. By combining both screen recording data from iMotions and forecasting history from Cultivate, we determined how often each participant made a new forecast, provided new rationales and/or visited the platform. From the record on Cultivate, the number of updates (both numerical and verbal) per question were tallied. The average number of updates for top forecasters compared to non-top forecasters was compared, thereby, answering H8a and H8b.

Using the same data set, the magnitude of each forecast update (i.e., numerical update) was then determined by computing the absolute difference between every two subsequent forecasts. Comparing the absolute value of differences between forecasts for top forecasters and non-top forecasters addressed H9.

To examine forecasting consistency (H10), we summarized how often a participant concentrated their time in the same parts of the platform across sessions. For each session, we leveraged calculations on the percentage of total session time spent (from earlier hypotheses) in every predefined AOA and ranked these areas by that percentage to identify

the most-time-spent (T1), second-most-time-spent (T2), third-most-time-spent (T3) areas. For each participant, we then calculated a repetition rate for each rank (i.e., the share of that participant's sessions in which the same area held that rank). For example, if area B (Forecast Page) received the most time in four out of six sessions, the T1 repetition rate is 66.7%. We also recorded which area occupied each repeated rank to clarify where time tended to concentrate. This rank-based summary treats consistency as repeating one's own time allocation template (i.e., prioritizing the same top areas across sessions), and allowed us to compare how structured or flexible session focus was across the different groups.

Forecast Page (Area B).

The Forecast Page area was not preregistered as a hypothesis because, initially, we treated it as the execution interface for providing forecasts. However, the data showed substantial, heterogeneous time investment in area B across all participants. Consequently, we included analyses of area B to contextualize how gathered information translates into edits and updates, and how this translation, through area B, shapes overall forecasting behavior.

Results and Discussion

To highlight the contrast between different subsets of forecasters (i.e., top forecasters, average forecasters, as well as the bottom forecasters), we selected seven participants (out of the ten recorded) for analysis. For the seven participants, performance rankings ranged from 1st to 54th place out of 55 total forecasters, which coincidentally represented a near perfect sampling of the performance distribution.

Data collection yielded 44 total sessions across seven participants, resulting in approximately 1612 minutes, or 26 hours and 53 minutes, of usable data (see Appendix D for a breakdown of collected data). Notable data limitations include one unusable session for participant TF2[7] totaling 34 mins and two sessions for BF2 totaling 57 minutes, due to failure to enable screen sharing permissions. Moreover, TF2, BAF1 and BF1 each faced an instance of session recording upload failure to iMotions' cloud; the duration of the

[7] TF = top forecaster/performer, BF = bottom forecaster/performer, AF = average forecaster/performer, and BAF = below-average forecaster/performer

recordings is unknown. Lastly, AF1 and BF2 omitted the required steps for enabling eye-tracking for one and two sessions, respectively; the duration of these visits is also unknown.

Overview of On and Off-platform Behaviors

Comparison Between Top Forecasters (TF1 and TF2)

First, we must examine the top forecasters closely and isolate what the similarities or differences are between those who belong to this subgroup. Despite both TF1 and TF2 achieving elite forecasting accuracy, their information foraging behaviors differed markedly. TF1 spent nearly half of their gaze time (49%) performing external research, making them a heavy off-platform information gatherer. TF1 complemented this with substantial time (33%) editing and/or reviewing their own forecast (area B). As such, TF1 spent more than 80% of their time in these areas (i.e., patches) that they deemed of high value and demonstrated effective integration across information patches. This pattern reflects what we could term a central-place forager[8] – using the platform as a home base while making extensive forays into outside sources for validation, fact checking, and specialized knowledge. TF1 spent approximately 12% of their time in the Activity section (area C), absorbing the insights of others. The total amount of research, both external and on-platform, surpassed 60% of total time. Beyond these three buckets, the time investment in all other AOAs is marginal. Overall, TF1 was heavily research-focused, paying particular attention to information patches through external, off-platform, research.

TF2, in contrast, allocated a substantial proportion of their time (47%) directly contemplating and self-editing within Forecasts Page (area B), supported by some referencing of their past predictions (My Forecasts, area G, 7%). These patterns indicate that TF2 takes an iterative approach in continuously reflecting on and refining previous forecasts. The 7% of time reviewing old forecasts (area G), compared to TF1's approx. 2%, suggests that TF2 employs a unique strategy whereby they use their own historical predictions as an additional information patch for information gathering. Albeit more than

[8] In behavioural ecology, central-place foraging theory describes animals (e.g., birds) that collect resources and then return to a central place (nest or hive). Information foragers can sometimes act as "central-place foragers" when they return repeated to a main hub (e.g., search page or dashboard) after exploring other resources.

TF1, TF2 also spent a modest amount of their time (17%) on the Activity (area C) section, highlighting that their judgment was largely influenced from within-platform insights.

Unsurprisingly, TF2 spent considerably less proportion of total time, roughly 9%, on conducting external research compared to TF1 who spent almost half their time. The total amount of research, both external and on-platform, only accounted for 26% of TF2's time and is significantly lower than the 60% of TF1. The large difference in research time between TF1 and TF2 indicates a divergence of strategy. TF1 seems to focus heavily on thorough gathering information across different and mostly external patches to articulate a forecast, whereas TF2 seems to focus on understanding the gist of information across crowd forecasts (area C) to cleverly formulate and refine forecasts. These distinct archetypes demonstrate that multiple effective foraging strategies can co-exist, extending IFT's framework by suggesting that optimal information gain may be achieved through several locally optimal strategies adapted to user goals and environmental constraints (Pirolli, 2007).

Both top forecasters employed the use of reminders and were able to flexibly switch between different questions during a single session, indicating a higher level of cognitive control and adaptability.

Top Forecasters vs. Bottom Forecasters Group (BF1, BF2)

While top forecasters focused mainly on some form of research (external or on-platform crowd), coupled with editing and reviewing their forecasts, bottom forecasters showed substantial heterogeneity, reflecting multiple potentially maladaptive foraging behaviors.

BF1 dedicated 53% of total session time exclusively to editing and articulating personal forecasts (area B), surpassing TF2's 47% and TF1's 33% investment in forecast formulation. Moreover, BF1 spent 24% of their time on external research, and 15% of time on within-platform crowd research. Interestingly, the total amount of research time (both off and on-platform) accounted for roughly 40% of total time investment, which lies between the time investment of TF1 and TF2. These behavioral tendencies indicate that while BF1 may have understood what was necessary to be a successful forecaster, the

execution of these tasks is unproductive. For example, BF1's excessive editing of forecasts reflects an overgrazing pattern – remaining within a single information patch well beyond the optimal timeframe and potentially violating IFT principles on timely patch abandonment, as well as over-exploitation of a single patch (Pirolli & Card, 1999; Pirolli, 2007). BF1's behavioral tendencies likely indicate the inability to integrate information across multiple patches. The screen recording hinted at excessive rumination, characterized by typing, deleting, and retyping, during forecast composition in area B. The qualitative contrast between TF2 and BF1 in area B is noteworthy: while TF2's 47% of time investment in forecast formulation reflects productive iteration with systematic refinement and information integration, BF1's 53% of time investment reflects more unconstructive repetitive processing that may impede synthesis (Nolen-Hoeksema et al., 2008; Watkins, 2008), indicating potential analysis paralysis, rather than synthesis (e.g., Iyengar & Lepper, 2000).

