Preregistration Report (10/10/2017)
Title: The A/B Illusion: Full Experiment on Autonomous Vehicles and Genetic Testing
Authors: Michelle N. Meyer, Patrick R. Heck, Geoffrey S. Holtzman, Christopher F. Chabris, Duncan J. Watts, William Cai, and Stephen Anderson.

Research Questions:

The A/B illusion is a hypothesized phenomenon in which individuals perceive the decision to run a randomized, controlled experiment (e.g., comparing two interventions, policies, or practices) on human subjects as less appropriate than simply implementing one of those alternatives without testing its effects. The A/B Illusion was previously anecdotally observed and described (Meyer, 2015), but it has never been experimentally investigated. The research questions we are asking include:

1. Can we demonstrate the A/B illusion in naive research participants?
2. Assuming we are able to detect an effect, do any demographic variables or other individual differences either amplify or attenuate the A/B illusion?
3. What kinds of reasons do participants give for endorsing the A/B illusion, and what kinds of reasons do participants give for approving of unilateral implementation of untested policies?

The A/B illusion hypothesis was supported in a previous experiment preregistered and run in our lab on catheterization practices in a medical/hospital setting. The goal of the next experiment is to replicate this effect in two additional domains.

Pilot Research:

We have run a series of preregistered, small sample (N = 20-30) pilot surveys to gauge perceptions of appropriateness in several popular domains. We are interested in running a full-sample experiment (N = 100 per condition) on policy questions in two specific domains: programming autonomous vehicles and returning results of genetic testing.

Hypotheses:

1.) We predict that participants who read vignettes describing an unequivocal policy change decision (to either 'Policy A' or 'Policy B') will rate this decision as more appropriate than those participants who read a similarly unequivocal decision to run a randomized, controlled experiment designed to test the comparative effectiveness of 'Policy A' and 'Policy B.' We predict that this effect will emerge in two distinct domains: autonomous vehicles and genetic testing.

Data Collection Procedures:

Sample & Sample Rationale.

Participants (≥ 18 years; restricted to the United States) will be recruited via Amazon Mechanical Turk. Unless otherwise specified, we will recruit approximately 100 participants per vignette condition. The effect sizes observed in the pilot study suggest that samples of this size will adequately power our experiment (> 95% power).

Stimuli (Vignettes) and Additional Questions.

Genetic testing - Prevention

A: Some genetic mutations lead to health conditions that can make a person sick, or even cause them to die. Many of these health conditions can be prevented or slowed by taking certain steps once a person knows they have a genetic condition, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origins, but the CEO wants to help as many people as he can. So, he decides that he will offer all of his clients the option to see if they have any genetic risks for health conditions that can be prevented or reduced. Customers will have the option of viewing these results or not.

B: Some genetic mutations lead to health conditions that can make a person sick, or even cause them to die. Many of these health conditions can be prevented or slowed by taking certain steps once a person knows they have a genetic condition, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origins, but the CEO wants to help as many people as he can. So, he decides that he will offer all of his clients the option to see if they have any genetic risks for health conditions, whether or not these health conditions can be prevented or reduced. Customers will have the option of viewing these results or not.

A/B SHORT: Some genetic mutations lead to health conditions that can make a person sick, or even cause them to die. Many of these health conditions can be prevented or slowed by taking certain steps once a person knows they have a genetic condition, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origins, but the CEO wants to help as many people as he can. He thinks of two different ways to help customers, so he decides to run an experiment by randomly assigning them to one of two test conditions. He will offer half of his clients the option to see if they have any genetic risks for health conditions that can be prevented or reduced. He will offer the other half the option to see if they have any genetic risks for health conditions, whether or not these health conditions can be prevented or reduced. Customers in both test conditions will have the option of viewing the additional results offered to them.

