PNAS
www.pnas.org

Supplementary Information for

Objecting to experiments that compare two unobjectionable policies or treatments

Authors: Michelle N. Meyer,[1] Patrick R. Heck,[1,2†] Geoffrey S. Holtzman,[1,2,3†] Stephen M. Anderson,[1,2,4] William Cai,[4,5] Duncan J. Watts,[6] Christopher F. Chabris[2,7]

Corresponding author: Michelle N. Meyer
Email: michellenmeyer@gmail.com

Author Affiliations:
[1] Center for Translational Bioethics and Health Care Policy, Geisinger Health System, Danville, PA 17821.
[2] Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA 17837.
[3] Present address: Department of Psychology, Franklin & Marshall College, Lancaster, PA 17603.
[4] Present address: Department of Psychology, The Pennsylvania State University, University Park, PA, 16802.
[5] Present address: Department of Management Science and Engineering, Stanford University, Stanford, CA, 94305.
[6] New York City Lab, Microsoft Research, New York, NY 10011.
[7] Institute for Advanced Study in Toulouse, 31015 Toulouse, France.
[†] These authors contributed equally.

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1820701116

Table of Contents

References cited in this document are listed on page 110.

Methods

For all research conducted through MTurk, all participants accessed our study from American IP addresses. Complete demographic information for each study is available in the results section of this supplement.

Except for Study 6, participants who completed our questionnaires were compensated for their time, regardless of the quality of the data they provided. All analyses in the manuscript were conducted with the complete set of participants who completed each study and had not participated in another study or pilot described in this document. For our MTurk experiments, individuals who had taken one of our prior experiments or pilots were excluded from participating in future experiments and pilots on those platforms. Exclusions for repeat participants recruited using Pollfish were made manually. Participants in Study One, for which data from MTurk participants were collected via the data collection platform SurveyMonkey, were excluded manually from all subsequent MTurk studies on the basis of their MTurk IDs. Participants in subsequent MTurk studies, all of which were run on the platform Qualtrics, were automatically excluded via the intermediary platform TurkGate.

For all pilots, experiments, and replications consisting of a single questionnaire, MTurk and Geisinger participants were randomly assigned to experimental conditions by algorithms deployed by those questionnaires' host websites. For MTurk pilots and studies consisting of multiple questionnaires, assignment to a given condition was automatically randomized within questionnaires, though the availability of those questionnaires varied over time. For Pollfish replications, random assignment to conditions within a questionnaire was not possible. Instead, separate surveys were launched for each condition.

All analyses reported in the main text were preregistered with the OSF prior to collection of data. Each study (except for Study 1) was fully preregistered before being run. For analysis purposes, every domain within a study—rather than the study as a whole—was the subject of an analysis. The only exception to these two general rules was our hierarchical linear model (HLM) analysis, described in greater detail later in this supplement**.**

Experiments and Replications

Study 1 was conducted with 413 participants, corresponding to roughly 100 participants per condition. Using G*Power *(1)*, we determined that *t*-tests for $\alpha = .05$ and $\beta = .20$ would be able to detect effect sizes of $d \geq 0.40$ for differences between the two policy conditions ($n \sim 200$), or between the two A/B conditions ($n \sim 200$). For the overall A/B effect—that is, the difference between the pooled policy condition means and the pooled A/B condition means ($n \sim 400$)—an effect size of $d \geq 0.28$ was detectable with 80% power.

Participants in this and all other studies were recruited via a link on Amazon Mechanical Turk (MTurk) unless otherwise noted. Participants in this study were then redirected to a questionnaire hosted by SurveyMonkey. After seeing an initial screen on SurveyMonkey stating that they would be compensated $0.25 USD each, participants moved to a second screen on which they were presented with of one of four vignettes.

Each vignette on this second screen involved a hospital director who provided ID badges and/or posters listing safety precautions for central venous catheterization. Full text of these and all other vignettes are listed below in the materials section of this document. Below each vignette, participants rated the appropriateness of the director's decision and provided a written explanation of their rating. A third screen collected demographic information on sex, race, age,

education, and income, and a fourth screen provided a confirmation code for payment and included a text box if participants (optionally) wanted to provide further comments.

Study 2 was conducted with 1,628 participants across two research platforms and consisted of three distinct replications of Study 1. On the basis of their MTurk worker IDs, 74 participants across two of these replications were determined to have participated in Study 1, and 146 participants were found to have taken the Pollfish survey more than once. These participants were manually coded for exclusion in our uploaded data files and automatically excluded by our uploaded R scripts, resulting in a final n = 1,408.

For the first replication of Study 2, 402 participants were recruited to complete a questionnaire that was identical to that used in Study 1, except for three exploratory questions—about God, intuition, and science—at the end of the demographics screen. As in Study 2, participants were paid $0.25 each. Sixteen participants in this replication were determined to have participated in Study 1 and were coded for exclusion, leading to a final n = 386 for this replication.

The second replication, for which 401 MTurk participants were paid $0.35 each, presented participants with one of four vignettes describing roughly the same situations as those described in Study 1. However, the wording of these vignettes differed from the wording in Study 1. Survey flow for this replication was identical to that in Study 1, except for the inclusion of nine science literacy questions *(2)*, two questions about participants' own science education, and three questions about God, intuition, and science at the end of the demographics screen. This questionnaire and all subsequent questionnaires (unless otherwise noted) were hosted on Qualtrics. A total of 58 participants in this replication were determined to have participated in the first experiment and were coded for exclusion, resulting in a final n = 343.

The third replication in Study 2 was hosted by Pollfish, which recruited 825 participants directly (rather than through MTurk). We excluded all 146 survey responses from participants who were inadvertently permitted to take the survey more than once, resulting in a final n = 679. Because Pollfish is a marketing company that exclusively recruits consumers via mobile devices, we made minor changes in the design of our questionnaire when adapting the version run on Qualtrics.

First, because Pollfish collects demographic data on its users, participants were only asked to rate one of the four vignettes, to provide a written explanation of their ratings, and to answer one question about God, one about intuition, and one about science. Due to concerns that mobile participants would be especially inclined to provide ratings that did not require any scrolling on their devices, we randomized participants to be presented with rating options in either increasing or decreasing order of appropriateness. Additionally, Pollfish's interface required that we present each questionnaire item on a different screen (Fig. 1).

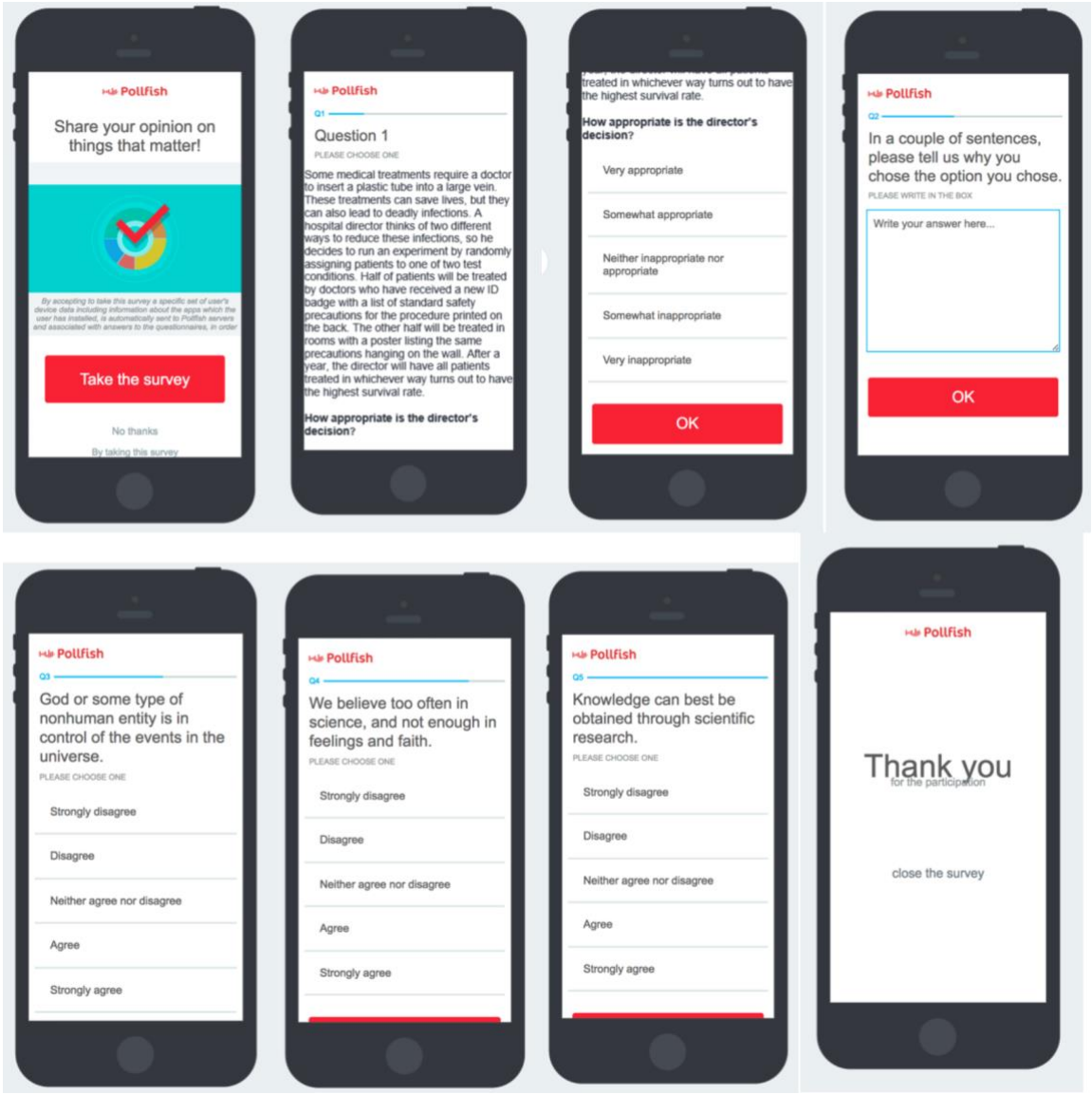


Fig. S1. Survey appearance in Pollfish

Because of concerns that mobile users might simply select the first option on their screens, half of these participants were presented with rating options in ascending order of appropriateness, and half with options in descending order.

Study 3 was conducted in two parts with 2,312 MTurk participants (n = 2,270 after exclusions). In the first part, 803 participants (n = 775 after exclusion of repeat participants from Study One) were paid $0.35 each and randomly assigned one of eight vignettes. These vignettes were comprised of two policy conditions each concerning autonomous vehicles and genetic testing, and two A/B conditions each for both of these domains. After rating and explaining their ratings of these vignettes, participants saw five more screens. These consisted of nine science literacy questions, two questions, about science education, three questions about God, intuition, and science, five demographic questions, and a thank-you screen with a link to their completion code for payment.

In the second part of Study 3, 1,509 participants were paid $0.40 each and randomly assigned to one of fifteen vignettes (n = 1,495 after exclusions of repeat participants). Note that we discontinued A/B short during Study 3 after confirming that the two experimental conditions did not differ from each other in any case where both were run. Therefore, the vignettes run in the second part of Study 3 (and all subsequent studies) included two policy conditions and one A/B condition. In Study 3, these conditions were run each in the following domains: Retirement Plans, Teacher Wellbeing, Poverty Alleviation, Health Worker Recruitment, and Basic Income. Participants then proceeded through screens and items identical to those in the first part of Study 3.