In stark contrast, BF2 spent an extreme 93% of their time within the Activity section, engaging almost exclusively with crowd rationales. This pattern exemplifies an over-exploitation of a high information scent patch; the strong perceived informational value of crowd rationales sustained attention well beyond the point of diminishing returns (Pirolli, 2007). The Activity area, at least beyond a certain point, can operate as a "supernormal stimulus" (Tinbergen & Perdeck, 1951), whereby the continuous stream of peer rationales prevents the natural decline in returns that would normally trigger patch abandonment, transforming BF2 from an active information forager into a relatively passive information consumer. Additionally, BF2's negligible investment in forecast generation (2%), external research (3%), as well as absence of any other value-added behaviors (e.g., setting reminders) reinforces a suboptimal strategy consistent with IFT's predictions regarding overcommitment to a single high-scent patch despite declining informational yield.

Ultimately, the two bottom forecasters, BF1 and BF2, were ineffective at forecasting in fundamentally different ways. BF1's poor performance appears to stem from excessive rumination and fixation on forecast formulation, coupled with limited integration of insights from either external sources or peer rationales. In contrast, BF2's ineffectiveness resulted from overcommitment to a single high-scent patch (Activity, area C), where

sustained engagement with crowd rationales supplanted original analysis and synthesis. In IFT terms, BF1 exemplifies overexploitation within one patch, in which we later characterize as integration and synthesis failure, whereas BF2 represented entrapment by an overly salient information source (a form of patch selection failure). Both patterns underscore how deviations from balanced, exploratory foraging can impede the development of accurate, well-calibrated forecasts. Moreover, while recent research has shown that deliberation time is moderately correlated with accuracy (Atanasov & Himmelstein, 2023), our findings suggest that excessive deliberation is associated with lower prediction accuracy.

Top Forecasters vs. Expanded-Bottom Forecasters Group (Extending to BAF1)

Given the diversity of results shown among bottom forecasters, we included an additional below-average participant (BAF1) to further refine our understanding of non-top forecaster performance. Like TF1, BAF1 displayed heavy external research behavior (41%). Additionally, comparable to TF1 and TF2, BAF1 allocated a moderate portion of time to Activity (14%). However, BAF1 devoted comparatively less time to iterative composition and refinement of personal forecasts (29% in area B) but showed modest engagement with the Analysis section (area E), an area rarely used by top forecasters. Notably, BAF1 did not employ any forecast management tools (e.g., switch questions, reminders, etc.).

Despite a seemingly productive distribution of time across patches, BAF1's behavior yielded limited benefit. Notably, they invested substantial effort both outside platform (41% in external research) and within it (14% in Activity area C) – a pattern resembling top forecaster-level engagement. Yet this extensive information gathering produced minimal returns in forecasting accuracy. The result is high engagement without commensurate information yield, a form of extended foraging where effort is expended across diverse sources but not sufficiently synthesized into actionable insight.

This paradox underscores that breadth of information access (internal or external) is insufficient without the cognitive integration necessary to transform diverse inputs into coherent judgments. As such, BAF1's behaviors fundamentally challenge our understanding of forecasting expertise, suggesting three possible explanations: 1) BAF1

may recognize behaviors associated with forecasting success but do not understand their functional purposes; 2) they may have collected information effectively but struggle to integrate or synthesize information for unified reasoning (i.e., integration failure); and/or 3) attentional, confirmatory, or other biases that may limit the extent of information processed.

Including BAF1 amongst the bottom forecasters temper some of the stark contrast observed with BF1 and BF2, revealing that underperformance can arise not only from maladaptive overfocus, but also from inefficient integration of otherwise adaptive behaviors. A central insight emerges and highlights that forecasting success depends less on the intensity within any single patch and more on balanced, integrative, and diverse cross-patch foraging.

Top Forecasters vs. Average Forecasters Group (AF1, AF2)

Comparing average forecasters (AF1 and AF2) against top forecasters provides hints into behavioral nuances that distinguish between "acceptable" forecasters from exceptional ones. At a broad level, both groups demonstrated comparable proportional engagement with high-utility areas, including forecast formulation (area B), crowd rationales (area C) and external research. This suggests that average forecasters, like top forecasters, possess an implicit understanding of which patches yield informational value and the importance of multi-source information gathering.

The quantitative similarities are striking – average forecasters spent 29.3% of their time on external research compared to 28.8% for top forecasters, and 41.3% on the Forecast page (area B) compared to 40.3% for top forecasters. These near-identical allocations indicate that average forecasters successfully identified where to invest time but may lack the deeper cognitive processing required to extract and synthesize value from it. As a group, the average forecasters show a similar pattern of behavior when compared to our below-average forecaster, BAF1.

Interestingly, average forecasters devoted proportionally more time to the comparatively lower-yield Activity patch (23% of total time) than top forecasters (14%), representing roughly a 65% increase; nevertheless, time investment is significantly less

than all other groups (lower performers from approx. 41% to 54%). Although far from BF2's extreme focus on crowd rationales (93%), this elevated focus may reflect some difficulty in filtering or prioritizing peer input amid potential information overload (Eppler & Mengis, 2004).

The comparative overinvestment in peer rationales could reflect an overgeneralized wisdom of the crowds assumption whereby increased consultation with the crowd necessarily improves calibration (Lorenz et al., 2011; Surowiecki, 2004), particularly given evidence that social influence can reduce independence and degrade crowd accuracy (Larrick & Soll, 2006). Average forecasters appear caught between two extremes: they avoided BF2's trap of total crowd dependence but did not yet achieve the selective extraction of unique, high-value insights seen amongst top forecasters. Instead, they seem to consume rationales more exhaustively rather than strategically.

Moreover, the average forecasters diverged meaningfully in how they approached their tasks. AF1 displayed relatively limited platform engagement overall, relying sparingly on internal tools and allocating minimal time to exploration, suggesting an efficiency-oriented, but somewhat shallow foraging strategy. In contrast, AF2 was highly active, engaging across multiple patches within and outside the platform. Despite this, AF2's pattern appeared less selective and more exhaustive, reflecting extensive but diffuse information consumption. This divergence demonstrates that average forecasters, despite recognizing which patches hold information value, can still exhibit contrasting inefficiencies in execution, manifesting as either under-exploration (AF1) or overextended foraging without synthesis (AF2).

The near absence of time spent in the Analysis section, area E (0.01% for average forecasters vs. 3.08% for top forecasters) suggests average forecasters may either lack statistical literacy to carefully interpret the charts, or the recognition that high-level quantitative analysis could improve forecasts. This omission is particularly revealing, as average forecasters otherwise distributed time sensibly across patches, implying that the leap from average to top forecaster performance depends less on diligence and more on analytical capabilities.

Ultimately, average forecasters exemplify a knowing-doing gap; they seem to mirror the structural behaviors of success yet seem to falter in timing, synthesis and evaluative depth, exhibiting milder forms of integration and synthesis failures observed more acutely in lower performing groups. AF1 demonstrates how limited engagement constraints information gain, while AF2 reveals the correct on- and off-platform utilization without integration leads to diminishing returns. Together, these patterns underscore that forecasting success hinges not merely on the quantity or distribution of patch engagement, but on the quality of cognitive integration that transforms foraged information into well-calibrated judgment. See Table 2 for a comparison between groups in every AOA.

