Kiersten Henderson
Essay 2: Experimental Research Proposal

Does Providing People with a Balanced View of the Benefits and Risks of Artificial Intelligence Alter their Opinion of Artificial Intelligence?

Background

The rise of Artificial Intelligence (AI) in increasingly clever software and machines promises unprecedented gains to social and economic welfare. It is hypothesized that AI-based superintelligence will create revolutionary new technologies that could help us eradicate war, disease, and poverty. However, there is the potential for AI systems to cause great harm intentionally or unintentionally[1].

Hollywood frequently portrays AI as "robots taking over the world" and comments from Science and Tech giants Bill Gates, Elon Musk, and Steven Hawking warn that AI could be the biggest existential threat that humanity faces[2]. Although the news media does often portray a balanced view of AI, we wondered how well-exposed the average person is to both the pros and cons of AI-based technologies and to what extent this shapes their attitudes toward AI.

A recent study found that people who watched sci-fi films that either depicted AI in a positive or negative light actually became more extreme in the view they held of AI prior to watching the film[3] (more supportive or more skeptical of AI) . Thus, when consuming information about AI from a film source, people exhibited confirmation bias. We wondered how presenting people with both the pros and cons of AI-based technologies (a balanced view) in a non-entertainment setting would affect their attitudes toward AI.

To address this question, we will ask:

Does Providing People with a Balanced View of the Benefits and Risks of Artificial Intelligence Alter their Opinion of Artificial Intelligence?

Answering this question is important because how people feel about AI-based technologies will likely affect how technologies are implemented. Will there be resistance to the technologies because they are viewed as disruptive to society or too dangerous? Does the average person think that oversight of AI-based technologies should be established? If we found that opinions changed based on learning the pros and cons of AI, we could apply this knowledge to shape the approach we take to educating the public so they can make informed decisions.

Experimental Design

The experimental design is that two surveys with the same survey questions (described below) will be administered to different sets of people. Both surveys will have an

introductory paragraph that defines AI for laymen and that defines the scope of AI that we are addressing when soliciting opinions in the survey. For the scope of the survey, “AI” will refer to: self-driving vehicles, decision-making robotics for both every day and industrial applications, automated language recognition/processing that responds to users or makes recommendations based on input or translates/ transcribes, and algorithms that interpret complex data to make decisions that affect people.

The control survey will only include the introductory paragraph that introduces AI. The treatment survey will be identical to the control survey, but will have a paragraph prior to each survey question that provides a short summary of current expert opinions of potential impacts of AI (both the pros and cons) related to the survey question.

Questions will be posed to evaluate the opinions of respondents to the following topics:

1) How beneficial to society will the widespread introduction of AI technology be?
2) How large is the possibility of AI having unintended consequences that are dangerous or undesirable?
3) How concerned are they about the potential for increased unemployment due to replacement of workers by AI-enabled automation?
4) Should regulatory oversight be established prior to the widespread adoption of particular kinds of AI technologies?

For example, to address question 3, the survey question could be:

On a scale of 1 to 5 (with 5 being most concerned), how concerned are you that widespread adoption of Artificial Intelligence will lead to major job losses across several employment sectors?

I don’t know/No opinion 1 2 3 4 5

In the treatment group survey, the following paragraph will precede the survey question above:

Historically, the relationship between automation and employment has been complicated. Automation has eliminated old jobs, but it has also created new jobs. Studies vary on their estimates of what the risk is that AI will eliminate jobs. One study estimates that 47% of U.S. jobs are at “high risk” of potential automation, while another classifies 9% of U.S. jobs as “high risk”[4]. Jobs categorized as “high risk” include paralegals and fast food cooks, to name a few. However, the demand for care-related professionals who work in fields ranging from personal healthcare to the clergy is likely to increase. Economists point out that in the past, technology has tended to increase rather than reduce total employment, but acknowledge that "we're in uncharted territory" with Artificial Intelligence technologies[5].

The outcome measure we care most about is whether the attitude of people in the treatment and control group differ with respect to particular questions we address. Meaning, does giving people more information about AI change their opinion of AI?

We anticipate that the control group will represent the "largely uninformed" popular opinions towards AI. However, this group could contain respondents with significant amounts of knowledge of AI technologies. If this is the case, we anticipate that these people will be equally represented in our treatment group. We will collect information about how informed the respondents are about Science and Technology (described below in Covariates) in order to evaluate whether well-informed respondents are equally represented in the treatment and control groups.