Study 4 was conducted on 304 participants recruited from MTurk (n = 302 after exclusion of repeats). Participants were compensated $0.40 each to provide 13 responses across five screens. The first two responses were ratings and explanations of one of three vignettes related to the prescription one of two nameless, indistinguishable blood pressure drugs ("Drug A" and "Drug B"). On the next four screens, there were nine questions related to science literacy, two questions about science education, three questions about belief in God, intuition, and science, and five demographic questions. Participants then saw a screen thanking them for their participation, and followed a link to a separate screen with a code for payment.

Study 5 was identical to Study 4, except that in Study 5 (n = 1027; final n = 1023), the vignette described prescription of Drug A vs. Drug B as essentially random. In the A and B conditions of Study 5, blood pressure patients were treated by a doctor who assigned Drug A to all of his patients or by a doctor who assigned Drug B to all of his patients at a walk-in clinic. However, their assignment to that particular doctor was arbitrary, a result of whatever doctor happened to be available at the time they walked into the clinic.

First, 307 participants were recruited via MTurk for a survey hosted on Qualtrics (after exclusion of repeats, final n = 303). After results of this experiment were analyzed, 720 additional participants were recruited to participate on Pollfish. Exactly half saw rating options in ascending order, while the other half saw them in descending order.

Study 6 was conducted on 457 healthcare providers (physicians, physician assistants, nurse practitioners, and nurses) employed by Geisinger, an integrated health system in Pennsylvania and New Jersey. As part of a larger survey about unrelated topics, each of these participants was shown one condition each from the Safety Checklist and Drug Effectiveness Walk-in vignettes used in studies described above, separated by several questions about the other topics of the survey. Participants were incentivized by the chance to win one of eight $50 Amazon gift cards. Randomization assigned each participant to view a policy condition (A or B) of one scenario and the A/B condition of the other scenario. Here, for comparability to the purely between-subjects design of Studies 1–5, we report results for only the first vignette that each participant saw. In addition to the A/B effect studies reported here, we collected other, nonoverlapping, data during this survey. One part of this survey solicited social-perceptual ratings of a hypothetical patient's decision to ignore or learn genetic health information (reported in *3*), while another measured Geisinger providers' attitudes toward genetic testing for actionable health results and nudges designed to increase its uptake (reported in *4*).

In all scenarios across all six studies, A/B condition vignettes were longer than those in the A or B policy condition, though reading level was sometimes lower and sometimes higher in the A/B conditions (Table S1).

Table S1. Readability statistics for all experimental vignettes

Scenario	Condition	Words	Words per sentence	Flesch-Kincaid grade level
Safety Checklist	A	83	20.7	10.1
Safety Checklist	B	75	18.7	9.3
Safety Checklist	A/B learn	107	20.4	10.2
Safety Checklist	A/B short	126	21.0	10.2
Safety Checklist 2	A	116	19.3	9.4
Safety Checklist 2	B	113	18.8	8.9
Safety Checklist 2	A/B learn	136	19.4	9.2
Safety Checklist 2	A/B short	164	20.5	9.8
Autonomous Vehicles	A	100	25.0	12.5
Autonomous Vehicles	B	104	26.0	13.2
Autonomous Vehicles	A/B learn	157	26.1	13.1
Autonomous Vehicles	A/B short	181	25.8	12.8
Genetic Testing	A	115	23.0	11.7
Genetic Testing	B	120	24.0	12.1
Genetic Testing	A/B learn	173	24.7	12.4
Genetic Testing	A/B short	196	24.5	12.5
Retirement Plans	A	71	23.6	12.7
Retirement Plans	B	76	25.3	13.5
Retirement Plans	A/B learn	142	23.6	12.5
Health Worker Recruitment	A	90	30.0	14.4
Health Worker Recruitment	B	88	29.3	15.1
Health Worker Recruitment	A/B learn	167	27.8	14.3
Poverty Alleviation	A	45	22.5	11.8
Poverty Alleviation	B	49	24.5	12.0
Poverty Alleviation	A/B learn	105	21.0	11.0
Teacher Wellbeing	A	43	21.5	12.0
Teacher Wellbeing	B	48	24.0	13.6
Teacher Wellbeing	A/B learn	96	19.2	11.4
Basic Income	A	88	22.0	10.2
Basic Income	B	88	22.0	10.0
Basic Income	A/B learn	166	27.6	12.9
Drug Effectiveness	A	58	16.6	9.0
Drug Effectiveness	B	58	19.3	9.6
Drug Effectiveness	A/B learn	107	21.4	10.2
Drug Effectiveness (Walk-in)	A	91	15.1	8.3
Drug Effectiveness (Walk-in)	B	91	15.1	8.2
Drug Effectiveness (Walk-in)	A/B learn	139	19.8	9.8

Materials Pretesting and Pilot Testing

In the process of conducting Studies 3–5 (all domains but Safety Checklist) we developed stimuli by iteratively *pretesting* policy conditions and *pilot testing* these policies in comparison with an A/B test condition. We first *pretested* several policy vignettes with small samples (~20–30 participants per condition). The purpose of this preliminary materials testing was to ensure that both policies were themselves rated as above the midpoint in appropriateness (that is, that the policies were not thought to be unreasonable), that both were rated as similarly appropriate (to mimic the requirement of equipoise that is a precondition for the effect as we have defined it), and that participants understood the materials.

Policies that met the pretesting stage criteria were then *pilot tested* by comparing the rated appropriateness of each of these policies against the appropriateness rating of an A/B test condition (~20–30 participants per condition). Appropriateness ratings were typically collected independently, though in some cases participants rated several policies or A/B test conditions (see preregistrations for details).

Of those sets of vignettes that reached the pilot testing stage, eight were abandoned at this stage and not tested in full samples. In one of those sets, one of the policies was rated as substantially more inappropriate than the other (Autonomous Vehicles Pilot 1 – yellow lights). In the remaining seven cases (Online Dating Pilot 1, Online Dating Pilot 4, Simple SUPPORT 1, Genetic Testing Pilot 3, Colonoscopies Pilot, Resident Hours Pilot, and Basic Income Pilot 1) the policy conditions were rated as reasonably similar in appropriateness, and at or above the scale midpoint, but we did not observe a large, immediately apparent A/B effect. Note that these pilot tests were not appropriately powered to detect small or medium effects, and so we cannot infer the presence or absence of an effect based on them. We conducted additional analyses to ensure that our results were robust to potential bias introduced by proceeding with some pilot studies and not others. Two hierarchical linear models, which found that our effects remained significant even after including the data from vignettes that were pilot tested, are described at the end of this supplement (see pp. 102–105). Finally, we note that we do not expect all descriptions of possible A/B tests to result in an effect. Learning where the boundary conditions are for people's objections to A/B testing is a task for future research.

Full text of all pilots is presented after the full text of all full experiments. All vignettes are presented in the chronological order in which they were run. Vignettes marked with a * appeared verbatim in subsequently run experiments listed above, and their full text appears below in duplicate in order to help illustrate pilot chronology. **Tables S68–S69 display detailed information for all pilot testing and materials pretesting.**

Materials Pretesting and Pilot Testing: Participants and Procedure

After replicating our initial findings with Study 2, we conducted a series of preregistered pilots to explore other candidate domains in which to study the effect. 83 MTurk participants rated one of four randomly assigned vignettes concerning the algorithm implemented or tested by an online dating site. Participants in this pilot were paid $0.25, reported their demographic information as in our main studies, and answered three questions about God, intuition, and science.

For our next pilot, we recruited 82 MTurk participants to rate one of four vignettes about self-driving vehicles. These four vignettes included one each involving the unilateral instantiation of one of two self-driving algorithms, and two different vignettes in which

assignment of algorithms to cars was randomized as part of an A/B test. Participants in this pilot were paid $0.25, reported their demographic information as in our main studies, and answered three questions about God, intuition, and science.

Next, 363 MTurk participants were recruited to rate one of 12 vignettes. These vignettes were comprised of two policy conditions each (without a corresponding A/B condition), across three domains, where the vignettes in each domain were worded in two different ways. Participants did not provide any information besides rating and explanation, and were paid $0.15 each.

Following this, an additional 363 MTurk participants were recruited to rate one of 12 vignettes. These included two policy conditions and one A/B condition concerning autonomous vehicles, online dating algorithms, and two sets of genetic testing vignettes. Participants were paid $0.15 each for their participation and were asked only to rate the vignettes and explain their ratings.

We then paid 183 MTurk participants $0.80 each to rate and explain their ratings of five vignettes, and to answer five demographic questions. Each participant in this study was presented with one of two policy vignettes in each of five domains, from among 15 domains. No A/B conditions were included. We later paid 331 new participants $0.25 to rate and explain ratings of an A/B test in 11 of these 15 domains. Responses to five demographic questions were collected in this pilot as well.

Hierarchical linear model

Pilot testing was used as bases for experiments only in cases where it seemed that the difference between policy conditions and A/B conditions might be noteworthy. To control for any selection bias or multiple testing concerns that may have been introduced in this process, and more generally because of the large number and diversity of vignettes in our experiments and pilots, we conducted a hierarchical linear model to verify whether the effect remained significant across all participants in our research. Details of this model are available in the section entitled "Robustness Checks on Selection Bias and Multiple Comparisons."

Materials

Below is the complete set of materials from all pretesting, pilots, experiments, and replications. Preregistered documents containing all materials can be accessed via our archive on the Open Science Framework (OSF) (https://osf.io/5y4f9/). Full details of survey flow and visual appearance of the questionnaires can be gathered from those OSF materials.

Probes
Beneath each vignette was a list of ratings ranging from 'Very inappropriate' (recorded as '1') to 'Very appropriate' (recorded as '5'):

- O Very inappropriate
- O Somewhat inappropriate
- O Neither inappropriate nor appropriate
- O Somewhat appropriate
- O Very appropriate

Rating options were presented in the order above to all MTurk participants. They were also presented in the above order to half of Pollfish participants, and in the reverse order to the other half of Pollfish participants. Participants were required to select exactly one rating from the list above. They were also required to write in a response box below the list of ratings. Above the response box was the following prompt:

> In a couple of sentences, please tell us why you chose the option you chose.

Experiment Vignettes
Below is the full text for each experimental or pilot vignette, as copied directly from the SurveyMonkey, Qualtrics, and Pollfish questionnaires.

Study 1

Safety Checklist

A
Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director wants to reduce these infections, so he decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure printed on the back. All patients having this procedure will then be treated by doctors with this list attached to their clothing.
How appropriate is the director's decision?

B

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director wants to reduce these infections, so he decides to hang a poster with a list of standard safety precautions for this procedure in all procedure rooms. All patients having this procedure will then be treated in rooms with this list posted on the wall.
How appropriate is the director's decision?

A/B short
Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by a doctor who has received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in a room with a poster listing the same precautions hanging on the wall.
How appropriate is the director's decision?

A/B learn
Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall. After a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.
How appropriate is the director's decision?

Study 2

First Replication

Vignette materials and questions were identical to those above.

Pollfish Replication (mobile users)

Vignette materials and questions were identical to those above.