Table 2: *Breakdown of Time Investment (% of Total) for Each Area of Attention*

Areas of Attention (% of total time)	**Top Forecasters (*n* =2)**	**Average Forecasters (*n* = 2)**	**Δ**	**Exp. Bottom Forecaster (*n* = 3)**	**Δ**	**Bottom Forecasters (*n* = 2)**	**Δ**
Landing Page (Area A)	0.86	2.39	-1.53	0.50	0.36	0.50	0.36
Forecast Page (Area B)	40.32	41.30	-0.98	28.06	12.26	27.78	12.54
Activity Page (Area C)	14.30	23.32	-9.02	40.86	-26.56	54.08	-39.78
Crowd Forecast Page (Area D)	1.05	1.28	-0.23	0.64	0.41	0.34	0.71
Analysis Page (Area E)	3.08	0.01	3.07	4.78	-1.70	1.68	1.40
Source Links Page (Area F)	0.00	0.00	-	0.00	-	0.00	-
My Forecasts (Area G)	4.53	2.19	2.34	1.88	2.65	1.50	3.03
External Research	28.80	29.33	-0.53	22.71	6.09	13.44	15.36
Review Closed/Pending Questions	6.70	4.36	2.34	0.00	6.70	0.00	6.70
Miscellaneous	0.13	0.03	0.10	0.56	-0.43	0.69	-0.56
Switch Questions	0.15	0.08	-	0.01	-	0.00	-
Set Reminders	0.09	0.07	-	0.01	-	0.01	-

Note: Switching questions, and setting reminders are qualtiative asessments where Δ would not compute meaningful answers

Hypotheses Summary

Using mean proportion of session time across the sub-groups, we find strong support against H1: the top forecaster group spent 14.30% in Activity (area C) versus 23.32% for average-forecaster group (Δ = -9.02 pp), 40.86% for expanded bottom-forecaster group (Δ = -26.56 pp), and 54.08% for bottom-forecaster group (Δ = -39.78 pp). H1 is rejected; top forecasters avoided prolonged crowd rationale browsing, and as prediction accuracy decreases, more time is spent in area C. For H2 (Crowd Forecasts, area D), top forecasters devoted 1.05% of time to aggregate trendline and crowd forecast insights, compared to 1.28% (average; Δ = -0.23 pp), 0.64% (expanded bottom; Δ = +0.41 pp), and 0.34%

(bottom; Δ = +0.71 pp). Support for H2 is mixed and appears only for the bottom groups. H3 (Analysis, area E) shows mixed results. Top forecasters spent 3.08% of time, compared to 0.01% (average; Δ = +3.07 pp) and 1.68% (bottom; Δ = +1.40 pp), but less than 4.78% (expanded bottom; Δ = -1.70 pp), driven by the inclusion of BAF1. H4 (Source links, Area F) is not testable. Usage is non-existent across all groups. H5 (My Forecasts, Area G) is supported. Top forecasters spent 4.53% of their time, compared to 2.19% (average; Δ = +2.34pp), 1.88% (expanded bottom; Δ = +2.65 pp), and 1.50% (bottom; Δ + 3.03pp). A clear trend is shown here: as forecasting performance decreases, less time is spent in reviewing prior forecasts. For H6 (External Research), there is a pattern of parity with average-forecaster group (28.80% vs. 29.33%, Δ = -0.53 pp), but higher than expanded bottom-forecaster group (22.71%, Δ = +6.09 pp) and bottom-forecaster group (13.44%**,** Δ = +15.36 pp) groups; H6 is partially supported.

The H1 to H6 findings, coupled with the extensive overview of on- and off-platform forecast behaviour in the prior section, describe where attention is allocated across on- and off- platform patches, and how these patches work together to support accurate forecasts. H7 to H10 shifts the lens from allocation to execution dynamics, whereby we examine how forecasting unfolds over time: the average length of a session (H7), the cadence and frequency of updates and visits (H8a/H8b), the magnitude of forecast adjustments (H9), and the stability of the forecasting process across sessions (H10).

Top forecasters' sessions averaged 32.65 mins, compared with 21.17 (average, Δ = +11.48 min), 37.69 (expanded bottom; Δ = -5.04 min), and 43.94 (bottom; Δ = -11.29 min). Thus, H7 shows inconclusive results from our study, as top forecasters have shorter sessions than both groups of lower performers, but more than average forecasters. Interpreting alongside update productivity (top forecasters at 1.50 updates per visit, vs 0.86, 0.70, and 0.73 updates per session visit respectively for average, expanded bottom, and bottom group of forecasters), the pattern suggests that top forecasters avoid the very long, low-yield sessions in bottom forecaster groups while maintaining enough time to integrate evidence and act.

Top forecasters showed a more frequent update tempo per session: they made 12 forecast updates over 8 platform visits (1.5 updates per visit), compared with 0.86 for

average forecasters, 0.70 for expanded bottom group, and 0.73 for bottom forecasters. This indicates support for H8a; when top forecasters log in, sessions have prompted more forecast updates, whether numerical or verbal; similar conclusions have been drawn by other scholars (e.g., Mellers, Stone, Atanasov, et al., 2015). By contrast, H8b is not supported. Top forecasters made more platform visits than average forecasters (8 vs. 7), but less often than bottom groups (10 for expanded bottom, and 11 for bottom), suggesting quantity alone does not drive accuracy, and may instead reflect indecision or incomplete sessions that prompt return visits.

Critically, the average magnitude of forecast adjustment was smallest for top forecasters (6%) and larger for every other group (average at 8.30%, expanded bottom at 14.20% and bottom at 13.50%), consistent with claims that incremental changes in forecast updates lead to superior results due to fine-grained calibration (e.g., Baron et al., 2014; Mellers et al., 2017). H9 is supported in this study and replicates prior work.

We defined consistency as how often a participant's most-time-spent area (T1), second-most time-spent area (T2), and third-most time-spent area (T3) recurs across that participant's sessions. On this measure, H10 is not supported. Top forecasters only show moderate T1 repetition (66.7% on average), with TF1 investing most time into External Research, and TF2 for Forecast Page (area B). The top forecasters' T2 and T3 positions rotate across Forecast Page (area B), external research, and Activity Page (area C). By contrast, the expanded bottom forecasters group shows very high T1 repetition for the most time spent area (approx. 93%), and the bottom group is 100%; the repeated area is participant-specific, BF1 spent most time in Forecast Page (area B) across all sessions, whereas BF2's focus was on Activity Page (area C). In other words, lower performers are more repetitively consistent (i.e., often locked into a single preferred area), while top forecasters exhibited structured variety, whereby they cycle among a small set of productive areas (patches). Interpreted through IFT, top forecasters show adaptive foraging routines, whereas bottom forecasters show patch lock-in. We note, however, that our operationalization of consistency – the repetition of time-allocation across sessions – may capture process consistency rather than judgement consistency. For an overview of all hypothesis results, see Table 3.

