Preregistration Report (10/27/2017)
Title: The A/B Illusion: Full Experiment on Piloted Domains (Drug Prescriptions, Retirement Savings, Charity, Education, Community Healthcare, and Welfare Services).
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 has been supported across three domains in two previous experiments preregistered and run in our lab. The goal of the present experiment is to replicate this effect across a broad range of additional domains. These domains include prescription drugs, retirement savings, charity, education, community healthcare, and welfare.

Pilot Research:

We ran two preregistered, small sample ($N = 30$) pilot surveys to gauge perceptions of (1) appropriateness of unilateral policy implementation and (2) A/B policy testing in a variety of domains. Based on the results of these pilot surveys, we will proceed to run a full-sample experiment ($N = 300$ per domain; 100 per condition) to determine if the A/B illusion arises in the six independent domains listed above, across seven sets of conditions.

Hypotheses:

We predict that participants who read vignettes describing either of two unequivocal policy change decisions will rate these decisions as more appropriate than participants who read a similarly unequivocal decision to run a randomized, controlled experiment designed to test the comparative effectiveness of these same two policies.

Data Collection Procedures*:*

Sample & Sample Rationale.

Participants (≥ 18 years; restricted to the United States) will be recruited via Amazon Mechanical Turk. For this experiment, we will recruit approximately 100 participants per vignette condition (i.e., we will set Mechanical Turk to recruit 2100 participants to our survey, who will be randomized across the 21 conditions). The effect sizes observed in our previous studies suggest that samples of this size will adequately power our experiment (> 90% power).

Notable Changes from Previous Experiments

Experiment structure: The structure of this experiment slightly deviates from the two experiments we have already run. In the present experiment, we reduced the number of experimental (A/B) conditions per domain from two to one. Previously, we included a short and long version of the experimental condition, where the long version included an additional sentence clarifying what the policymaker (doctor, CEO, etc.) in each condition would do after learning the results of the experiment. In our previous experiments, we have never observed a statistically significant difference in rated appropriateness between these short and long conditions. Because these two versions of the experimental condition do not appear to differ in the appropriateness ratings they elicit, we decided to proceed with the complete description of the experiment in all cases (previously referred to as the 'AB-Long' version).

Drug prescription vignettes: We decided to include two versions of a similar vignette within the domain of drug prescriptions ("Best Drug: Teaching" and "Best Drug: No Teaching"). One of these versions is information-sparse: the A and B policy conditions only describe the option that the policymaker (in these cases, "Doctor Jones") chooses to implement (for example, mentioning only Drug A as a treatment before choosing Drug A). In the second version, the A and B policy conditions mention the existence of a possible alternative (for example, mentioning both Drug A and Drug B as a treatment before choosing Drug A). We decided to include both versions because in later stages of this project, we are interested in learning how varying vignette structure may reduce or intensify the A/B illusion. A working, though exploratory, hypothesis at this time is that introducing possible alternative policies into the A and B policy conditions may reduce the A/B illusion effect by increasing disapproval or unilateral policy implementation. This is the first time we will have run two similar vignettes in the same domain in an experiment.

Stimuli (Vignettes) and Additional Questions.

Drug – No Teaching (Prescriptions 1)

A: Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug A. It is affordable and patients can tolerate its side effects.

B: Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug B. It is affordable and patients can tolerate its side effects.

A/B: Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones thinks of two different ways to provide good treatment to his patients, so he decides to run an experiment by randomly assigning his patients who need high blood pressure medication to one of two test conditions. Half of patients will be prescribed drug A, and the other half will be prescribed drug B. Both drugs are affordable and patients can tolerate their side effects. After a year, he will only prescribe to new patients whichever drug has had the best outcomes for his patients.

Drug – Teaching (Prescriptions 2)

A: Several drugs have been approved by the US. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones works in a multi-doctor walk-in clinic where patients see whichever doctor is available. Some doctors in the clinic prescribe drug A for high blood pressure, while others prescribe drug B. Both drugs are affordable and patients can tolerate their side effects. Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug A.