Alternate Replication (with wording changes)

Safety Checklist 2

A

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director comes up with an idea to reduce these infections. He decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure printed on the back. All doctors performing this procedure will then have this list attached to their clothing, so they can look at it while performing the procedure. The director thinks that these new badges might help doctors remember all of the safety steps they were trained to take during the procedure.

How appropriate is the director's decision?

B

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director comes up with an idea to reduce these infections. He decides to hang a poster with a list of standard safety precautions for this procedure in all the rooms where it is performed. All doctors performing this procedure will then work in rooms with the poster on the wall, so they can look at it while performing the procedure. The director thinks that these posters might help doctors remember all of the safety steps they were trained to take during the procedure.

How appropriate is the director's decision?

A/B short

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director comes up with two ideas to reduce these infections. He decides to run an experiment to compare these two ideas by randomly assigning patients to one of two groups. Half of the patients will be treated by a doctor who has received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in a room with a poster listing the same precautions hanging on the wall. The director thinks that these badges and posters might help doctors remember all of the safety steps they were trained to take during the procedure.

How appropriate is the director's decision?

A/B learn

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director comes up with two ideas to reduce these infections. He decides to run an experiment to test compare two ideas by randomly assigning patients to one of two groups. Half of the patients will be treated by a doctor who has received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in a room with a poster listing the same precautions hanging on the wall. The director thinks that these badges and posters might help doctors remember all of the safety steps they were trained to take during the

procedure. After a year, the director will check which option, badges or posters, is most effective, and make it standard for all patients and doctors throughout the entire hospital.
How appropriate is the director's decision?

Study 3

Genetic Testing

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.
How appropriate is the CEO's decision?

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.
How appropriate is the CEO's decision?

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.
How appropriate is the CEO's decision?

A/B learn

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 one year, the CEO will provide all new customers with the option that led to the highest customer satisfaction during the experiment.

How appropriate is the CEO's decision?

Autonomous Vehicles

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.

How appropriate is the CEO's decision?

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.

How appropriate is the CEO's decision?

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.
How appropriate is the CEO's decision?

A/B learn
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.
How appropriate is the CEO's decision?

Retirement Plans

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.
How appropriate is the CEO's decision?

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.
How appropriate is the CEO's decision?

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.
How appropriate is the CEO's decision?

Health Worker Recruitment

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.
How appropriate is the congress's decision?

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.
How appropriate is the congress's decision?

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.
How appropriate is the congress's decision?

Poverty Alleviation

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.
How appropriate is the director's decision?

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.
How appropriate is the director's decision?

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.
How appropriate is the director's decision?

Teacher Wellbeing

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.
How appropriate is the superintendent's decision?

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.
How appropriate is the superintendent's decision?

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.
How appropriate is the superintendent's decision?

Basic Income

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.
How appropriate is the congress's decision?

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.
How appropriate is the congress's decision?

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.
How appropriate is the congress's decision?

Study 4

Drug Effectiveness

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.
How appropriate is Doctor Jones's decision?

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.
How appropriate is Doctor Jones's decision?

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.
How appropriate is Doctor Jones's decision?

Study 5

Drug Effectiveness Walk-in

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.
How appropriate is Doctor Jones's decision?

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.
How appropriate is Doctor Jones's decision?

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.

How appropriate is Doctor Jones's decision?

Studies 6a and 6b

Materials were identical to those used in Study 1 and Study 5, as described above.

Pretesting and Pilot Vignettes

See pp. 8–9 for explanation and justification of material pretesting and pilot testing.

Online Dating Pilot 1

A

Customers who sign up for an online dating service answer questions about their tastes and habits. The system enables its users to search and view every member's profile and lets them contact whomever they want. The CEO of the dating service wants to improve customer satisfaction by suggesting potential matches to all users. He thinks of a way to do this. So, based on the idea that "opposites attract,"— so long as they aren't *too* different—he decides to program his website to suggest users are a match when they answer 61–80% of questions in the same way.

How appropriate is the CEO's decision?

B

Customers who sign up for an online dating service answer questions about their tastes and habits. The system enables its users to search and view every member's profile and lets them contact whomever they want. The CEO of the dating service wants to improve customer satisfaction by suggesting potential matches to all users. He thinks of a way to do this. So, based on the idea that "birds of a feather flock together," he decides to program his website to suggest users are a match when they answer 81–100% of questions in the same way.

How appropriate is the CEO's decision?

A/B short

Customers who sign up for an online dating service answer questions about their tastes and habits. The system enables its users to search and view every member's profile and lets them

contact whomever they want. The CEO of the dating service wants to improve customer satisfaction by suggesting potential matches to all users. He thinks of two ways to do this. So, he decides to run an experiment by randomly assigning customers to one of two test conditions. For half of users, based on the idea that "opposites attract,"—so long as they aren't *too* different—he programs his website to suggest users are a match when they answer 61–80% of questions in the same way. For the other half of users, based on the idea that "birds of a feather flock together," he programs his website to suggest users are a match when they answer 81–100% of questions in the same way.
How appropriate is the CEO's decision?

A/B learn
Customers who sign up for an online dating service answer questions about their tastes and habits. The system enables its users to search and view every member's profile and lets them contact whomever they want. The CEO of the dating service wants to improve customer satisfaction by suggesting potential matches to all users. He thinks of two ways to do this. So, he decides to run an experiment by randomly assigning customers to one of two test conditions. For half of users, based on the idea that "opposites attract,"—so long as they aren't *too* different—he programs his website to suggest users are a match when they answer 61–80% of questions in the same way. For the other half of users, based on the idea that "birds of a feather flock together," he programs his website to suggest users are a match when they answer 81–100% of questions in the same way. After one year, all users will be matched according to whichever program results in greater customer satisfaction.
How appropriate is the CEO's decision?

Autonomous Vehicles Pilot 1

A
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of a way to do this. He decides that in order to reduce rear-end collisions in heavy traffic, all of the company's cars will be programmed to accelerate whenever they are close to a stoplight as it turns yellow.
How appropriate is the director's decision?

B
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of a way to do this. He decides that in order to reduce side-impact collisions in heavy traffic, all of the

company's cars will be programmed to brake whenever they are close to a stoplight as it turns yellow.
How appropriate is the director's decision?

A/B short
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of two ways to do this. So, he decides to run an experiment by randomly assigning cars to be programmed in one of two ways. Half of the company's cars will be programmed to brake whenever they are close to a stoplight as it turns yellow, in order to reduce side-impact collisions in heavy traffic. The other cars will be programmed to accelerate whenever they are close to a stoplight as it turns yellow, in order to reduce rear-end collisions in heavy traffic.
How appropriate is the director's decision?

A/B learn
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of two ways to do this. So, he decides to run an experiment by randomly assigning cars to be programmed in one of two ways. Half of the company's cars will be programmed to brake whenever they are close to a stoplight as it turns yellow, in order to reduce side-impact collisions in heavy traffic. The other cars will be programmed to accelerate whenever they are close to a stoplight as it turns yellow, in order to reduce rear-end collisions in heavy traffic. After the self-driving cars have been on the road for one year, all new cars will be programmed to use whichever safety measure was more effective.
How appropriate is the director's decision?

Autonomous Vehicles Pilot 2

A
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of a way to do this. He decides that in order to reduce collisions in heavy traffic, all of the company's cars will be programmed to match the accelerating or braking behavior of nearby cars when they are close to a traffic light as it turns yellow.
How appropriate is the director's decision?

B
Self-driving cars, which are currently being developed by several companies, are programmed to control all aspects of driving without any human involvement. However, there are a number of difficult problems in programming self-driving cars. For instance, cars in heavy traffic face significant danger at yellow traffic lights. The director of software engineering at a company developing self-driving cars wants to reduce accidents at yellow lights, and he thinks of a way to do this. He decides that in order to reduce collisions in heavy traffic, all of the company's cars will be programmed to brake when they are close to a traffic light as it turns yellow.
How appropriate is the director's decision?

Online Dating Pilot 2

A
Customers who sign up for an online dating service are matched with each other based on their answers to questions about their tastes and habits. The system is set up to only allow customers to contact people who are recommended as a match. The CEO of the dating service wants to improve customer satisfaction by introducing a new way to match people together. He thinks of a way to do this. So, he decides to program his website to report users are a match when they are very similar to each other, based on the questions they answered.
How appropriate is the CEO's decision?

B
Customers who sign up for an online dating service are matched with each other based on their answers to questions about their tastes and habits. The system is set up to only allow customers to contact people who are recommended as a match. The CEO of the dating service wants to improve customer satisfaction by introducing a new way to match people together. He thinks of a way to do this. So, he decides to program his website to report users are a match when they are **somewhat** similar (but not TOO similar) to each other, based on the questions they answered.
How appropriate is the CEO's decision?

Genetic Testing Pilot 1

A
People sometimes get genetically tested to learn more about their ancestry. However, genetic testing can also reveal important health risks that people wouldn't otherwise learn they have. The research director at a popular genetic testing company wants to help customers learn more about their health risks, but the company's testing system only collects enough genetic material to run one health-related test. So, the research director decides that in addition to testing for ancestry, the company will also test customers for their risk of developing certain kinds of cancer later in life. Customers will have the option of viewing these results or not.
How appropriate is the CEO's decision?

B
People sometimes get genetically tested to learn more about their ancestry. However, genetic testing can also reveal important health risks that people wouldn't otherwise learn they have. The research director at a popular genetic testing company wants to help customers learn more about their health risks, but the company's testing system only collects enough genetic material to run one health-related test. So, the research director decides that in addition to testing for ancestry, the company will also test customers for their risk of developing dementia later in life. Customers will have the option of viewing these results or not.
How appropriate is the CEO's decision?

Genetic Testing Pilot 2

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 the genetic mutation, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origin, 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.
How appropriate is the CEO's decision?

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 the genetic mutation, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origin, 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.
How appropriate is the CEO's decision?

Autonomous Vehicles Pilot 3

A
Many people like the idea of completely self-driving cars, which are capable of navigating the road without 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 an emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible 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.
How appropriate is the CEO's decision?

B
Many people like the idea of completely self-driving cars, which are capable of navigating the road without 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 an emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible 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.
How appropriate is the CEO's decision?

Online Dating Pilot 3

A
The founders of a new dating app are deciding how to suggest potential matches to its users. All users are able to view every other user's profile and message him or her, but every day the app suggests to each user one new "match," or person that he or she might like. The founders come up with a computer program for generating these matches. They decide that match suggestions will be based on the number of friends and friends-of-friends that users have in common on Facebook.
How appropriate is the CEO's decision?

B
The founders of a new dating app are deciding how to suggest potential matches to its users. All users are able to view every other user's profile and message him or her, but every day the app suggests to each user one new "match," or person that he or she might like. The founders come up with a computer program for generating these matches. They decide that match suggestions will be based on the percentage of "profile questions" that users answer in the same way.
How appropriate is the CEO's decision?

Online Dating Pilot 4

A
Customers who sign up for an online dating service are matched with each other based on their answers to questions about their tastes and habits. The system is set up to only allow customers to contact people who are recommended as a match. The CEO of the dating service wants to improve customer satisfaction by introducing a new way to match people together. He thinks of a way to do this. So, he decides to program his website to report users are a match when they are very similar to each other, based on the questions they answered.
How appropriate is the CEO's decision?