Table 3: *Summary of Hypotheses and Results*

H#	Hypothesis	Result	Evidence	Group			
				Top (n = 2)	Avg. (n = 2)	Exp. Bottom (n = 3)	Bottom (n = 2)
H1	Superforecasters spend more time in area C	Not supported	Superforecasters spend the least time here, and time in this area rises steadily as performance decreases.	14.30%	23.32%	40.86%	54.08%
H2	Superforecasters spend more time in area D	Inconclusive	Superforecasters spend less time than average forecasters, but more than both bottom groups.	1.05%	1.28%	0.64%	0.34%
H3	Superforecasters spend more time in area E	Inconclusive	Superforecasters spend more time than average and bottom groups, but less than expanded-bottom (driven by BAF1).	3.08%	0.01%	4.78%	1.68%
H4	Superforecasters spend more time in area F	-	Usage is zero across all groups; not testable.	-	-	-	-
H5	Superforecasters spend more time in area G	**Supported**	Superforecasters spend the most time here, and time declines as performance decreases.	4.53%	2.19%	1.88%	1.50%
H6	Superforecasters spend more time in area on ext. research	**Partially supported**	Superforecasters are roughly on par with average forecasters, and exceed both bottom groups.	28.80%	29.33%	22.71%	13.44%
H7	Superforecasters spend more time per forecasting session	Inconclusive	Superforecasters have longer sessions than average forecasters, but shorter sessions than both bottom groups.	32.65 min	21.17 min	37.69 min	43.94 min
H8a	Superforecasters make updates more often	**Supported**	Highest cadence for superforecasters and declines with performance.	1.50	0.86	0.70	0.73
H8b	Superforecasters visit the platform more often	Not supported	Superforecasters visit more often than average forecasters, but less often than both bottom groups.	8	7	10	11
H9	Superforecasters make smaller adjustments	**Supported**	Superforecasters make the smallest adjustments, and magnitudes increase as performance decreases.	6.00%	8.30%	14.20%	13.50%
H10	Superforecasters show consistent forecasting patterns	Not supported	Repetition of the most-visited area is the highest in the bottom groups, while superforecasters only show moderate reptition.	66.67%	67.86%	93.33%	100.00%

Overall Implications for Cultivate and Information Foraging Theory

Our findings are preliminary signals from a small pilot sample (n = 7) with 44 recorded sessions. Estimates are descriptive and not intended for population generalization. We report insights as patterns that can guide future large-scale studies.

Overall, there are pockets of clear trends when looking at the results holistically. Top forecasters engaged in areas A through G for approximately 64% of their time investment, allocating roughly 29% on external research, 7% to reviewing previous questions, as well as tackling multiple questions per session and setting forecast reminders. Critically, when top forecasters did engage with the platform, their sessions were characterized by higher update productivity, whereby they made 1.5 forecast updates per session compared to 0.86 or fewer updates per session for all other groups. This pattern suggests top forecasters not only allocated their time more strategically across on-platform areas, but also extracted more value for time invested, translating information gathering into concrete forecast adjustments. The quality of these updates differed substantially; top forecasters made smaller magnitude adjustments (averaging 6% changes) compared to larger swings by other groups (average forecasters at 8.30%, and bottom forecasters at 13.50%), indicating that their forecast refinements reflected fine-grained calibration rather than reactive overcorrection.

Average forecasters spent approximately 71% of their time in area A through G, and while the average forecasters do spend similar time on phrasing forecasts (area B) and external research, they spend 63% more time on crowd rationales and 50% less time re-visiting prior forecasts. They also tended to focus on one question during each forecasting session and set reminders less often. Notably, average forecasters also demonstrated lower update productivity (0.86 updates per session) and made larger magnitude adjustments (8.30%) when they did update, suggesting they may have struggled to convert their research time into actionable insights with the same efficiency as top forecasters. These behavioral tendencies highlight that the average forecasters may understand what is required to succeed but may either lack the cognitive and/or forecasting expertise to discern between relevant information and integrate all sources of information.

Comparing against lower performers sharpens the contrast. The expanded bottom group spent almost 77% of their time in areas A to G; relative to them, top forecasters spent 44% more time focused on formulating or adjusting forecasts, 27% more on external research (outside A to G), 141% more on reviewing prior forecasts, 64% more studying the crowd trendline, and 65% less on crowd rationales. Despite more platform time, the expanded bottom group demonstrated markedly lower update productivity (approx. 0.7 updates per session) and made substantially larger forecast adjustments (14.20%), suggesting these extended sessions yielded diminishing returns, especially in the crowd forecasts area. The forecast adjustment patterns, characterized by dramatic swings rather than incremental refinements, indicate poor information integration and potentially reactive decision-making, driven by the most recent information encountered rather than synthesis across all areas.

When comparing solely against the two bottom forecasters, the pattern is even more striking. The proportion of time spent on A to G by bottom forecasters is more extreme at approximately 86%. On average, top forecasters spent more time formulating and perfecting forecasts (45% more), more on the crowd trendline (209% more), more on the analysis area (83% more), more time on reviewing prior forecasts (202% more), and less time on crowd forecasts (74% less). Moreover, beyond the platform, top forecasters spent 114% more time, proportionally, on external research, and 81% less time on miscellaneous activities, such as using Google Translate. Bottom forecasters did not switch questions within sessions and rarely set reminders. The behavioral rigidity of bottom forecasters extended beyond area selection to their temporal patterns whereby both participants showed extreme consistency. The same area was the dominant focus point in 100% of their sessions; in contrast, top forecasters adapted their approach across sessions, suggesting that they employ flexible, context-dependent strategies rather than formulaic routines.

Our data reveals a striking pattern: more on-platform browsing does not translate to better forecasts. On-platform usage (areas A through G) rises as accuracy falls, from 64% (top forecasters) to 71% (average forecasters) to 77% (expanded bottom forecasters) to 86% (bottom forecasters only) yet update productivity and accuracy drop. This inverse relationship suggests that Cultivate may inadvertently trap lower performers in

unproductive consumption patterns. The issue is not insufficient engagement (i.e., total time investment), but misallocated distribution of attention, as lower performers forage longer in comparatively low profitability patches. Top forecasters, in contrast, demonstrate selective engagement by effectively exploring high-value patches, often shifting to external validation and, crucially, engage in synthesis – an integration of disparate pieces of information into a coherent forecast and a key facet of intelligence[9] (e.g., Bateman et al., 2019; Oberauer et al., 2008).

As prediction accuracy drops, we see distinct failures modes of late patch exits and weak conversion of information into forecast updates. On-platform patterns include excessive rumination in forecast formulation (area B) and being trapped in crowd rationales (area C), indicating a potential failure to integrate, synthesize, and convert various information into new forecasts. These maladaptive patterns are accentuated by decreased external research, abandonment of supportive tools (e.g., reminders to forecast), and single question focus per session. This heterogeneity suggests poor performance stems not from a single deficit, but rather various obstacles during the entire process: patch selection failure (overconsumes single or only a few sources), integration failure (cannot process multiple pieces of information), or synthesis failure (cannot convert information into relevant forecasts). Moreover, bottom forecasters demonstrate behavioral rigidity, maintaining their specific maladaptive pattern across all sessions rather than adapting behavior, indicating a lack of metacognitive flexibility to recognize the need to adapt strategies, as all participants were given updates on their performance periodically.

Nevertheless, the behavioral tendencies of top forecasters partially support IFT's predictions about optimal foraging, as top forecasters demonstrate strategic patch selection and timely abandonment. However, our preliminary findings challenge IFT's traditional emphasis on comprehensive information gathering, revealing that in forecasting contexts, selective efficiency outperforms exhaustive thoroughness. This unexpected finding – that less can be more in information consumption – has profound implications for both platform

[9] As a proxy for intelligence, we report the Strategic Management midterm (mix of knowledge and applied questions) scores of participants, where results may suggest higher intelligence in top forecasters compared to non-top forecasters (32 vs. 28.8 out of 36). We note overlap in scores (e.g., a non-top forecaster had the highest exam score, and a bottom forecaster matched a top forecaster). With $n = 7$, we report this only descriptively.

design and forecaster training. Leveraging these initial findings, we offer concrete suggestions for Cultivate to potentially boost forecaster performance.