B: Several drugs have been approved by the US. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones works in a multi-doctor walk-in clinic where patients see whichever doctor is available. Some doctors in the clinic prescribe drug A for high blood pressure, while others prescribe drug B. Both drugs are affordable and patients can tolerate their side effects. Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug B.

A/B: Several drugs have been approved by the U.S. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones works in a multi-doctor walk-in clinic where patients see whichever doctor is available. Some doctors in the clinic prescribe drug A for high blood pressure, while others prescribe drug B. Both drugs are affordable and patients can tolerate their side effects. Doctor Jones thinks of two different ways to provide good treatment to his patients, so he decides to run an experiment by randomly assigning his patients who need high blood pressure medication to one of two test conditions. Half of patients will be prescribed drug A, and the other half will be prescribed drug B. After a year, he will only prescribe to new patients whichever drug has had the best outcomes for his patients.

Funds (Retirement Savings)

A: Most people in the U.S. save too little for retirement, causing them financial problems later in life. Company retirement plans are a good way to save, but too few employees choose to enroll in them. The CEO of a company wants to encourage newly hired employees to enroll in the company retirement savings plan, so he decides that he will increase the number of available investment funds from 10 to 15.

B: Most people in the U.S. save too little for retirement, causing them financial problems later in life. Company retirement plans are a good way to save, but too few employees choose to enroll in them. The CEO of a company wants to encourage newly hired employees to enroll in the company retirement savings plan, so he decides that he will provide enrollment paperwork that highlights the most popular of the 10 investment funds the company offers.

A/B: Most people in the U.S. save too little for retirement, causing them financial problems later in life. Company retirement plans are a good way to save, but too few employees choose to enroll in them. The CEO of a company thinks of two different ways to encourage newly hired employees to enroll in the company retirement savings plan, so he decides to run an experiment by randomly assigning new hires to one of two test conditions. For half of new hires, he will provide enrollment paperwork that highlights the most popular of the 10 investment funds the company offers. For the other half of new hires, he will

increase the number of available investment funds from 10 to 15. After a year, the CEO will adopt whichever practice turns out to lead the most employees to enroll in the company's retirement program.

Teacher Motivation (Education)

A: Research has shown that increasing teacher happiness and well-being can result in better student outcomes. The superintendent of a school district wants to improve how well his elementary school students do, so he decides that the school's teachers will receive a yearly bonus.

B: Research has shown that increasing teacher happiness and well-being can result in better student outcomes. The superintendent of a school district wants to improve how well his elementary school students do, so he decides that the school's teachers will receive additional vacation days during summer and winter breaks.

A/B: Research has shown that increasing teacher happiness and well-being can result in better student outcomes. The superintendent of a school district thinks of two different ways to improve how well his elementary school students do, so he decides to run an experiment by randomly assigning the school's teachers to one of two test conditions. Half of the school's teachers will receive a yearly bonus. The other half will receive additional vacation days during summer and winter breaks. After a year, the superintendent will give all teachers whichever benefit turns out to result in better student outcomes.

Escaping poverty (Charity)

A: Last year, a charity received a large number of donations. The director of this charity wants to help people in a low-income country escape extreme poverty, so he decides that all adults below a certain income level will receive a sturdy roof for their home.

B: Last year, a charity received a large number of donations. The director of this charity wants to help people in a low-income country escape extreme poverty, so he decides that all adults below a certain income level will receive one month of training in a trade of their choice.

A/B: Last year, a charity received a large number of donations. The director of this charity thinks of two different ways to help people in a low-income country escape extreme poverty, so he decides to run an experiment by randomly assigning people to one of two test conditions. Half of all adults below a certain income level will receive a sturdy roof for their home. The other half will receive one month of training in a trade of their choice. After a year, the director will begin providing everyone in the country whichever resource (roof or training) turns out to help more people escape extreme poverty.

Healthcare Recruitment (Community Healthcare)

A: A poor nation with a severe shortage of nurses and doctors is creating a new civil service position: Health Assistant. These Health Assistants will undergo one year of training, then become the first line of healthcare for people living in the most remote regions of the country. The nation's congress wants to recruit the best people it can to become Health Assistants, so it decides to have recruitment posters made that emphasize the social benefits of becoming a Health Assistant, such as serving and being a leader in one's community.