B
Customers who sign up for an online dating service are matched with each other based on their answers to questions about their tastes and habits. The system is set up to only allow customers

to contact people who are recommended as a match. The CEO of the dating service wants to improve customer satisfaction by introducing a new way to match people together. He thinks of a way to do this. So, he decides to program his website to report users are a match when they are **somewhat** similar (but not TOO similar) to each other, based on the questions they answered.
How appropriate is the CEO's decision?

A/B short
Customers who sign up for an online dating service are matched with each other based on their answers to questions about their tastes and habits. The system is set up to only allow customers to contact people who are recommended as a match. The CEO of the dating service wants to improve customer satisfaction by introducing a new way to match people together. He thinks of two ways to do this. So, he decides to run an experiment by randomly assigning customers to one of two test conditions. For half of customers, his website will report users are a match when they are very similar to each other, based on the questions they answered. For the other half, his website will report users are a match when they are somewhat similar (but not TOO similar) to each other, based on the questions they answered.
How appropriate is the CEO's decision?

Genetic Testing Pilot 3

A
People sometimes get genetically tested to learn more about their ancestry. However, genetic testing can also reveal important health risks that people wouldn't otherwise learn they have. The research director at a popular genetic testing company wants to help customers learn more about their health risks, but the company's testing system only collects enough genetic material to run one health-related test. So, the research director decides that in addition to testing for ancestry, the company will also test customers for their risk of developing certain kinds of cancer later in life. All customers will have the option of viewing their result or not.
How appropriate is the CEO's decision?

B
People sometimes get genetically tested to learn more about their ancestry. However, genetic testing can also reveal important health risks that people wouldn't otherwise learn they have. The research director at a popular genetic testing company wants to help customers learn more about their health risks, but the company's testing system only collects enough genetic material to run one health-related test. So, the research director decides that in addition to testing for ancestry, the company will also test customers for their risk of developing dementia later in life. All customers will have the option of viewing their result or not.
How appropriate is the CEO's decision?

A/B short
People sometimes get genetically tested to learn more about their ancestry. However, genetic testing can also reveal important health risks that people wouldn't otherwise learn they have. The research director at a popular genetic testing company wants to help customers learn more about their health risks, but the company's testing system only collects enough genetic material to run

one health-related test. So, the research director decides to run an experiment by randomly assigning customers to one of two test conditions. In addition to testing for ancestry, the company will test half of customers for their risk of developing dementia later in life. The company will test the other half of customers for their risk of developing certain kinds of cancer later in life. All customers will have the option of viewing their result or not.
How appropriate is the CEO's decision?

Genetic Testing Pilot 4

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 the genetic mutation, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origin, 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.
How appropriate is the CEO's decision?

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 the genetic mutation, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origin, 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.
How appropriate is the CEO's decision?

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 the genetic mutation, but others cannot. A certain genetic testing company currently only returns "genealogy" results, about customers' family tree and national origin, 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.
How appropriate is the CEO's decision?

Autonomous Vehicles Pilot 4

A*
Many people like the idea of completely self-driving cars, which are capable of navigating the road without 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 an emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible 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.
How appropriate is the CEO's decision?

B*
Many people like the idea of completely self-driving cars, which are capable of navigating the road without 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 an emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible 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.
How appropriate is the CEO's decision?

A/B short*
Many people like the idea of completely self-driving cars, which are capable of navigating the road without 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 an emergency. The CEO of a company developing self-driving cars wants people to have as much freedom as possible 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.
How appropriate is the CEO's decision?

Resident Hours Pilot 1

A
"First-year residents" are new doctors who practice medicine under the supervision of more experienced doctors. The body that accredits U.S. hospital residency programs has already determined that first-year residents must work no more than 80 hours per week, averaged over 4 weeks. Now it must decide, within that constraint, how long first-year residents may work in any one shift. Second-year and later residents may work up to 24 hours at a time. The accrediting body is concerned that residents working longer hours might get less sleep and that sleep-deprived residents might make errors that hurt patients. The accrediting body wants to adopt the

policy that will best protect patients, so it decides that first-year residents at all hospitals it accredits may work no more than 16 hours at a time.
How appropriate is the accrediting body's decision?

B
"First-year residents" are new doctors who practice medicine under the supervision of more experienced doctors. The body that accredits U.S. hospital residency programs has already determined that first-year residents must work no more than 80 hours per week, averaged over 4 weeks. Now it must decide, within that constraint, how long first-year residents may work in any one shift. Second-year and later residents may work up to 24 hours at a time. The accrediting body is concerned that shorter work hours mean more patient hand-offs, which are dangerous because it is easy for important patient information not to be relayed between care teams. Shorter hours might also mean less education, or socialization into a kind of "shift mentality" that reduces professionalism—either of which might result in less competent and less committed doctors for patients in the future. The accrediting body wants to adopt the policy that will best protect patients, so it decides that first-year residents at all hospitals it accredits may work no more than 24 hours at a time.
How appropriate is the accrediting body's decision?

Drug Effectiveness Walk-in Pilot 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 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 whenever his patients need blood pressure medication, they will be prescribed drug A.
How appropriate is Doctor Jones' decision?

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 wants to provide good treatment to his patients, so he decides that whenever his patients need blood pressure medication, they will be prescribed drug B.
How appropriate is Doctor Jones' decision?

Drug Effectiveness Pilot 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 to prescribe drug A for high blood pressure. It is affordable and patients can tolerate its side effects.
How appropriate is Doctor Jones' decision?

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 to prescribe drug B for high blood pressure. It is affordable and patients can tolerate its side effects.
How appropriate is Doctor Jones' decision?

Basic Income Pilot 1

A
A state government's department of human services is tasked with distributing welfare and unemployment resources to struggling families. The head of this department wants to reduce poverty and unemployment, so he decides to increase welfare services offered to the poorest families in the state: those who are the farthest below the poverty line.
How appropriate is the department head's decision?

B
A state government's department of human services is tasked with distributing welfare and unemployment resources to struggling families. The head of this department wants to reduce poverty and unemployment, so he decides to increase welfare services equally to everyone in the state below the poverty line.
How appropriate is the department head's decision?

Retirement Plans Pilot 1

A* (with minor edit)
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 his employees to enroll in the company retirement savings plan, so he decides to increase the number of available investment funds from 10 to 15, in the hope that employees will find some of the new options more attractive.
How appropriate is the CEO's decision?
Note: In the A/B condition pilot (Retirement Plans Pilot 3) and full experiment (Retirement Plans) that used this policy, we removed the text, "in the hope that employees will find some of the new options more attractive."

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 his employees to enroll in the company retirement savings plan, so he decides to decrease the number of available investment funds from 10 to 5, in the hope that employees won't be overwhelmed by too many choices and suffer from "decision paralysis."
How appropriate is the CEO's decision?

Teacher Wellbeing Pilot 1

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 to give all of the school's teachers a yearly bonus.
How appropriate is the superintendent's decision?

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 to give all of the school's teachers additional vacation days during summer and winter breaks.
How appropriate is the superintendent's decision?

Health Worker Recruitment Pilot 1

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 the public and being a leader in one's community.
How appropriate is the congress' decision?

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.
How appropriate is the congress' decision?

Colonoscopies Pilot 1

A
The American Cancer Society recommends that starting at age 50, healthy adults should get colonoscopies every 10 years. Colonoscopies can save lives by detecting colon cancer early, but they are not fun and many people avoid them. The CEO of a health system would like to increase patients' likelihood of having a potentially life-saving colonoscopy, so he decides to send every patient a birthday card on their 50th birthday, which reads "It's your 50th birthday – treat yourself to a colonoscopy!" The card will come with coupons for a bowel cleanser that patients need to drink before a colonoscopy, and a book of crossword puzzles to entertain them while they're on the toilet. The idea is that this little bit of whimsy might make the procedure seem less "icky" and lead more patients to get colonoscopies.
How appropriate is the CEO's decision?

B
The American Cancer Society recommends that starting at age 50, healthy adults should get colonoscopies every 10 years. Colonoscopies can save lives by detecting colon cancer early, but they are not fun and many people avoid them. The CEO of a health system would like to increase patients' likelihood of having a potentially life-saving colonoscopy, so he decides to send every patient a birthday card on their 50th birthday, which reads "It's your 50th birthday – treat yourself to a colonoscopy!" The card will come with a notice of a pre-scheduled colonoscopy appointment, with options for selecting a more convenient time. Patients won't be charged if they don't show up. The idea is that removing the chore of scheduling will lead more patients to get colonoscopies.
How appropriate is the CEO's decision?

Poverty Alleviation Pilot 1

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 to spend this money on providing everyone below a certain income level with one piece of livestock (such as a donkey or a goat).
How appropriate is the director's decision?

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 to spend this money on providing everyone below a certain income level with a sturdy roof for their home.
How appropriate is the director's decision?

Music Streaming Pilot 1

A
The CEO of a music-streaming company wants to increase the likelihood that users of the company's free service will pay to upgrade to its Premium service. So, he decides to increase the number of advertisements that run on the free service.
How appropriate is the CEO's decision?

B
The CEO of a music-streaming company wants to increase the likelihood that users of the company's free service will pay to upgrade to its Premium service. So, he decides to remove several of the most popular songs from the free service.
How appropriate is the CEO's decision?

Poverty Alleviation Pilot 2

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 to spend this money on providing all adults below a certain income level with a cash payment equal to the cost of three months of food.
How appropriate is the director's decision?

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 to spend this money on providing all adults below a certain income level with one month of training in a trade of their choice.
How appropriate is the director's decision?

Retirement Plans Pilot 2

A* (with minor edit)
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 the enrollment paperwork will highlight the most popular of the three investment funds the company currently offers.
How appropriate is the CEO's decision?
Note: In the A/B condition pilot (Retirement Plans Pilot 3) and full experiment (Retirement Plans) that used this policy, we changed the text, "so he decides that the enrollment paperwork will highlight the most popular of the three investment funds the company currently offers," to instead say, "so he decides that he will provide enrollment paperwork that highlights the most popular of the 10 investment funds the company offers."

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 to give them enrollment paperwork in which they are automaticlaly [sic] enrolled in the most popular of the three investment funds the company offers. Employees are then free to make a difference choice or to decline the program entirely.
How appropriate is the CEO's decision?

Music Streaming Pilot 2

A
The CEO of a music-streaming company wants to increase the likelihood that users of the company's free service will pay to upgrade to its Premium service. He thinks that running short but frequent advertisements will encourage users to upgrade, so he decides to place a short, 30 second advertisement after every 2nd song users listen to.
How appropriate is the CEO's decision?

B
The CEO of a music-streaming company wants to increase the likelihood that users of the company's free service will pay to upgrade to its Premium service. He thinks that running infrequent but long advertisements will encourage users to upgrade, so he decides to place a long, 2 minute block of advertisements after every 8th song users listen to.
How appropriate is the CEO's decision?

Basic Income Pilot 2

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.
How appropriate is the congress's decision?

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.
How appropriate is the congress's decision?