Practical Implications for Platform Design

We suggest several design modifications that could enhance information foraging, integration and synthesis across different forecaster groups. Our UI/UX suggestions aim to tune users' information diet, strengthening information scent for high-yield patches, and support patch-foraging decisions that reflect patch profitability (Charnov, 1976; Pirolli, 2007; Pirolli & Card, 1999). The recommendations are grounded on our explicit assumption that there is a platform objective of improving individual forecasts to lift the aggregate (e.g., Budescu & Chen, 2015; Mellers et al., 2014), rather than optimizing the crowd forecast calculation only (e.g., weighting, selection, and/or aggregation methods).

Strengthen Information Scent for High-value Patches

Currently, all menu headings (Activity, Crowd Forecast, Analysis, etc.) appear visually identical despite vastly different usage by top forecasters. Cultivate should visually emphasize AOAs showing highest profitability based on our data (Wolfe & Horowitz, 2017). For example, the Crowd Forecast trendline (area D), which top forecasters visit 209% more than bottom forecasters, could feature a subtle highlight or icon (e.g., "Highly recommended") indicating if the user has not yet accessed this area after a few logins. The argument can be made for the Analysis Page (area E). Conversely, the crowd forecast section (area C), where bottom forecasters spend excessive time, could display a muted appearance (i.e., scent decay) after initial viewing, signaling depleted patch value (Charnov, 1976; Pirolli & Card, 1999; Pirolli, 2007)

Show "Diminishing Returns" Explicitly

After a user reads more than a pre-determined proportion (e.g., 50-75%) of the available rationales in the Activity section, Cultivate could display a small counter showing "x out of y rationales viewed", or offer even more tailored suggestions (e.g., "You have reviewed 50 rationales, 90% of top forecasters transition to external research at this point") (e.g., Oinas-Kukkonen & Harjumaa, 2009). To complement the notification, the

background color of already-read rationales should gradually fade, creating a visual depletion effect (Wolfe & Horowitz, 2017).

For excessive rumination when formulating on the Forecast Page (area B), as observed with BF1's typing-deleting-retyping pattern, the platform could detect when users have revised the same sentence three times without substantial changes and display a gentle prompt (e.g., "Your current forecast is x%. Consider gathering additional information or reflecting on findings before refining further"), with quick links to Google Gemini/ChatGPT Deep Research or crowd forecast section (Oinas-Kukkonen & Harjumaa, 2009; Weinmann et al., 2016). Additionally, even more personalized phrasing such as "You have spent 30 minutes trying to construct your forecast, perhaps it's time for a breather and visit other sections!" can be added. These interventions could interrupt the rumination cycle and redirect toward information gathering or further reflection rather than perpetual editing.

Create Visible Foraging Trails

Rather than abstract visualization, Cultivate could implement a simple session history bar (i.e., foraging trail) at the top showing the user's navigation history (e.g., side bar with headings that light up in different colors for visited, not yet visited, or excessively visited areas) (Oinas-Kukkonen & Harjumaa, 2009; Weinmann et al., 2016). Depending on data availability, the side bar can be compared to an ideal top forecaster pattern beside it. When users deviate significantly, such as narrow information diet (i.e., no external research), the comparison could highlight this divergence. Additionally, browsing patterns of top forecasters could be suggested through subtle arrows on the menu bar as "next step" recommendations based on current location and time spent (Weinmann et al., 2016).

Implement Exit Interventions for Areas Missed

When users attempt to close Cultivate, the platform could highlight unvisited high-value areas (e.g., with arrows) and prompt them to set reminders for next forecasting session (Einstein & McDaniel, 1990; Weinmann et al., 2016) – behaviors consistently observed in top forecasters but often absent in average and bottom forecasters. This intervention would promote comprehensive patch exploration and scheduled revisitation, maximizing one's chance for high forecasting accuracy.

Additional Area-specific Recommendations

Moreover, structured templates for extracting key insights from each area could address a critical gap identified in our analysis. Given that average forecasters (AF1, AF2 and AF3) demonstrate similar patch visitation patterns to top forecasters – spending comparable time on external research (approx. 29% for both) and forecast formulation (41% vs 40%) – yet achieve inferior accuracy, we argue their deficit lies not in knowing where to forage but in how to extract and integrate insights. This gap suggests that passive consumption of information, even from correct patches, fails to generate forecasting improvements without value generating synthesis. Therefore, implementing behaviors that could transform passive information gathering into tangible insights may lead to better forecasting accuracy.

Activity section (Area C)

After reading a certain number of rationales (e.g., 30-50%), a pop-up template could prompt: "Before continuing, identify: 1) the strongest argument for your prediction, 2) the strongest argument against, 3) any base rate or historical precedent mentioned, and 4) one perspective that was not considered (Lord et al., 1984; Heuer, 1999). This would force users to actively process and synthesize rather than endlessly scroll (e.g., BF2). This new mid-session template could save these insights to a workspace (i.e., a potential new tab) for later integration.

External Research Departures

When users return from a long period of external research (e.g., threshold at >15 mins), a re-entry form could ask: What did you find out? 1) List 2-3 facts or statistics, 2) Does this confirm or contradict the crowd consensus, and 3) What uncertainty is remaining for you (Heuer, 1999; Oinas-Kukkonen & Harjumaa, 2009)? This would help BAF1-types who conduct extensive research, but fail to integrate findings, to crystallize their discovery into actionable insights rather than accumulating, or being bombarded by, unfocused information. Again, this could be insights saved to a workspace tab for mind mapping and integration.

Crowd Forecast Trendline (Area D)

To support forecasts in extracting key information from the trendline (Cleveland & McGill, 1984), a simple tool could prompt: "The crowd currently predicts X%". Based on the trend, Cultivate can ask e.g., 1) is the consensus strengthening or weakening? 2) When did the major shift occur? 3) What might have triggered this change? The availability of prompts can transform passive observation into active pattern recognition.

Forecast Page (Area B)

Given the domination of time investment in the Forecast Page, and the potential difficulties in deciphering signal from noise (e.g., Kahneman et al., 2021; Silver, 2012), the platform could implement a questionnaire before allowing extended editing. For example, questions such as "Before writing your rationale: 1) state your numerical forecast, 2) list your top 3 reasons, and 3) identify your biggest uncertainty, 4) what evidence would change your mind?". This could prevent BF1's endless rumination, as well as encourage information synthesis and critical thinking by forcing commitment to specific claims before wordsmithing begins. Alternatively, we can consider a lightweight "Update Assistant" that can surface a forecaster's last two forecasts (area G), if applicable, and encourage small, frequent revisions with a slider (e.g., offer options of 1% to 5%).

The above question templates could potentially accumulate in a new tab (e.g., Insights), creating a visual representation of gathered insights that users must actively integrate rather than passively consuming information without synthesis. These interventions, collectively, represent examples of potential options to prevent information scent traps and overgrazing patterns that characterize unsuccessful forecasting, as well as tools that may help average forecasters and bottom forecasters to development reasoning skills.