A/B LONG: Some genetic mutations lead to health conditions that can make a person sick, or even cause them to die. Many of these health conditions can be prevented or slowed by taking certain steps once a person knows they have a genetic condition, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origins, but the CEO wants to help as many people as he can. He thinks of two different ways to help customers, so he decides to run an experiment by randomly assigning them to one of two test conditions. He will offer half of his clients the option to see if they have any genetic risks for health conditions that can be prevented or reduced. He will offer the other half the option to see if they have any genetic risks for health conditions, whether or not these health conditions can be prevented or reduced. Customers in both test conditions will have the option of viewing the additional results offered to them. After a year, the CEO will provide all new customers with whichever option turns out to lead to the highest customer satisfaction.

Autonomous vehicles - Control

A: Many people like the idea of completely self-driving cars, which are capable of navigating the road without any input from a human driver. These kinds of cars can make people's lives easier and reduce accidents, but some of them prevent people from taking control of their car in the case of emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible while on the road, while also remaining safe. So, he decides that all of the company's cars will have a lever that allows drivers to switch between self-driving and human-driving modes.

B: Many people like the idea of completely self-driving cars, which are capable of navigating the road without any input from a human driver. These kinds of cars can make people's lives easier and reduce accidents, but some of them prevent people from taking control of their car in the case of emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible while on the road, while also remaining safe. So, he decides that any use of the brakes, gas pedal, or steering wheel by a human driver will automatically override self-driving mode on the company's cars.

A/B SHORT: Many people like the idea of completely self-driving cars, which are capable of navigating the road without any input from a human driver. These kinds of cars can make people's lives easier and reduce accidents, but some of them prevent people from taking control of their car in the case of emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible while on the road, while also remaining safe. The CEO thinks of two different ways to balance freedom and safety, so he decides to run an experiment by randomly assigning cars to one of two test conditions. Half of cars the company sells will have a lever that allows drivers to switch between self-driving and human-driving modes. The other half will be programmed so that any use of the brakes, gas pedal, or steering wheel by a human driver will automatically override self-driving mode on the company's cars.

A/B LONG: Many people like the idea of completely self-driving cars, which are capable of navigating the road without any input from a human driver. These kinds of cars can make people's lives easier and reduce accidents, but some of them prevent people from taking control of their car in the case of emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible while on the road, while also remaining safe. The CEO thinks of two different ways to balance freedom and safety, so he decides to run an experiment by randomly assigning cars to one of two test conditions. Half of cars the company sells will have a lever that allows drivers to switch between self-driving and human-driving modes. The other half will be programmed so that any use of the brakes, gas pedal, or steering wheel by a human driver will automatically override self-driving mode on the company's cars. After a year, the CEO will have all of the company's cars built using whichever design turns out to lead to the fewest accidents.

Primary Dependent Measure: How appropriate is the CEO's decision?

Participants will be asked to rate the decision on a 1-5 scale measuring appropriateness. Participants will also be asked to complete a free-response item asking them why they chose the rating they gave.

Additional items: We will ask participants to answer three questions about their beliefs in the relative importance of God, faith, feelings, and science:

1.) God or some type of nonhuman entity is in control of the events in the universe.
2.) We believe too often in science, and not enough in feelings and faith.
3.) Knowledge can best be obtained through scientific research.

Finally, we will ask several questions designed to assess science literacy and education. “Experimental Understanding” is adapted from the 2014 NSF Science and Engineering Indicators survey (available from https://www.nsf.gov/statistics/seind14/). “Genetic Understanding” and “Scientific Knowledge” are taken directly from this survey.

Experimental Understanding: Two scientists want to know if a certain new plant food will increase plant growth. The first scientist wants to give the new food to 1,000 plants and see how many of them grow larger than they were before they received the new food. The second scientist wants to give the new food to 500 plants and give the normal, standard food to another 500 plants, and see whether the plants in the first group grow more than the plants in the second group. Which is the better way to test this plant food?

(a) Give the new food to all 1,000 plants.

(b) Give the new food to 500 plants and give the normal, standard food to another 500 plants.

Genetic Understanding: A doctor tells a couple that their genetic makeup means that they have a one in four chance of having a child with an inherited illness.

(1) Does this mean that if their first child has the illness, the next three will not?