Simple SUPPORT Pilot 1

A
In order to survive, babies born very prematurely need special machines that control how much oxygen is in their blood. Doctors know that both too little and too much oxygen can hurt these babies. They know that oxygen levels below 85% are too low, and they know that oxygen levels above 95% are too high. However, doctors do not know what level is best between 85% and 95%. The head of a hospital unit that cares for premature babies wants to give the babies the best chance of survival possible, so he decides to set all of the machines in his unit to keep these babies' oxygen levels between 85% and 89%.
How appropriate is the hospital unit head's decision?

B
In order to survive, babies born very prematurely need special machines that control how much oxygen is in their blood. Doctors know that both too little and too much oxygen can hurt these babies. They know that oxygen levels below 85% are too low, and they know that oxygen levels above 95% are too high. However, doctors do not know what level is best between 85% and 95%. The head of a hospital unit that cares for premature babies wants to give the babies the best chance of survival possible, so he decides to set all of the machines in his unit to keep these babies' oxygen levels between 91% and 95%.
How appropriate is the hospital unit head's decision?

Resident Hours Pilot 2

A
"First-year residents" are new doctors who practice medicine under the supervision of more experienced doctors. The body that accredits U.S. hospital residency programs has already determined that first-year residents must work no more than 80 hours per week, averaged over 4 weeks. Now it must decide, within that constraint, how long first-year residents may work in any one shift. Second-year and later residents may work up to 24 hours at a time. The accrediting body is concerned that residents working longer hours might get less sleep and that sleep-deprived residents might make errors that hurt patients. The accrediting body wants to adopt the policy that will best protect patients, so it decides that first-year residents at all hospitals it accredits may work no more than 16 hours at a time.
How appropriate is the accrediting body's decision?

B
"First-year residents" are new doctors who practice medicine under the supervision of more experienced doctors. The body that accredits U.S. hospital residency programs has already determined that first-year residents must work no more than 80 hours per week, averaged over 4

weeks. Now it must decide, within that constraint, how long first-year residents may work in any one shift. Second-year and later residents may work up to 24 hours at a time. The accrediting body is concerned that shorter work hours mean more patient hand-offs, which are dangerous because it is easy for important patient information not to be relayed between care teams. Shorter hours might also mean less education, or socialization into a kind of "shift mentality" that reduces professionalism—either of which might result in less competent and less committed doctors for patients in the future. The accrediting body wants to adopt the policy that will best protect patients, so it decides that first-year residents at all hospitals it accredits may work no more than 24 hours at a time.
How appropriate is the accrediting body's decision?

A/B learn
"First-year residents" are new doctors who practice medicine under the supervision of more experienced doctors. The body that accredits U.S. hospital residency programs has already determined that first-year residents must work no more than 80 hours per week, averaged over 4 weeks. Now it must decide, within that constraint, how long first-year residents may work in any one shift. Second-year and later residents may work up to 24 hours at a time. The accrediting body thinks of two different policies that could best protect patients, so it decides to run an experiment by randomly assigning hospitals to one of two test conditions. Residents at half of the hospitals the body accredits will be assigned to work no more than 16 hours at a time. The accrediting body is concerned that residents working longer hours might get less sleep and that sleep-deprived residents might make errors that hurt patients. Residents at the other half of the hospitals will be assigned to work no more than 24 hours at a time. The accrediting body is concerned that shorter work hours mean more patient hand-offs, which are dangerous because it is easy for important patient information not to be relayed between care teams. Shorter hours might also mean less education, or socialization into a kind of "shift mentality" that reduces professionalism—either of which might result in less competent and less committed doctors for patients in the future. After a year, the accrediting body will require all hospitals it accredits to adopt whichever policy turns out to produce the best patient outcomes.
How appropriate is the accrediting body's decision?

Drug Effectiveness Pilot 2

A/B learn*
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 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.
How appropriate is Doctor Jones' decision?

Simple SUPPORT Pilot 2

A/B learn
In order to survive, babies born very prematurely need special machines that control how much oxygen is in their blood. Doctors know that both too little and too much oxygen can hurt these babies. They know that a saturation level below 85% is too low, and they know that a saturation level above 95% is too high. However, doctors do not know what level is best between 85% and 95%. The head of a hospital unit that cares for premature babies thinks of two different ways to give babies the best chance of survival, so he decides to run an experiment by randomly assigning babies in his unit to one of two test conditions. Half of the babies will be treated by machines set to keep their oxygen levels between 85% and 89%. The other half will be treated by machines set to keep their oxygen levels between 91% and 95%. After a year, he will set all oxygen machines of the babies he treats to whichever range turns out to have the highest survival rate.
How appropriate is the hospital unit head's decision?

Poverty Alleviation Pilot 3

A/B learn
A state government's department of human services is tasked with distributing welfare and unemployment resources to struggling families. The head of this department thinks of two different ways to reduce poverty and unemployment, so he decides to run an experiment by randomly assigning counties in his state to one of two test conditions. In half of counties, the state government will substantially increase welfare services offered to the poorest families: those who are the farthest below the poverty line. In the other half of counties, the state government will moderately increase welfare services equally to everyone below the poverty line. After a year, the state government will adopt for all counties in the state whichever policy turns out to best reduce poverty and unemployment.
How appropriate is the department head's decision?

Teacher Wellbeing Pilot 2

A/B learn*
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.
How appropriate is the superintendent's decision?

Health Worker Recruitment Pilot 2

A/B learn*
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, they 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.
How appropriate is the congress's decision?

Colonoscopies Pilot 2

A/B learn
The American Cancer Society recommends that starting at age 50, healthy adults should get colonoscopies every 10 years. Colonoscopies can save lives by detecting colon cancer early, but they are not fun and many people avoid them. The medical director of a hospital thinks of two different ways to increase patients' likelihood of having potentially life-saving colonoscopies, so he decides to run an experiment by randomly assigning every patient to one of two test conditions. All patients will receive a birthday card on their 50th birthday, which reads "It's your 50th birthday—treat yourself to a colonoscopy!" For half of patients, the card will come with coupons for a bowel cleanser that patients need to drink before a colonoscopy, and a book of crossword puzzles to entertain them while they're on the toilet. The idea is that this little bit of whimsy might make the procedure seem less "icky" and lead more patients to have colonoscopies. For the other half of patients, the card will come with a notice of a pre-scheduled colonoscopy appointment, with options for selecting a more convenient time. Patients won't be charged if they don't show up. The idea is that removing the chore of scheduling will lead more patients to have colonoscopies. After a year, the medical director will have all patients turning 50 receive whichever birthday package turns out to lead the most patients to have their recommended colonoscopies.
How appropriate is the medical director's decision?

Poverty Alleviation Pilot 4

A/B learn*
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.
How appropriate is the director's decision?

Basic Income Pilot 3

A/B learn*
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.
How appropriate is the congress's decision?

Retirement Plans Pilot 3

A/B learn*
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.
How appropriate is the CEO's decision?

Drug Walk-in Pilot 2

A/B learn*
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 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.

How appropriate is Doctor Jones' decision?

Additional questions

Listed below are the various scales and demographic questions used in different pilots and experiments as described above. Note that the science literacy scale used exact or adapted versions of questions from a scale developed by the National Science Board *(2)*. Where changes were made to these questions, they were made only to avoid asking participants further questions with obvious relationships to A/B testing on human participants.

Science literacy scale

1. 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?
O Give the new food to all 1,000 plants.
O Give the new food to 500 plants and give the normal, standard food to another 500 plants.

2. 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.
Does this mean that if their first child has the illness, the next three will not?
O No, the next three still might inherit the illness.
O Yes, none of the next three children will have the illness.

3. The center of the earth is very hot.
O True
O False

4. All radioactivity is man-made.
O True
O False

5. Lasers work by focusing sound waves.
O True
O False

6. Electrons are smaller than atoms.
O True
O False

7. The continents have been moving their location for millions of years and will continue to move.
O True
O False

8. It is the father's gene that decides whether the bay is a boy or a girl.
O True
O False

9. Antibiotics kill viruses as well as bacteria.
O True
O False

Science education questions

1. Do you have a college or graduate school degree in any science or engineering field?
O Yes
O No

2. If you answered 'yes' to the question above, please enter the specific science or engineering field below.
(Free response)

<u>God, Intuition, and Science (GIS) scale</u>

Please indicate how strongly you agree or disagree with the following statements.

1. God or some type of nonhuman entity is in control of the events in the universe.
O Strongly disagree
O Somewhat disagree
O Neither disagree nor agree
O Somewhat agree
O Strongly agree

2. We believe too often in science, and not enough in feelings and faith
O Strongly disagree
O Somewhat disagree
O Neither disagree nor agree
O Somewhat agree
O Strongly agree

3. Knowledge can be best obtained through scientific research
O Strongly disagree
O Somewhat disagree
O Neither disagree nor agree
O Somewhat agree
O Strongly agree

Demographic questions

1. What is your sex?
O Male
O Female
O Other

2. What race/ethnicity do you consider yourself? Check all that apply.
O Black/African-American
O Hispanic or Latino
O White
O Asian
O Other

3. What is your age (in years)?
(Free response)

4. What is the highest level of education you have received?
O Less than High School
O High School Degree
O Some College
O Four-Year College Degree
O Some Graduate School
O Graduate Degree

5. What is the total annual income from everyone in your household?
O Less than $20,000
O Between $20,000 and $40,000
O Between $40,000 and $60,000
O Between $60,000 and $80,000
O Between $80,000 and $100,000
O More than $100,000

Procedures

Qualitative Coding

For Studies 1, 2a, 4, and 5a, a list of codes was developed to categorize the content of participants' free text responses. To develop the codebook, a preliminary list of expected codes was created and preregistered. Following data collection of Study 1, one author read through each free response and noted themes not included in the preliminary list, which were then integrated to create a second, finer-grained version of the codebook.

Since Studies 4 and 5 used vignettes in different domains, and may have contained responses not captured by the codebook, one author read through the free text responses of the pilot studies pertaining to Studies 4 and 5. The author added two codes – Work Has Already Been Done and Not Enough Information – to the codebook, to be used specifically for these two studies. In the codebook provided below, these two codes were marked with a * to indicate that they were only used for Studies 4 and 5.

Two coders independently coded the free text responses of Studies 1, 2a (the direct replication of Study 1), 4, and 5a (the initial MTurk sample). Each discrepancy between coders was flagged, and coders met afterwards to discuss and resolve disagreements. The final, reconciled coding was used as the basis for all analyses. Overall interrater reliability was $\kappa = .83$, with respective interrater reliabilities for each study of $\kappa = .87$, $\kappa = .93$, $\kappa = .87$, and $\kappa = .66$.

The full list of codes indicating reasons given for appropriateness ratings is listed below. Note that the codebook references the catheterization scenario we ran in our first experiment. Codes for exclusions were adjusted only to reflect the details of each vignette, most of which did not involve this specific scenario (see note on exclusions section, p. 46, for details). Note also that in the codebook below, "Agent" refers to the primary decision-maker in a vignette. "Patient" refers to the intended beneficiary, recipient, or subject of a policy or test treatment and therefore may include not only medical patients but also, e.g., employees or consumers as "patients." "Expert" refers to an expert party other than the agent (e.g., line physicians in a vignette in which a hospital director is the agent).

Codebook

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 apply to any condition, and in A/B conditions, may or may not specifically mention randomization or experimentation as methodologies).