Limitations and Future Research

Key limitations

Several limitations constrain our interpretations and generalizability. The most significant is our small sample size of seven analyzed participants, including only two proxy top forecasters. Although we report group means for H1 to H10 to address our

hypotheses, participants within each group have some divergent foraging behaviours. Consequently, a small number of hypothesis verdicts are sensitive to membership. For example, BAF1 (part of expanded bottom-group) devoted unusually high time to Analysis Page (area E), which raised the expanded bottom-group mean (4.78%) to above top forecasters (3.08%). If BAF1 is excluded, H3 would be supported. A similar argument can also be made for BAF1 in the context of H6 (external research), strengthening the support meaningfully. Another example would be BF2's extreme 93% fixation on Activity might represent a statistical outlier or bias due to missing data, rather than common failure modes. As such, individual variations may masquerade as systematic patterns. Moreover, while our top forecasters have provided an overview of desirable traits, with our limited participants, we cannot definitively determine whether TF1 and TF2 represent distinct successful strategies, or individual variations within a broader top forecaster profile. Although our findings show initial support for certain adaptive on-platform behaviours, a more extensive study is needed to estimate their direct impact on accuracy, given evidence that such behavioural observations, while useful, explain only some variance (Atanasov & Himmelstein, 2023).

Measurement challenges present additional constraints. Our webcam-based eye tracking required manual coding of screen recordings after automated tracking showed poor calibration across all participants. While we could identify which section was displayed, we could not pinpoint gaze location within pages; this prevented us from determining nuances of participants' attention (e.g., did they read full rationales versus just scanning usernames and percentages, which specifics [parts of] charts in the Analysis section were examined, or what values were focused on in Crowd Forecast). This limitation is particularly problematic for distinguishing information extraction strategies within patches. The qualitative distinction between productive iteration (TF2's 47% in Forecast Page) versus unproductive rumination (BF1's 53%) relied solely on observable typing-deleting-and-retyping patterns. Laboratory-grade eye tracking could reveal whether BF1 repeatedly fixated on single sentences or was scanning their entire draft; thereby, highlighting behavioral markers that can better support our argumentation for synthesis efficiency.

Data collection issues compound these limitations. Missing sessions for TF2 (34 minutes), BF2 (57 minutes), compounded with failed uploads for TF2, BAF1, and BF1, as well as with forgotten eye-tracking setup for AF1 and BF2, bias our proportional time calculations and behavioral generalizations. These missing data points are problematic given our small sample, as each lost session represents substantial information about behavioral patterns.

Despite these limitations, this pilot study adequately serves its intended purpose: establishing feasibility for larger-scale investigations and revealing unexpected behavioral patterns that challenge our initial hypotheses. The heterogeneity in our data demonstrate the value of exploratory research in understudied domains. Our small sample allowed deep qualitative analysis of individual foraging patterns, revealing nuances such as rumination cycles and crowd entrapment that might be obscured in larger quantitative studies. The technical challenges encountered, from webcam tracking accuracy to participant compliance with protocols, provide crucial methodological insights for future research design.

Our findings also suggest that information foraging in forecasting environments operates differently than traditional IFT predictions, with successful forecasters demonstrating efficiency and selectivity of information, rather than comprehensive coverage of all available information patches. These preliminary insights, while requiring validation through larger samples and more controlled experiments, offer a foundation for reconceptualizing how expertise develops in complex information environments, and how forecasting platforms might better support this development.

Future Research Directions

First, our proposed design modifications require empirical validation with a focus on behavioral outcomes. Visual hierarchy changes emphasizing high-profitability areas need testing (i.e., A/B testing) (Kohavi et al., 2020) to determine such changes in information scent (Pirolli, 2007) can improve productivity and/or forecasting accuracy or merely redistributed time. The information extraction methods (i.e., new Cultivate features) proposed earlier should also be evaluated for their impact on both update frequency and adjustment magnitude; for example, can they help average forecasters engage in more

frequent but small incremental adjustments? Exit interventions prompting setting reminders and comprehensive exploration also require assessment of both compliance rates and subsequent forecasting accuracy.

Second, the paradox of some forecasters mimicking top forecaster foraging patterns through extensive external research (e.g., BAF1's 41% matching top forecasters' levels) and moderate investment in crowd rationales, but still achieving poor performance, fundamentally challenges our understanding of expertise. Interventions such as think-aloud protocols (Fox et al., 2011) during active forecasting could reveal whether top forecasters gather confirming or diverse information, how source credibility is evaluated, and where synthesis breaks down. The stark contrast in update productivity (1.5 updates per session for top forecasters vs. 0.7 for bottom forecasters) despite similar time allocation suggests fundamental differences in information processing. Lab-based eye tracking at the sentence/word level could determine whether poor forecasters read different content than top forecasters or simply process identical information ineffectively.

Third, systematically varying information access on the platform could test whether patterns represent optimal adaptations or arbitrary preferences based on availability of choices. Key experiments should examine, for example, how does removing external research ability affect different types of forecasters? How does limiting Activity section impact update productivity? Do over-consumers forced to diversify achieve similar updates per session seen in top forecasters? Can artificial time constraint transform lengthy, low-yield sessions (e.g., 44 minutes for bottom forecasters) into efficient, high-productivity sessions? Adding artificial barriers may help further isolate the utility of each area and the interplay with time allocated to said areas.

Fourth, our intriguing finding that top forecasters make smaller and more frequent adjustments (6% magnitude, 1.5 forecasts per session) versus larger magnitudes and less productive sessions of non-top forecasters suggest a critical skill gap. Future research could explore e.g., what triggers the decision to update versus continue researching? How do top forecasters determine adjustment magnitude? Can training in incremental calibration improve accuracy even without changing foraging patterns? (e.g., Mellers et al., 2014;

Mellers, Stone, Atanasov, et al., 2015). The relationship between session duration, update frequency, and forecast adjustment magnitude may reveal optimal intervention points.

Finally, scaling to larger samples with diverse populations is essential. Our pilot study with graduate students at an elite Swiss university may not generalize to broader populations. Future studies should include participants with varying educational backgrounds, domain expertise, and professional experience to determine whether identified patterns for on- and off-platform represent universal phenomena, or population-specific effects. This is particularly noteworthy given that Tetlock (2005) and Mellers, Stone, Atanasov, et al. (2015) have shown that many seemingly relevant predictors (e.g., education levels, professional experience) do not influence prediction accuracy. However, it may be plausible that the design of the forecasting platform could better enable participants with certain characteristics to perform differently. Additionally, further investigations into validated experts from forecasting tournaments (i.e., superforecasters) (e.g., Mellers et al., 2014) could confirm whether the observed patterns persist.

These research directions would advance both IFT's application in forecasting, as well as practical understanding of how to develop forecasting expertise through platform design and on- or off-platform training interventions. Findings from this pilot, particularly the idea that success emerges from efficient, calibrated updates through targeted information foraging on- and off-platform, as opposed to exhaustive and thorough information consumption, suggest rich opportunities for reconceptualizing how we understand and support expert decision-making in information-abundant environments.

References

Atanasov, P., & Himmelstein, M. (2023). Talent spotting in crowd prediction. In M. Seifert (Ed.), *Judgment in predictive analytics* (pp. 135–184). Springer.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science, 63*(3), 691–706.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19–35.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2025). Crowd prediction systems: Markets, polls, and elite forecasters. *International Journal of Forecasting, 41*(2), 580-595.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis, 11*(2), 133–145.

Bateman, J. E., Thompson, K. A., & Birney, D. P. (2019). Validating the relation-monitoring task as a measure of relational integration and predictor of fluid intelligence. *Memory & Cognition, 47*, 1457–1468.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*(2), 267–280.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting. *Judgment and Decision Making, 11*(5), 509–526.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55–81.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions on the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 490–497). ACM.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554.

