Pre-Analysis Plan

Ryan Kennedy, Philip D. Waggoner, Matthew Ward,
University of Houston[1]

Title: Trust in Automation Experiments

Abstract: The past decade has seen incredible strides made in the ability for computers and algorithms to forecast social events. One of the most difficult questions in the use of these algorithms, however, is the degree to which humans trust them to assist in decision-making. If the algorithm does perform better, we want a human to use the algorithm's forecasts. We do not, however, want humans to be so reliant on the algorithm that they fail to note when the algorithm is leading them astray. These two alternatives have been labeled "algorithms aversion" and "automation bias", respectively. While the literature on these twin dangers is already large in aerospace and is growing in the medicine and computer science, little has been done to evaluate these issues with regards to political decision-making. Moreover, there are clear blind-spots in the literature, such as the difficulty in understanding the heterogeneity of effects based on when information is presented and the area of algorithm appropriateness.

To date, we have conducted a pilot analysis with preliminary inspection of the results across three waves of experiments asking students and respondents on Amazon's Mechanical Turk (MTurk) to evaluate forecasting tasks like those posed by the Good Judgment Project. Respondents are given advice from varying sources and at different times during the decision-making process. These experiments are designed to elicit the degree to which students trust information from algorithms in making their forecasts. The experiments are ongoing and are analyzed cumulatively in a Bayesian hierarchical framework. This allows us to increase the power of the experiments and evaluate heterogeneity across forecasting tasks. In addition to contributing to our understanding of automation on political decision-making, the results will also be utilized to improve forecasting in government and business.

Sample: The three completed waves included in the pilot analysis were comprised of MTurk respondents and undergraduate students at a large public research university. The result is nearly 1,000 Americans in three waves of surveys. The fourth wave, soon to be fielded and subsequent experiments will be conducted similarly.

Respondents in our samples are randomly assigned to one of four groups. The first group is told the average response of a group of individuals provided a given set of probabilities of the event occurring. The second group is told a group of randomly selected individuals

[1] Corresponding author: Ryan Kennedy, rkennedy@uh.edu. Authors listed in alphabetical order.

provided the following probabilities for the event occurring. The third group is told a computer algorithm was used to combine human responses, resulting in the set of probabilities for the event occurring. Finally, the fourth group is told a computer algorithm, developed by experts and scientists, generated the set of probabilities for the event occurring.

Treatments: The treatments offer probabilities from different sources of the given event occurring (stated in the IFP at the outset). Respondents are randomly assigned to one of the four treatment conditions. The text for these conditions are included as an attachment to this pre-analysis plan. Importantly, all probabilities are identical across all treatment conditions. The source of the probabilities is the only varied information in order to get at the underlying concept of interest, which is the level of trust (or not) in automation and computer algorithms to assist in making forecasts.

Hypotheses: The ultimate goal of our research program is to understand the degrees to which humans trust (or do not) algorithms and automation as they make decisions. As such, in pursuit of this ultimate understanding of the interaction between humans and automation, there are numerous hypotheses we will likely test with the current data across three waves, as well as with subsequent data in future waves of our experiments.

Main hypotheses:

- In general, we expect respondents to be most trusting of algorithmic aggregation of human responses, followed by averages of human responses, followed by algorithm-generated advice, and least trusting of the advice of a randomly selected human.

- When an event is less salient or clear, respondents should trust automation to a greater degree than humans.

- As the number of humans cited as giving advice in the treatment increases, respondents should be increasingly likely to trust human advice.

- Algorithms should be more trusted, relative to human advice, in situations where algorithm performance is likely to be higher (e.g. quantitative tasks, that occur relatively frequently, with long and accessible training data).

- When advice is presented prior to being asked to forecast an event, the respondent should be more likely to take advice across all treatment conditions (i.e., all sources).

Additional hypotheses:

- Respondents with higher political knowledge should be less likely to trust the advice given, regardless of source.

- Respondents with detail-oriented, duty-driven personalities (e.g., conscientiousness, emotional stability, etc.) should be more likely to trust automation over human advice.

- Respondents with more open personalities (e.g., extroverted, open to new experience) should be more likely to trust human advice over automation.

- Older respondents should trust automation less than younger respondents.

- Respondents with higher levels of education should be more likely to trust automation over human advice, compared to respondents with less education.