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 *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 *Physical risk/harm*: Comment that the intervention will or may place some or all patients at physical risk or will or may physically 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 or other stakeholders at non-physical risk or will or may harm some or all patients or other stakeholders in some non-physical way (e.g., by causing patients anxiety about whether their doctor is competent).

4. Ineffective

4.1. *Ineffective:* Comment that the intervention won't be, or might not be, effective in achieving the goal of reducing infections or helping patients (e.g., "won't work," "no point," "redundant" due to prior training).

4.2. *Work Has Already Been Done*:* Comments that research has already produced the results needed to achieve what the treatment or policy aims for, making the treatment or policy unnecessary.

5. Not Enough Information

5.1 *Not Enough Information**: Comment that the scenario does not provide enough information for the participant to make an informed judgment.

6. Research

6.1 *Randomization:* Specific mention of positive or negative aspects of 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"]).

6.2 *Experimentation:* Negative comments about experimentation, testing, practicing, 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.

6.3 *Inequality:* Negative comment that patients will be treated differently or unequally.

6.4 *No Equipoise:* Comment in A/B condition that one policy (A or B) is inferior to the other. Likely also coded as Inequality and Ineffective.

6.5 *Bad Design:* Comment that the A/B test lacks a control group, etc.

7. Consent

7.1 *Notice:* Comment on the importance of telling patients about the intervention.

7.2 *Consent:* Comment on the importance of patient choice to participate or not in the intervention

8. Action

8.1 *Too Long:* Comment on the temporal cost of research or of implementing an untested policy (e.g., conducting the experiment for one year before making a decision is too long).

8.2 *Pick a Policy:* Comment in an A/B condition that the director should immediately pick a policy rather than conducting an experiment (e.g., he should just use his best judgment, or just do what works best instead of running an experiment).

9. Intent

9.1 *Good intentions:* Comment that the director's intentions are, or his character is, good.

9.2 *Bad intentions:* Comment that the director's intentions are bad

10. Status Quo

10.1 *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.

11. Agent/expert/participant/patient Infallibility

11.1 *Patient Confidence*: Comment that the policy or test will or may reduce patient confidence in the agent or expert

11.2 *Coddled Experts:* Comment that experts will or may feel coddled by the policy or treatment because they already know or do what the policy or test treatment aims for.

11.3 *Agents Already Know:* Comment that agents already know or do what the policy or test treatment aims for.

11.4 *Agents Should Already Know:* Comment that agents *should* already know or do what the policy or test treatment aims for.

11.5 *Experts Already Know:* Comment that experts already know or do what the policy or test treatment aims for.

11.6 *Experts Should Already Know:* Comment that experts *should* already know or do what the policy or test treatment aims for.

11.7 *Participant Knows Best:* Comment in which the participant suggests an alternative, presumably superior, policy, or suggests that all patients receive both treatments

(which should also be coded as Inequality). Does *not* apply to comments suggesting a superior A/B test (which is coded as Bad Design) or to comments that one policy is superior to the other (which is No Equipoise).

11.8 *Patient Already (Should) Know:* Comment that the patients for whom the policy or test treatment is intended do not require it, because they already do or should know what to do (e.g., employees already know to save for retirement, or should, and so don't need to be prompted to do so).

12. Other

12.1 *Necessary Reveal:* Indicates that coder was forced to look at the condition and/or appropriateness rating associated with a comment in order to resolve ambiguity about the comment's meaning.

12.2 *Ambiguity:* Indicates that coder was unable to determine part or all of a comment's meaning (with or without revealing the condition or rating), and hence did not otherwise code that comment or part of it. Does not include intentional nonsense answers (e.g., meow meow meow).

12.3 *Coder Disagreement:* Indicates that the coders initially disagreed about a code for a comment and had to resolve this disagreement by discussion.

Data Exclusions

For Studies 1 and 2a, we analyzed our data both before and after excluding participants on the basis of free responses, following the procedures outlined in our preregistered codebook. Our exclusion criteria were preregistered as follows:

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, other than of participants who took part in more than one of our studies, were made.

As detailed in our comparison of pre- and post- exclusion results effect sizes remained roughly the same with and without exclusions, and exclusions did not affect the directionality or statistical significance of any effect. **We therefore ceased excluding participants for any reason other than repeat participation beginning with Study 3.**

To provide a methodologically conservative estimate, and to maintain consistency across studies, we report analyses of all participants who provided complete data for all results in the main text and who had not participated in a prior study described in this supplement. All materials used in all experiments, replications, and pilots are available in full in the materials section.

Results

Reporting Conventions

We proceed by reporting the following analyses within each condition of each vignette scenario, for the studies reported in the main manuscript.

Table S2. Study Names and Location in Document

Study		Scenario	Page #
Study 1	1	Safety Checklist	51
Study 2	2a	Safety Checklist (Direct Replication)	54
	2b	Safety Checklist 2	58
	2c	Safety Checklist (Mobile)	61
Study 3	3a	Genetic Testing	65
	3b	Autonomous Vehicles	68
	3c	Retirement Plans	71
	3d	Health Worker Recruitment	72
	3e	Poverty Alleviation	73
	3f	Teacher Wellbeing	75
	3g	Basic Income	77
Study 4	4	Drug Effectiveness	78
Study 5	5a	Drug Effectiveness Walk-in	80
	5b	Drug Effectiveness Walk-in (Mobile)	82
Study 6	6a	Safety Checklist (Healthcare)	85
	6b	Drug Effectiveness Walk-in (Healthcare)	87

Descriptive statistics. Within each condition, we report the number of participants (*n*) (excluding those who participated in multiple studies), the percentage of participants who objected to the decision (by providing a rating of 1 – Very inappropriate – or 2 – Somewhat inappropriate), the average appropriateness rating (on a scale of 1 – Very inappropriate – to 5 – Very appropriate), the standard deviation of this rating (*SD*), the standard error of the mean (SEM), and the 95% confidence interval range about the mean (95% CI +/–). We supplement these descriptive statistics with a histogram displaying the distributions of scale point responses within each condition.

Critical inferential tests on appropriateness ratings. For each sample within each vignette scenario, we report a series of four or five critical pairwise comparisons using two-tailed, independent-groups *t*-tests ($\alpha = .05$). We report exact *p*-values for all cases where $p > .001$. Although we display percentages in the main text, we conduct our inferential analyses on continuous appropriateness ratings. This is because percentages computed by dichotomizing continuous results, although descriptively useful, obscure statistical information when the underlying continuous data are available *(5)*.

The first comparison (**Policy Equivalence**) tests whether one of the two policy conditions (either A or B) is rated as more appropriate than the other.

The second comparison (**Experiment Equivalence**) establishes whether one of the two experimental conditions (either A/B short or A/B learn) is rated as more appropriate than the other. Starting in Study 3, we no longer report this comparison because A/B short was not run.

The three remaining comparisons provide evidence for the effect by comparing the average appropriateness rating in the A/B condition against the average appropriateness rating of the corresponding A condition (**A:A/B**), or B condition (**B:A/B**). In cases where we ran both A/B short and A/B learn conditions, we pooled these conditions together because they did not differ in appropriateness. Finally, we compared the average appropriateness rating of either the A or B condition (pooled together) against the average appropriateness rating of the corresponding A/B test. When no A/B short condition was run, this was necessarily the A/B learn condition. Whenever both A/B conditions were run, data from the A/B short and A/B learn conditions are pooled together (reported here as "A/B"). This comparison (**A/B Effect**) is the focal test of the effect and serves as the basis for the effect size estimates reported in the main text. Table S53 displays the results of these three comparisons across all studies.

Additional preregistered analyses. We report the analyses we preregistered in italicized text and note the results by pointing to one of the critical inferential tests. If a preregistered analysis differs from this approach, we report this additional analysis and result in this section.

Note that in some cases, we preregistered secondary analyses to be conducted within studies that are better suited (and better powered) to be conducted over all studies where the data are available. These analyses include testing for any effect of scientific literacy, the God, Intuition, and Science scale, and demographic characteristics (e.g., sex, race/ethnicity, age, education, and income). We report the results of these exploratory analyses in the section entitled "Additional Exploratory and Preregistered Analyses."

Summary of main result. We provide a brief summary of the results and note cases where they differed from our predictions or did not support our hypotheses.

Free response analyses. For studies where we coded participants' free response text (Studies 1, 2a, 4, and 5a), we report the results of the qualitative free response coding using conventional content analysis, including all preregistered and exploratory results.

Additional robustness checks. For studies where we coded participants for exclusion based on preregistered criteria, we report the critical results after exclusion to ensure that the findings are robust.

***Study 1**: Safety Checklist*

Descriptive statistics

Table S3. Descriptive Statistics (Study 1)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	75	14.67	3.93	1.14	0.13	0.26
Poster (B)	118	7.63	4.35	1.01	0.09	0.18
A/B short	118	43.22	2.84	1.32	0.12	0.24
A/B learn	102	50.98	2.63	1.30	0.13	0.26

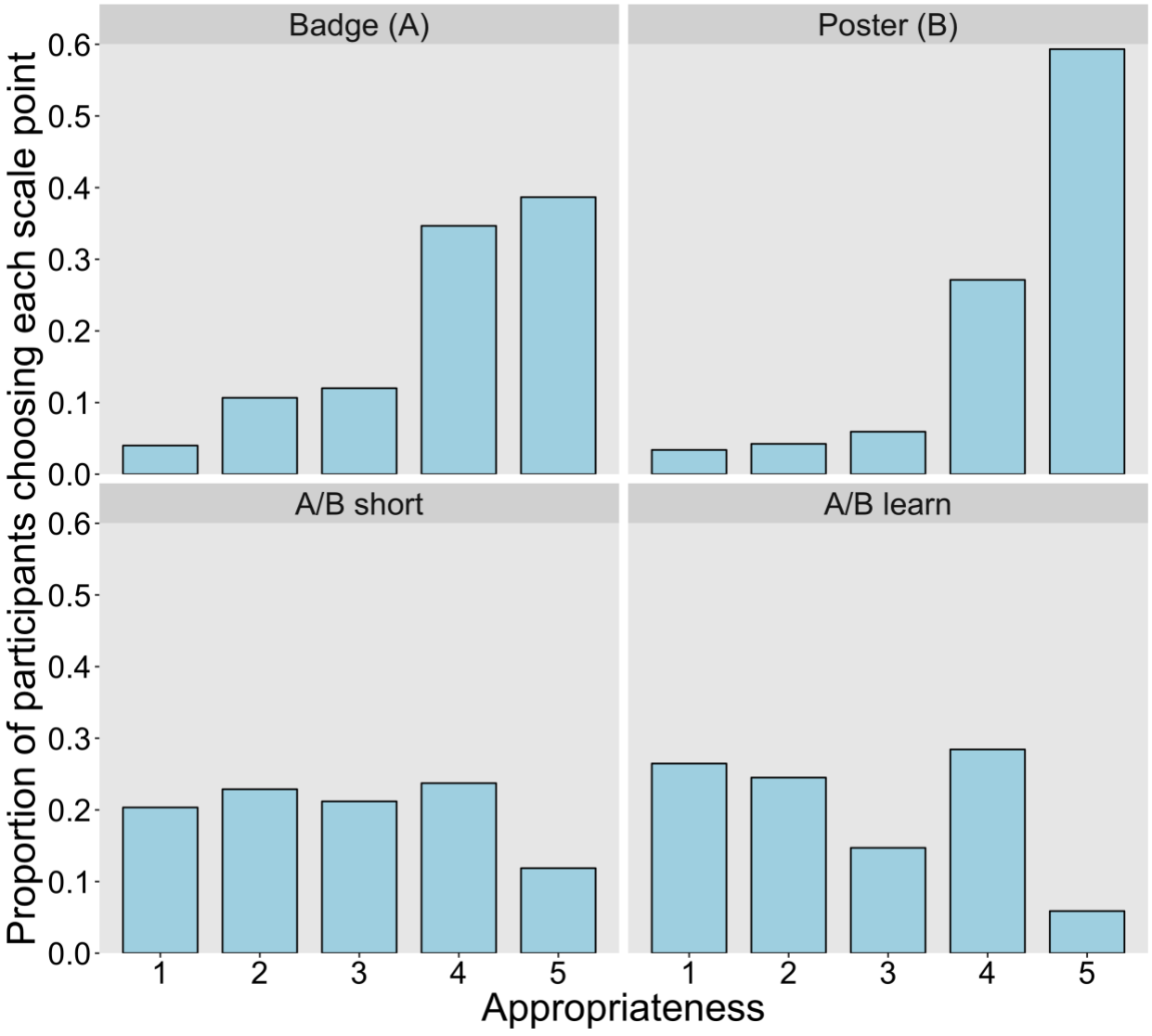


Fig. S2. Distributions of appropriateness responses within each condition (Study 1).