Csaszar, F. A., & Laureiro-Martinez, D. (2018). Individual and organizational antecedents of strategic foresight: A representational approach. *Strategy Science, 3*(3), 513–532.

Djamasbi, S. (2014). Eye tracking and web experience. *AIS Transactions on Human-Computer Interaction, 6*(2), 37–54.

Djamasbi, S., Siegel, M., & Tullis, T. (2010). Generation Y, web design, and eye tracking. *International Journal of Human-Computer Studies, 68*(5), 307–323.

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 717–726.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society, 20*(5), 325–344.

Faraday, P. (2000). Visually critiquing web pages. In N. Correia, T. Chambel, & G. Davenport (Eds.), *Multimedia '99* (pp. 155–166). Springer.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316–344.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555–588.

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Hollander, J. B., Sussman, A., Purdy Levering, A., & Foster-Karim, C. (2020). Using eye-tracking to understand human responses to traditional neighborhood designs. *Planning Practice & Research, 35*(5), 485–509.

Huberman, B. A., Pirolli, P. L., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in world wide web surfing. *Science, 280*(5360), 95–97.

iMotions. (2022). *Webcam eye-tracking accuracy whitepaper* (Version 3) [White paper]. https://go.imotions.com/WebET3Whitepaper

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995–1006.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39*(1), 17–31.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kapoor, R., & Wilde, D. (2023). Peering into a crystal ball: Forecasting behavior and industry foresight. *Strategic Management Journal, 44*(3), 704-736.

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2022). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting, 38*(2), 688–704.

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology, 242*(2), 396–402.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*(1), 111–127.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231–1243.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025.

Lyon, A., & Pacuit, E. (2013). The wisdom of crowds: Methods of human judgment aggregation. In *Handbook of human computation* (pp. 599-614). New York, NY: Springer New York.

Mellers, B., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making, 12*(4), 369–381.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of

intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267–281.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., & Swift, S. A. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science, 3*(5), 400–424.

Oberauer, K., Süß, H. M., Wilhelm, O., & Sander, N. (2008). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). Oxford University Press.

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems, 24*(1), 485–500.

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods, 50*, 1645–1656.

Pan, B., Hembrooke, H., Gay, G., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior: An eye-tracking study. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications* (pp. 147–154). ACM.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford University Press.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*(4), 643–675.

Pirolli, P., & Fu, W. T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *User modeling 2003* (pp. 45–54). Springer.

Schuler, B. A., Murmann, J. P., Beisemann, M., & Satopää, V. (2025). Individual foresight: Concept, operationalization, and correlates. *International Journal of Forecasting, 41*(4), 1521-1538.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the user interface: Strategies for effective human-computer interaction* (6th ed.). Pearson.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—but some don't*. Penguin Press.

Skåner, Y., Strender, L. E., & Bring, J. (1998). How do GPs use clinical information in their judgments of heart failure? A clinical judgment analysis study. *Scandinavian Journal of Primary Health Care, 16*(2), 95–100.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishers/Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science, 23*(4), 290-295.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2013). Choice architecture. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 428–439). Princeton University Press.

Tinbergen, N., & Perdeck, A. C. (1951). On the stimulus situation releasing the begging response in the newly hatched herring gull chick (Larus argentatus Pont.). *Behaviour, 3*(1), 1–39.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge University Press.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin, 134*(2), 163–206.

Weinmann, M., Schneider, C., & vom Brocke, J. (2016). Digital nudging. *Business & Information Systems Engineering, 58*(6), 433–436.

Wisiecka, K., Krejtz, K., Krejtz, I., Sromek, D., Cellary, A., Lewandowska, B., & Duchowski, A. (2022). Comparison of webcam and remote eye tracking. In *2022 Symposium on Eye Tracking Research and Applications* (pp. 1–7). ACM.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*, 0058.

Appendix

Appendix A

Consent Form for Eye Tracking

Title of Project
User Experience in Forecasting Platforms

Investigators
Thomas Li; Thomas.li@unisg.ch
J. Peter Murmann; j.peter.murmann@unisg.ch
Institute of Management & Strategy, University of St. Gallen, Dufourstr. 40A

Summary of the Project
The purpose of the project is to understand how users interact with and make forecasts using a popular platform (i.e., Cultivate Labs), with the goal of improving the usability and design of the platform. Your eye movements will be tracked via the webcam of your own personal computer, and inferences of where you are looking on the screen will be made.

Basic demographic information, such as age, gender, and ethnicity, will be recorded. All participant data will be anonymized. No personal identifying information will be shared. Your participation in this study is voluntary, and will have no effect on your course grade. You may withdraw your participation at any time.

Procedure
Whenever you make a choose to browse or to make a forecast on the University of St. Gallen forecasting platform, please use the special link provided to you.

There will be an exercise to calibrate the eye tracker to you and your viewing position both before and after your session on the forecasting platform. The exercise simply involves sitting in position and visually following a dot appearing at different places on the screen for a short period of time (<60 seconds in total). You will be redirected to the forecasting platform after the initial calibration; please return to initial page after exiting the forecasting platform to complete the second calibration. Please refrain from moving your bodily position too much and try to keep your eyes engaged with your computer screen as much as possible.

Confidentiality and Data Security
All information you provide is considered completely confidential. Your name will not appear in any publication resulting from this exercise. Data collected will be retained in a password protected and secure location. All data will not include any identifying demographic information, and will be stored for statistical analysis.

Renumeration
For your participation until Nov 28, 2022 (due date of the last forecasting question), you will receive a one-time renumeration of CHF 400.00 upon the completion of the study. Please provide your IBAN for disbursement. Special requests for cash payment will also be possible.

Consent Statement and Signature
I have read this consent form and understand the information that has been provided above. I have had the opportunity to have my questions answered to my satisfaction and understand the research and my rights as a participant. I voluntarily agree to participate in this study. I understand that a copy of this consent will be provided to me for future

_____
Participant Name

_____
Signature

_____
Date, Place

Appendix B

Forecasting Questions

1. **Question:** What will be the balance of power in the United States Congress after the 2022 election (end-of-day Nov 8, 2022 CET)?
   Possible answers: Republican House, Democratic Senate; Both Republican House & Senate; Both Democratic House & Senate; Democratic House, Republican Senate

2. Question: Will Xiaomi surpass Apple in the percentage (%) of units of mobile phones shipped in Q3 2022 (i.e. by end-of-day Sept 30, 2022)?
   Possible Answers: Yes or No

3. **Question:** Will Finland and/or Sweden become members of NATO before end-of-day Dec 14, 2022 (CET)?
   Possible Answers: Yes, Finland only; Yes, Sweden only; Yes, both Finland and Sweden; No

4. **Question:** Will Pfizer overtake Roche in market cap ($) at end-of-day Nov 23, 2022 (EST)?
   Possible Answers: Yes or No

5. **Question:** What will be the end-of-day TTF Natural Gas Futures price (€/MWh) on Dec 7, 2022 (CET)?
   Possible Answers: Less than or equal to 99.99; Between 100.00 and 149.99, inclusive; Between 150.00 and 199.99, inclusive; Between 200.00 and 249.99, inclusive; Between 250.00 and 299.99, inclusive; Between 300.00 and 349.99, inclusive; Between 350.00 and 399.99, inclusive; More than or equal to 400.00

6. **Question:** Will North Korea detonate a nuclear device on or before end-of-day Dec 7, 2022 (CET)?
   Possible Answers: Yes or No

Appendix C

Screen Recording Coding Procedure (iMotions 10)

1. **Define the "Whole Screen" AOI** (Session coverage)
   Create a manual AOI that covers the screen and apply it only from the moment the participant is on a screen that is relevant to forecasting (i.e., a Cultivate page or a relevant external source) until the last second on a forecasting-related screen before terminating the session[10]. This bounds total session duration and prevents inflation from unrelated activity (e.g., eye tracking calibration in the beginning, or forgetting briefly to turn off screen recording).