- No, the next three still might inherit the illness.
- Yes, none of the next three children will have the illness.

Scientific Knowledge True/False Questions:

Please indicate whether the following statements are true or false:

1. The center of the Earth is very hot.

2. All radioactivity is man-made.

3. Lasers work by focusing sound waves.

4. Electrons are smaller than atoms.

5. The continents have been moving their location for millions of years and will continue to move.

6. It is the father’s gene that decides whether the baby is a boy or a girl.

7. Antibiotics kill viruses as well as bacteria.

8. Do you have a college or graduate school degree in any science or engineering field? *Participants who answer yes to this question will be able to enter the specific science or engineering field in a free-response text box.*

Demographics:

Finally, we will ask participants to provide their sex, age, income, race/ethnicity, and educational attainment.

Data Exclusions.

We intend to exclude the following participants:

1. Those who demonstrate a clear misunderstanding of the vignette, as determined by their free-response explanation. An example of a response that excluded a participant under this criterion in the pilot study is a participant in an AB condition who stated "I believe it would give a detailed POV on how some patients would react to the readily available info." (In the initial design, the medical interventions were designed to remind doctors, not inform patients.)
2. Those whose free-response explanation makes clear that they did not read the task or take it seriously. Examples of responses that excluded participants under this criterion in the pilot study include: "meow meow meow," "blah blah blah," and "fdff." Participants who simply write "I don't know," or "because that's how I feel" will be excluded because these comments do not clearly demonstrate a lack of understanding.
3. Those whose appropriateness ratings are at odds with their free-response explanation. Examples of responses that excluded participants under this criterion in the initial pilot study include a participant assigned to the Policy A condition who stated "I think putting extra emphasis on prevention is a good thing, it would give doctors something to reference so that steps don't get skipped when they are busy," but who rated the director's decision as "very inappropriate"; another participant assigned to the condition who stated "Because somethign [sic] slightly tedious will save lives, it is worth it.," but rated the director's decision "somewhat inappropriate"; and a participant who stated "It could be a matter of life and death, it should be taken seriously. The poster with [sic] help medical workers be reminded of that.," but who rated the director's decision "very inappropriate". We suspect that these participants confused the poles of the scale.

No other exclusions are planned.

***Planned Analyses**:*

Critical (Step 1) analyses:

Independent groups *t*-test for mean-level differences in appropriateness between A and B. This is just for descriptive purposes.

Independent groups *t*-test for differences in appropriateness between the two AB conditions [AB_long] and [AB_short]. This is just for descriptive purposes.

Compare A appropriateness with the two AB conditions [AB_long] and [AB_short] using two separate independent groups *t*-tests. The AB Illusion hypothesis predicts that A will be rated more appropriate than either AB case.

Compare B appropriateness with the two AB conditions [AB_long] and [AB_short] using two separate independent groups *t*-tests. The AB Illusion hypothesis predicts that B will be rated more appropriate than either AB case.

To obtain an estimate of overall effect size *d*, we will compare the magnitude of the difference in appropriateness between [A] and [B] combined and [AB_long], and the difference between [A] and [B] combined and [AB_short].

Scientific Education/Knowledge (Step 2) analyses:

In each of the four conditions, regress appropriateness on the scientific knowledge/education variable. This is an exploratory question of whether scientific literacy or education can predict perceived inappropriateness of randomized experiments.

Exploratory Demographic (Step 3) analyses:

We will conduct exploratory demographic analyses on: sex, race/ethnicity, age, education, and income, to determine whether any subgroup of participants perceives the experimental (AB) conditions as more or less appropriate.

Free Response Question (Step 4) analysis:

Two trained coders will employ a coding scheme to mark the free-response descriptions as indicated in the codebook below. Individual responses will be coded with multiple labels if they fit the criteria for more than one category. We will measure inter-rater reliability. Disagreements will be resolved by discussion.

We hypothesize that, among respondents in all conditions who rate the decision as somewhat or very appropriate, the most common coding of their free responses will be *Benefit*.