Critical inferential tests

Table S4. Inferential Statistics (Study 1)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.93	4.35	2.64	191	0.01	0.39
Experiment Equivalence	A/B short	A/B learn	2.84	2.63	1.19	218	0.23	0.16
A:A/B	A	A/B	3.93	2.74	7.01	293	0.001	0.94
B:A/B	B	A/B	4.35	2.74	11.58	336	0.001	1.32
A/B Effect	**A or B**	**A/B**	**4.19**	**2.74**	**12.12**	**411**	**0.001**	**1.19**

Additional preregistered analyses. Study 1 was not preregistered; there are no analyses to report.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the effect in the venous catheterization scenario. The effect size was large (d = 1.19).

Free response analyses. Study 1 was not preregistered but served as the basis for 1) developing the content code book and 2) future free-response coding analyses and hypotheses. Here, we report the results of the analyses that were preregistered as predictions for future studies.

As hypothesized, Benefit was the most commonly coded response (70.65% of participants) among the 184 participants who rated the director's decision as somewhat or very appropriate. The next most frequent codes were Learning (14.13%) followed by Equality (11.96%).

Contrary to our hypothesis, Ineffective was not the most frequently coded response among the 12 participants in the Policy conditions who rated the director's decision as somewhat or very inappropriate. The most commonly coded response among these participants was Other Risk (58.33%, or 7 participants), followed by Ineffective (50.00%, or 6 participants).

As hypothesized, Negative Research (now simply referred to as "Research") was the most frequently coded response (68.60%) among the 86 participants in the AB conditions who rated the director's decision as either somewhat or very inappropriate. The next most frequent codes were Harm (40.70%) followed Consent (30.23%).

As hypothesized, a greater proportion of participants mentioned Consent in the AB conditions (17/190) than in the A and B conditions (0/142), $\chi^2 = 11.61$, $p < .001$.

Robustness check after hand-coded exclusions. Results are displayed below for the same analyses conducted again after excluding participants for reasons we later preregistered based on their free response answer text (reversing the scale pole order; providing a nonsense

response; demonstrating a clear misunderstanding of the vignette). These exclusions (n = 76) did not meaningfully affect the direction or statistical significance of any critical comparison.

Table S5. Descriptive Statistics After Removing Hand-Coded Exclusions (Study 1)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	65	13.85	3.95	1.15	0.14	0.29
Poster (B)	81	8.64	4.35	1.06	0.12	0.23
A/B short	100	40.00	2.90	1.35	0.14	0.27
A/B learn	91	50.55	2.65	1.32	0.14	0.27

Table S6. Inferential Statistics After Removing Hand-Coded Exclusions (Study 1)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.95	4.35	2.13	144	0.03	0.36
Experiment Equivalence	A/B short	A/B learn	2.90	2.65	1.30	189	0.20	0.19
A:A/B	A	A/B	3.95	2.78	6.31	254	0.001	0.91
B:A/B	B	A/B	4.35	2.78	9.34	270	0.001	1.24
A/B Effect	**A or B**	**A/B**	**4.17**	**2.78**	**10.14**	**335**	**0.001**	**1.11**

Additionally, we conducted the same analyses after excluding the 11% of participants whose free response comments indicated that they believed that either the badge or the poster policy was superior or who broadly objected that the A/B test comparing the badge and poster conditions treated people unequally (inequality and no equipoise codes). This was to rule out the possibility that the A/B Effect is driven by participants who view the two policies as unequally effective after learning about both. After excluding these participants, the A/B Effect size remained large (d = 1.14).

***Study 2a**: Safety Checklist (Direct Replication)*

Descriptive statistics

Table S7. Descriptive Statistics (Study 2a)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	97	24.74	3.73	1.31	0.13	0.26
Poster (B)	97	11.34	4.21	1.09	0.11	0.22
A/B short	101	44.55	2.84	1.24	0.12	0.24
A/B learn	91	49.45	2.90	1.43	0.15	0.30

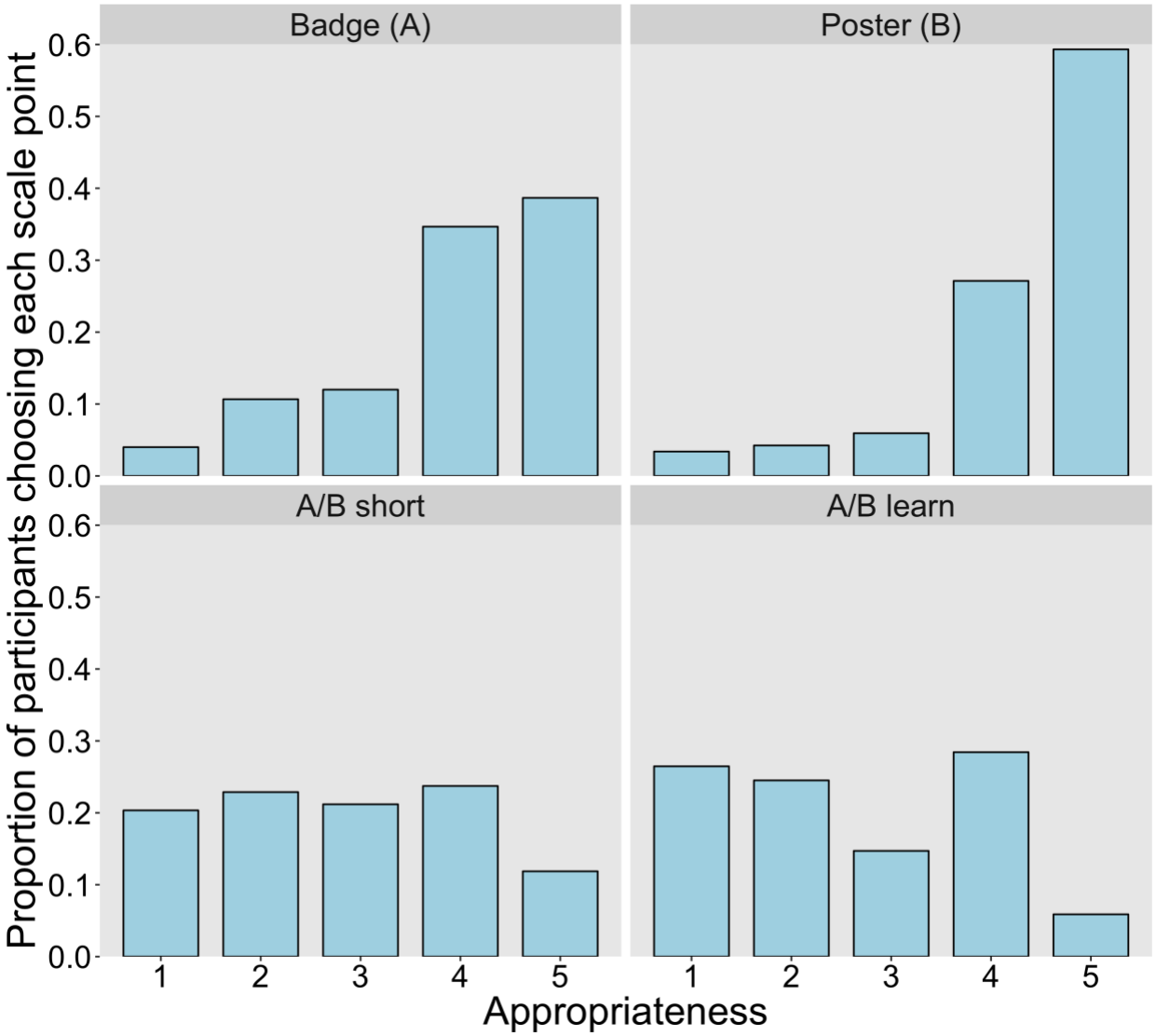


Fig. S3. Distributions of appropriateness responses within each condition (Study 2a).

Critical inferential tests

Table S8. Inferential Statistics (Study 2a)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.73	4.21	2.74	192	0.01	0.39
Experiment Equivalence	A/B short	A/B learn	2.84	2.90	0.31	190	0.76	0.04
A:A/B	A	A/B	3.73	2.87	5.23	287	0.001	0.65
B:A/B	B	A/B	4.21	2.87	8.55	287	0.001	1.07
A/B Effect	**A or B**	**A/B**	**3.97**	**2.87**	**8.45**	**384**	**0.001**	**0.86**

Additional preregistered analyses

Table S9. Additional Preregistered Analyses (Study 2a)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	3.73	2.84	4.91	196	0.001	0.70
B	A/B short	4.21	2.84	8.22	196	0.001	1.17
A	A/B learn	3.73	2.90	4.16	186	0.001	0.61
B	A/B learn	4.21	2.90	7.07	186	0.001	1.03

Italicized text copied from preregistration document.

Independent groups t-test for mean-level differences in appropriateness between A [badge] and B [poster]. This is just for descriptive purposes. **Policy equivalence** comparison reported in Table S8.

Independent groups t-test for differences in appropriateness between the two AB conditions [bp_long] and [bp_short]. This is just for descriptive purposes. **Experiment equivalence** comparison reported in Table S8.

Compare A [badge] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that A [badge] will be rated more appropriate than either AB case. See Table S9.

Compare B [poster] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that B [poster] will be rated more appropriate than either AB case. See Table S9.

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]. **A/B Effect** comparison reported in Table S8.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the effect in the direct replication of the venous catheterization scenario. The effect size was large (d = 0.86).

Free response analyses. Following from the preregistration document: As hypothesized, Benefit was the most commonly coded response (83.08% of participants) among the 195 participants who rated the director's decision as somewhat or very appropriate. The next most frequent codes were Learning (20.00%) followed by Absence of Harm (15.38%).