We will test whether, compared to the control group, the treatment group (given information about AI) becomes *more or less* likely to think that AI will benefit society*, more or less* concerned that AI would have unintended negative consequences, *more or less* concerned that AI could increase unemployment, or *more or less* likely to think that AI technologies should be regulated. Alternatively, the responses of those surveyed in the treatment group might not differ from those in the control group.

To perform these statistical comparisons, we will use randomization inference using the Sharp Null Hypothesis of no effect. This approach assumes that for every unit of observation there is no effect and that the potential outcomes in both treatment and control groups are identical. Once we have this distribution, we will determine the two-tailed p-value of the randomization inference test to determine the probability of achieving the actual average treatment effects we might observe by chance.

<u>Administering the Treatment and Control Surveys</u>
We will administer the surveys using Survey Monkey to Mechanical Turk workers. We will release the two surveys (control and treatment) separately and assume that a random set of people answer one survey and another random set of people answer the other. We have no reason to suspect that the treatment and control groups will be any different from each other, and thus assume this mimics random assignment to a treatment or control group. We will release the control survey first and the treatment survey two weeks later.

<u>Covariates</u>
We expect that if individuals surveyed are already well informed about AI, that the treatment (providing them with more information about the benefits and risks of AI) will not affect their opinion. However, we expect these AI-informed people will be randomly distributed amongst the control and treatment survey takers. Thus, unless these individuals make up a large percent of our total sample, they may not affect our ability to detect any difference between treatment and control if a difference exists. To evaluate whether our assumption that AI-informed respondents would be randomly distributed between treatment control groups, we will determine if survey respondents are already likely to be knowledgeable about AI. We will ask control and treatment survey

respondents: **“How often do you consume Science and Technology news media? (never, once per week, 3 times per week, every day)”**. We will consider that people who consume Science and Technology news media 3 to 7 times a week could be well informed about AI.

It is also possible that the cultural background or personal attributes of respondents could shape their preconceived attitudes to AI in ways we cannot predict a priori. For this reason we will also collect information from respondents regarding several other attributes: Gender, Age Group, Race, Country of Origin, Position on Political Spectrum, and Education level.

Potential Pitfalls

One potential concern for this study is that providing people with a “balanced view” of a subject is, ofcourse, a matter of subjective opinion. When we design our study will will be mindful of designing treatment paragraphs that containing balanced mainstream competing views - both the pros and cons of AI.

Another potential pitfall we will attempt to avoid with our study design is that of interference. For example, an efficient way to conduct this survey would be to post it on Facebook pages. However, it would be very difficult to manage treatment and control groups because individuals might see both versions of the survey and make comments that interfere with the responses of future survey takers (interference).

We can assume non-interference if we perform the survey on Mechanical Turk because there is no way for workers to make comments seen by other workers. However, when using Mechanical Turk, it is possible that one of the control respondents would also fill in the treatment survey (one individual is present in both the treatment and control). It may be possible in Mechanical Turk to block people who already responded to the control version of the survey from responding to the treatment version of the survey. If not, it is probably still wise to use Mechanical Turk workers as the survey subjects rather than Facebook friends because it is less likely that the same people will see both versions of the survey (Mechanical Turk workers are a larger group of people than a group of Facebook friends).

Another potential pitfall in our study design is that there could be selection bias in terms of the kind of people who are Mechanical Turk workers. If Mechanical Turk workers were already very well-informed about AI, then giving them slightly more information will not likely influence their opinions about AI. However, as mentioned we will evaluate selection bias by examining covariates. It is also possible that being a Mechanical Turk worker predisposes survey respondents to have strong preconceived opinions about AI in ways that we cannot anticipate.

Another potential pitfall of our study design could be that respondents either do not understand or have a strong reaction to the way we phrase survey questions or

treatment paragraphs. If this were the case, there might be a large number of I don't know/No opinion responses. We will run a pilot study involving 25 control and 25 treatment individuals to evaluate the efficacy of our survey prior to conducting a full survey with the goal of having 100 treatment and 100 control responses. If we find that we get a large percent of I don't know/No opinion responses, we will consider redesign our survey before administering it to the full treatment and control groups.

Sources

1) https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/
2) *Clever Computers: The Dawn of Artificial Intelligence.* May 2015. The Economist.
3) Vyacheslav Polonski. *People Don't Trust AI--Here's How We Can Change That.* January 2018. Scientific American.
4) Steve Lohr. *Robots will Take Jobs, but Not as Fast as Some Fear, New Report Says.* Jan 2017. The New York Times.
5) Martin Ford and Geoff Colvin. *Will Robots Create More Jobs than they Destroy?* Sept 2015. The Guardian.