We hypothesize that, among respondents assigned to the A and B conditions who rate the decision as somewhat or very inappropriate, the most common coding of their responses will be *Ineffective*.

We hypothesize that, among respondents in the AB conditions who rate the decision as somewhat or very inappropriate, the most common coding of their free responses will be *Negative Research*.

We hypothesize that more respondents in the AB conditions will object to the apparent lack of patient consent (meeting the conditions for coding their free responses as *Consent*) than will respondents in the A and B conditions.

Reasons Given for Appropriateness Ratings in all Conditions: Codebook

Note that the codebook references the catheterization scenario we ran in our first experiment. These references will be tailored where appropriate to fit the domain being studied (here, autonomous vehicles or genetic testing).

1. Benefit

1.1 Benefit Indication that director's intervention (badges, posters, or experiment comparing badges and posters) will or might be effective in reducing infections or helping patients.

1.2 Learning Specific mention that the intervention will or might help the director learn what will work or what will work best, that it will produce needed evidence, etc. (may or may not include specific positive mention of randomization or experimentation).

2. No harm

2.1 Absence of harm Comment that the respondent believes the intervention won't or is unlikely to do any harm.

2.2 Positive equality In the A or B conditions, a comment that all patients are being treated the same way. In an AB condition, a comment that the two different groups are actually more or less the same (e.g., patients receive the same treatment, which is what counts; or all doctors receive the same informational reminders, just displayed differently).

3. Harm

3.1 Medical risk/harm Comment that the intervention will or may place some or all patients at medical risk or will or may medically harm some or all of them (e.g., because half of patients may or will receive an inferior intervention leading to greater infection rates or because doctors handling badges to review safety procedures may or will compromise a sterile environment).

3.2 Other risk/harm Comment that the intervention will or may place some or all patients at non-medical risk or will or may harm some or all patients in some non-medical way (e.g., by causing patients anxiety about whether their doctor is competent).

4. No benefit

4.1 Ineffective Comments that the intervention won’t be, or is unlikely to be, effective in achieving the goal of reducing infections or helping patients (e.g., “won’t work,” “no point,” “redundant” due to prior training).

5. Negative research

5.1 Randomization Negative comments about randomization as a methodological approach. May include sound or unsound research method, sample size, reduced bias, dangers of randomization, concerns about study design (must clearly address randomization, either by name or proxy [i.e., “gold standard”]).

5.2 Experimentation Negative comments about experimentation, testing, research, or studies, including (for negative comments) “guinea pigs,” “lab rats,” “playing with lives,” “gambling with lives,” “playing God,” or wanting control over health care or medications.

5.3 Negative inequality Negative comment that patients will be treated differently or unequally.

6. Consent

6.1 Notice Comment on the importance of telling patients about the intervention, criticism of the apparent failure to disclose the intervention to patients, etc.

6.2 Consent Comment on the importance of patient choice to participate or not in the intervention, criticism of the apparent failure to obtain patient consent to the intervention, etc.

7. Action

7.1 Act now Comment that the director should act immediately rather than conducting an experiment, or that conducting the experiment for one year before making a decision is too long.

7.2 Best judgment

Comment that the director should “just use his best judgment” or that he should “just do what works best” instead of running an experiment.

8. Intent

8.1 Good intentions

Comment that the director’s intentions are good.

8.2 Bad intentions

Comment that the director’s intentions are bad.

9. Status quo

9.1 Status quo

Comment that the appropriateness of the director’s decision depends on how things (e.g., safety reminders to doctors) are currently done, or that all patients should receive “standard of care,” or words to that effect.

10. Other

10.1 Misunderstandings

Comments that reveal misunderstandings of the vignette (e.g., checklists designed to inform patients rather than remind doctors).

10.2 Irrelevant, unclear, other

Comments that are irrelevant (e.g., “meow meow mewo”) or insufficiently clear to interpret, or that make substantive points but do not fit any of the above categories.

Reference

Meyer, M. N. (2015). Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. *Colorado Technology Law Journal,13*(2), 273-331.