B: A poor nation with a severe shortage of nurses and doctors is creating a new civil service position: Health Assistant. These Health Assistants will undergo one year of training, then become the first line of healthcare for people living in the most remote regions of the country. The nation's congress wants to recruit the best people it can to become Health Assistants, so it decides to have recruitment posters made that emphasize the career benefits of becoming a Health Assistant, such as opportunities for promotion and professional development.

A/B: A poor nation with a severe shortage of nurses and doctors is creating a new civil service position: Health Assistant. These Health Assistants will undergo one year of training, then become the first line of

healthcare for people living in the most remote regions of the country. The nation's congress thinks of two different ways to recruit the best people it can to become Health Assistants, so it decides to run an experiment by randomly assigning the nation's districts to one of two test conditions. For half of the districts, the congress will have recruitment posters made that emphasize the social benefits of becoming a Health Assistant, such as serving and being a leader in one's community. For the other half, it will have recruitment posters made that emphasize the career benefits of becoming a Health Assistant, such as opportunities for promotion and professional development. After a year, the congress will have all districts in the nation use whichever kind of poster drew the highest-quality job applicants.

Unemployment Benefits (Welfare)

A: The congress of a small country wants to provide support for citizens who are unable to find work, while also encouraging those citizens to find and take on jobs. To do this, the congress decides on a plan. All citizens who have been out of work for at least 12 months will receive the equivalent of $1000 per month for 6 months, or until they find a job, whichever comes first. To be eligible for these payments, unemployed citizens must attend monthly job fairs run by the government.

B: The congress of a small country wants to provide support for citizens who are unable to find work, while also encouraging those citizens to find and take on jobs. To do this, the congress decides on a plan. All citizens who have been out of work for at least 12 months will receive the equivalent of $500 per month for 6 months, even if they find a job during that time. To be eligible for these payments, unemployed citizens must attend monthly job fairs run by the government.

A/B: The congress of a small country wants to provide support for citizens who are unable to find work, while also encouraging those citizens to find and take on jobs. The congress thinks of two different ways to do this, so it decides to run an experiment by randomly assigning citizens to one of two test conditions. Half of citizens who have been out of work for at least 12 months will receive the equivalent of $1000 per month for 6 months, or until they find a job, whichever comes first. The other half will receive the equivalent of $500 per month for 6 months, even if they find a job during that time. To be eligible for either of these payments, unemployed citizens must attend monthly job fairs run by the government. After a year, the congress will provide to all citizens who have been unemployed for at least 12 months whichever benefit system turns out to lead to lower unemployment among those who receive it.

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

Participants will be asked to rate the decision on a 1-5 Likert scale measuring ranging from very inappropriate to very appropriate. 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. These items are listed below; "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 that same 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 our first 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 not 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 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, even though they were clearly labeled.

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.

Compare A appropriateness with the AB condition using an independent groups *t*-test. The AB Illusion hypothesis predicts that A will be rated more appropriate than AB.

Compare B appropriateness with the AB condition using an independent groups *t*-test. The AB Illusion hypothesis predicts that B will be rated more appropriate than AB.

To obtain an estimate of overall effect size *d*, we will compare the magnitude of the difference in appropriateness between A and B combined with the AB condition. Effect sizes can also be computed by comparing A with AB separately, and again by comparing B with AB. This approach is useful when the A and B conditions differ from each other in rated appropriateness.

As a secondary and simpler way of presenting the results in common language, we will also compare the proportion of participants in each condition within a given domain who judged the decision in the vignette inappropriate. We will do this by coding all "very inappropriate" and "somewhat inappropriate" responses as objections to the decision, and three other responses as non-objection.

Scientific Education/Knowledge (Step 2) analyses:

In each of the four conditions, we will regress rated 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 than the other groups. We will also test whether difference scores between the combined non-experimental (A and B) conditions and the AB conditions can be predicted by any of these demographic measures.

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 interrater 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 condition 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 condition 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.

Codebook: Reasons Given for Appropriateness Ratings in all Conditions

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.