2. **Create area labels as custom annotations**
   Define a set of annotations (i.e., areas/actions of interests) with distinct labels and colors (e.g., green for Forecast Page)
   a. Cultivate annotations:
   Landing Page (Area A)
   Forecast Page (Area B)
   Activity Page (Area C)
   Crowd Forecast Page (Area D)
   Analysis Page (Area E)
   Source Links Page (Area F)
   My Forecasts (Area G)
   Review Closed/Pending Questions
   Switch Questions
   Set Reminders

   b. Off-platform annotations:
   External research
   Miscellaneous

3. **Annotate the timeline (primary coding)**
   Play the screen recording and apply the relevant annotations. For example, the moment a participant starts conducting external research, the "External Research" annotation starts and ends when the screen changes. All annotations are made manually. The threshold for annotating is for any segment lasting longer than one second.

4. **Ensure quality control for each session**
   Before the end of coding for each video recorded session, 1) confirm that there are no unintended overlaps across annotations and no gaps except ignored time, and 2) re-check ambiguous intervals (e.g., when a participant is using split screen, the attention must be properly designated). It is acceptable that there are instances where a small duration of the recording is not annotated.

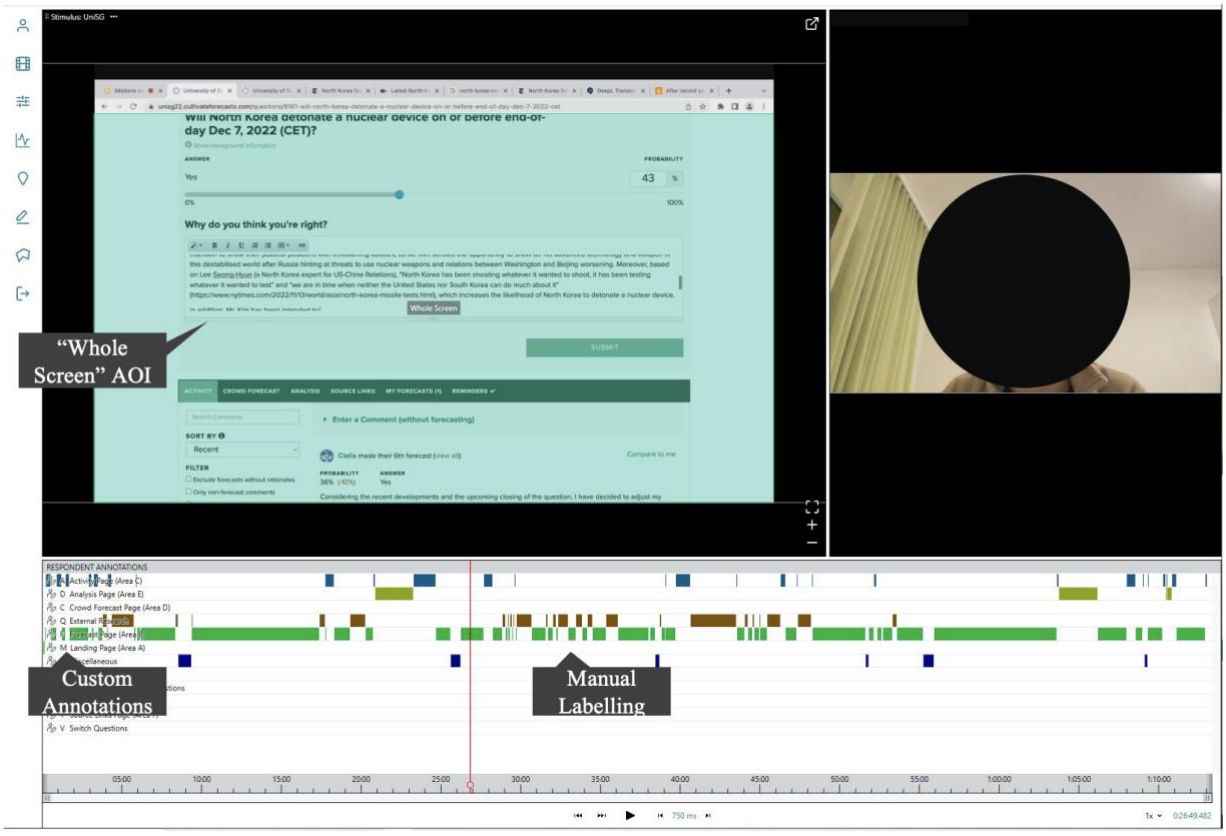
[10] Typically, only a limited 5-25 seconds from the beginning and/or end of each recording is cut.

5. **Double code each session to ensure reliability**
   Each session recording is double-coded manually.

6. **Resulting data is exported as CSV for further processing**
   iMotion exports contain information about "whole screen" AOI duration per session (default output in ms), as well as annotation durations per label (in ms) per session. This data export supports our calculations. See screenshot below for an overview.

Decision rules for edge cases:

- **Split-screen/multiple windows:** code the window with observable interaction (typing, scrolling, clicking etc.)
- **Idle detection:** if the same window remains foreground with no interaction, and the participant is clearly disengaged (from webcam recording), no coding is done.
- **Accidental clicks:** if the user clearly clicked on a wrong page/tab, and returned the previous page (e.g., duration less than one second), these are to be ignored.
- **Question Switching:** The annotations are used as a point event marker to count switches, and not duration.



Appendix D

Overview of Collected Behavioral Data

	Participants						
Overview of Eye Tracking Data	**TF1** (*R**: 1/55)	**TF2** (*R*: 3/55)	**AF1** (*R*: 18/55)	**AF2** (*R*: 23/55)	**BAF1** (*R*: 39/55)	**BF1** (*R*: 48/55)	**BF2** (*R*: 54/55)
Total # of platform visits	10	6	3	11	9	13	8
No access to eye tracking	1	1	1	2	1	1	1
Video recorded visits (w/ eye tracking)	9	3	2	9	7	11	3
Video recorded visits (screensharing forgotten)	0	1	0	0	0	0	2
No recording visits (upload error)	0	1	0	0	1	1	0
No recording visits (entire eye tracking forgotten)	0	0	0	0	0	0	2
Total % of platform visits	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
% No access to eye tracking	10.0%	16.7%	33.3%	18.2%	11.1%	7.7%	12.5%
% Video recorded visits (w/ eye tracking)	90.0%	50.0%	66.7%	81.8%	77.8%	84.6%	37.5%
% Video recorded visits (screensharing forgotten)	0.0%	16.7%	0.0%	0.0%	0.0%	0.0%	25.0%
% No recording visits (upload error)	0.0%	16.7%	0.0%	0.0%	11.1%	7.7%	0.0%
% No recording visits (entire eye tracking forgotten)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%

**R:* x/55 is forecasting accuracy rank out of 55 people