Partially consistent with our hypothesis, Ineffective, Experts Should Already Know, and Status Quo were the most frequently coded responses (each 28.57%, or 6 participants) among the 21 participants in the Policy conditions who rated the decision as somewhat or very inappropriate.

As hypothesized, Negative Research (now simply referred to as "Research") was the most frequently coded response (65.06%) among the 83 participants in the AB conditions who rated the director's decision as somewhat or very inappropriate. The next most frequent codes were Harm (43.37%) and Consent (30.12%).

As hypothesized, a greater proportion of participants mentioned Consent in the AB conditions (18/178) than in the A and B conditions (0/160), $\chi^2 = 15.14$, $p < .001$.

Robustness check after hand-coded exclusions. Results are displayed below for the same analyses conducted again after excluding participants for preregistered reasons based on their free response answer text (reversing the scale pole order; providing a nonsense response; demonstrating a clear misunderstanding of the vignette). These exclusions (n = 49) did not meaningfully affect the direction or statistical significance of any critical comparison.

Table S10. Descriptive Statistics After Removing Hand-Coded Exclusions (Study 2a)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	82	19.51	3.88	1.22	0.13	0.27
Poster (B)	78	6.41	4.38	0.84	0.10	0.19
A/B short	91	45.05	2.82	1.25	0.13	0.26
A/B learn	86	47.67	2.94	1.44	0.16	0.31

Table S11. Inferential Statistics After Removing Hand-Coded Exclusions (Study 2a)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.88	4.38	3.04	158	0.003	0.48
Experiment Equivalence	A/B short	A/B learn	2.82	2.94	0.58	175	0.56	0.09
A:A/B	A	A/B	3.88	2.88	5.71	257	0.001	0.76
B:A/B	B	A/B	4.38	2.88	9.11	253	0.001	1.24
A/B Effect	**A or B**	**A/B**	**4.13**	**2.88**	**9.29**	**335**	**0.001**	**1.01**

Table S12. Additional Preregistered Tests After Removing Hand-Coded Exclusions (Study 2a)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	3.88	2.82	5.59	171	0.001	0.85
B	A/B short	4.38	2.82	9.34	167	0.001	1.44
A	A/B learn	3.88	2.94	4.53	166	0.001	0.70
B	A/B learn	4.38	2.94	7.72	162	0.001	1.21

We once again conducted the same analyses after excluding the 17% of participants whose free response comments indicated that they believed that either the badge or the poster policy was superior or who broadly objected that the A/B test comparing the badge and poster conditions treated people unequally ("inequality" and "no equipoise" codes). This was to rule out the possibility that the A/B Effect is driven by participants who view the two policies as unequally effective after learning about both. After excluding these participants, the A/B Effect size remained large (d = 0.78).

***Study 2b**: Safety Checklist 2 (Alternate Replication with Wording Changes)*

Descriptive statistics

Table S13. Descriptive Statistics (Study 2b)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	90	21.11	3.82	1.28	0.13	0.27
Poster (B)	90	12.22	3.98	1.12	0.12	0.23
A/B short	85	29.41	3.48	1.31	0.14	0.28
A/B learn	78	28.21	3.51	1.26	0.14	0.28

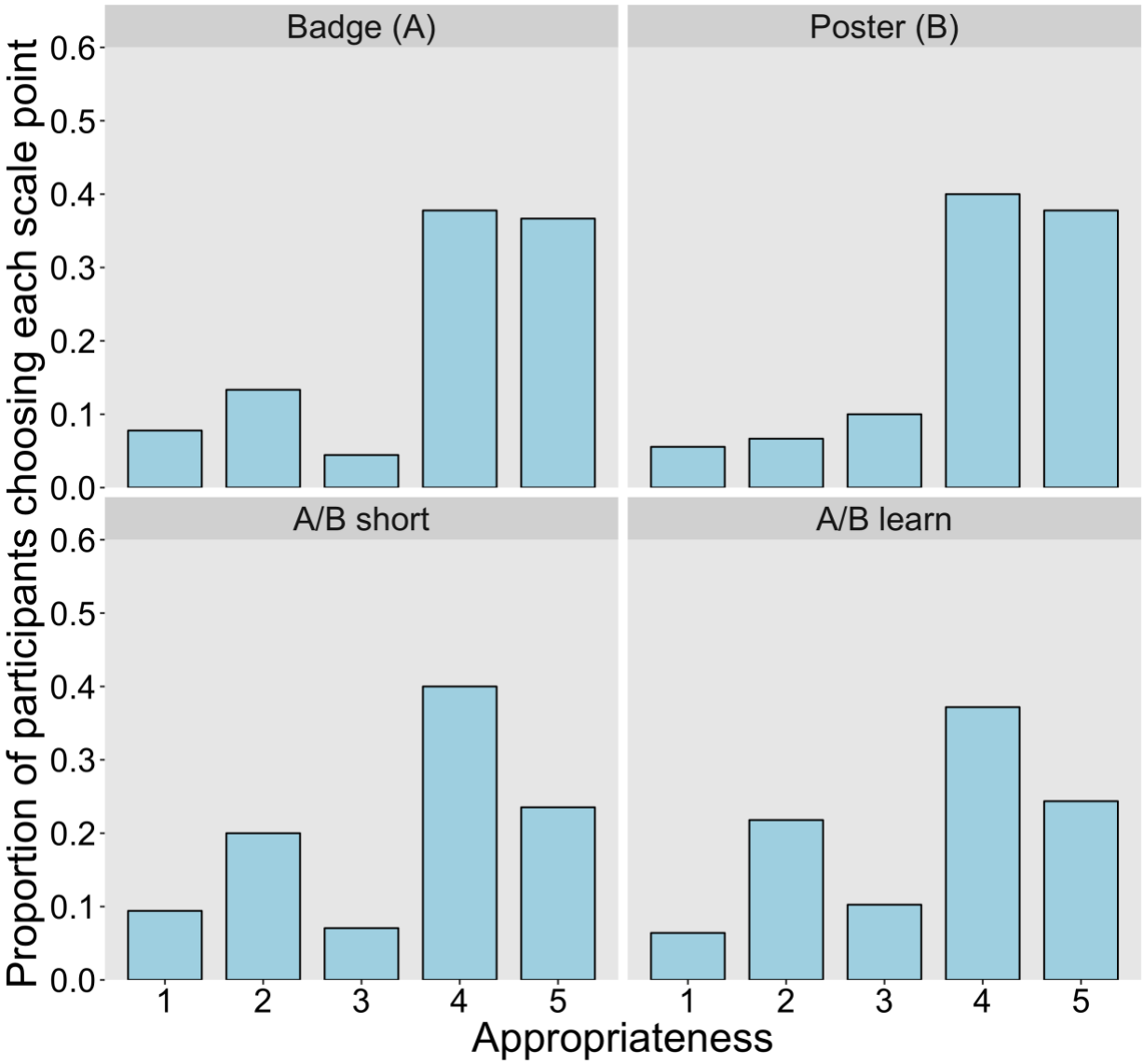


Fig. S4. Distributions of appropriateness responses within each condition (Study 2b).

Critical inferential tests

Table S14. Inferential Statistics (Study 2b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.82	3.98	0.87	178	0.39	0.13
Experiment Equivalence	A/B short	A/B learn	3.48	3.51	0.15	161	0.88	0.02
A:A/B	A	A/B	3.82	3.50	1.94	251	0.054	0.25
B:A/B	B	A/B	3.98	3.50	2.99	251	0.003	0.39
A/B Effect	**A or B**	**A/B**	**3.90**	**3.50**	**3.01**	**341**	**0.003**	**0.33**

Additional preregistered analyses

Table S15. Additional Preregistered Analyses (Study 2b)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	3.82	3.48	1.74	173	0.08	0.26
B	A/B short	3.98	3.48	2.70	173	0.01	0.41
A	A/B learn	3.82	3.51	1.58	166	0.12	0.24
B	A/B learn	3.98	3.51	2.53	166	0.01	0.39

Italicized text copied from preregistration document.

Independent groups t-test for mean-level differences in appropriateness between A [badge] and B [poster]. This is just for descriptive purposes. **Policy equivalence** comparison reported in Table S14.

Independent groups t-test for differences in appropriateness between the two AB conditions [bp_long] and [bp_short]. This is just for descriptive purposes. **Experiment equivalence** comparison reported in Table S14.

Compare A [badge] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that A [badge] will be rated more appropriate than either AB case. See Table S15.

Compare B [poster] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that B [poster] will be rated more appropriate than either AB case. See Table S15.

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]. **A/B Effect** comparison reported in Table S14.

Summary of main result. The A/B effect was robust to a series of wording changes in the venous catheterization stimuli, but the effect size decreased to $d = 0.33$. Additionally, the A:A/B comparison was only marginally significant, $p < .054$. Note here that a large number of participants were excluded ($n = 58$) because they also participated in Study 1. The results do not depend on whether these participants were excluded or not, but their exclusion does result in this study having a smaller sample size than intended.

Robustness check after hand-coded exclusions. Results are displayed below for the same analyses conducted again after excluding participants for preregistered reasons based on their free response answer text (reversing the scale pole order; providing a nonsense response; demonstrating a clear misunderstanding of the vignette). These exclusions ($n = 9$) did not meaningfully affect the direction or effect size of any critical comparison. We inferred that a comparatively small number of participants were marked for exclusion in this study compared to the previous studies because the vignette text was longer and more descriptive in this case, likely resulting in fewer participants misunderstanding the vignette.

Table S16. Descriptive Statistics After Removing Hand-Coded Exclusions (Study 2b)

Condition	N	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	89	20.22	3.85	1.25	0.13	0.26
Poster (B)	90	12.22	3.98	1.12	0.12	0.23
A/B short	79	29.11	3.54	1.29	0.15	0.29
A/B learn	76	28.95	3.47	1.25	0.14	0.29

Table S17. Inferential Statistics After Removing Hand-Coded Exclusions (Study 2b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.85	3.98	0.70	177	0.49	0.10
Experiment Equivalence	A/B short	A/B learn	3.54	3.47	0.35	153	0.73	0.06
A:A/B	A	A/B	3.85	3.51	2.06	242	0.04	0.27
B:A/B	B	A/B	3.98	3.51	2.91	243	0.004	0.39
A/B Effect	**A or B**	**A/B**	**3.92**	**3.51**	**3.03**	**332**	**0.003**	**0.33**

Table S18. Additional Preregistered Tests After Removing Hand-Coded Exclusions (Study 2b)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	3.85	3.54	1.58	166	0.12	0.24
B	A/B short	3.98	3.54	2.34	167	0.02	0.36
A	A/B learn	3.85	3.47	1.95	163	0.05	0.30
B	A/B learn	3.98	3.47	2.74	164	0.01	0.43

***Study 2c**: Safety Checklist (Direct Replication on Pollfish)*

Note. In this study, 146 participants were found to have taken the survey more than once. As per our MTurk study exclusion rules (exclude anyone who took more than one survey), these participants were excluded from all analyses.

Descriptive statistics

Table S19. Descriptive Statistics (Study 2c)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	177	20.34	3.75	1.39	0.10	0.21
Poster (B)	147	13.61	4.14	1.25	0.10	0.20
A/B short	158	44.94	2.87	1.48	0.12	0.23
A/B learn	197	40.10	2.91	1.41	0.10	0.20

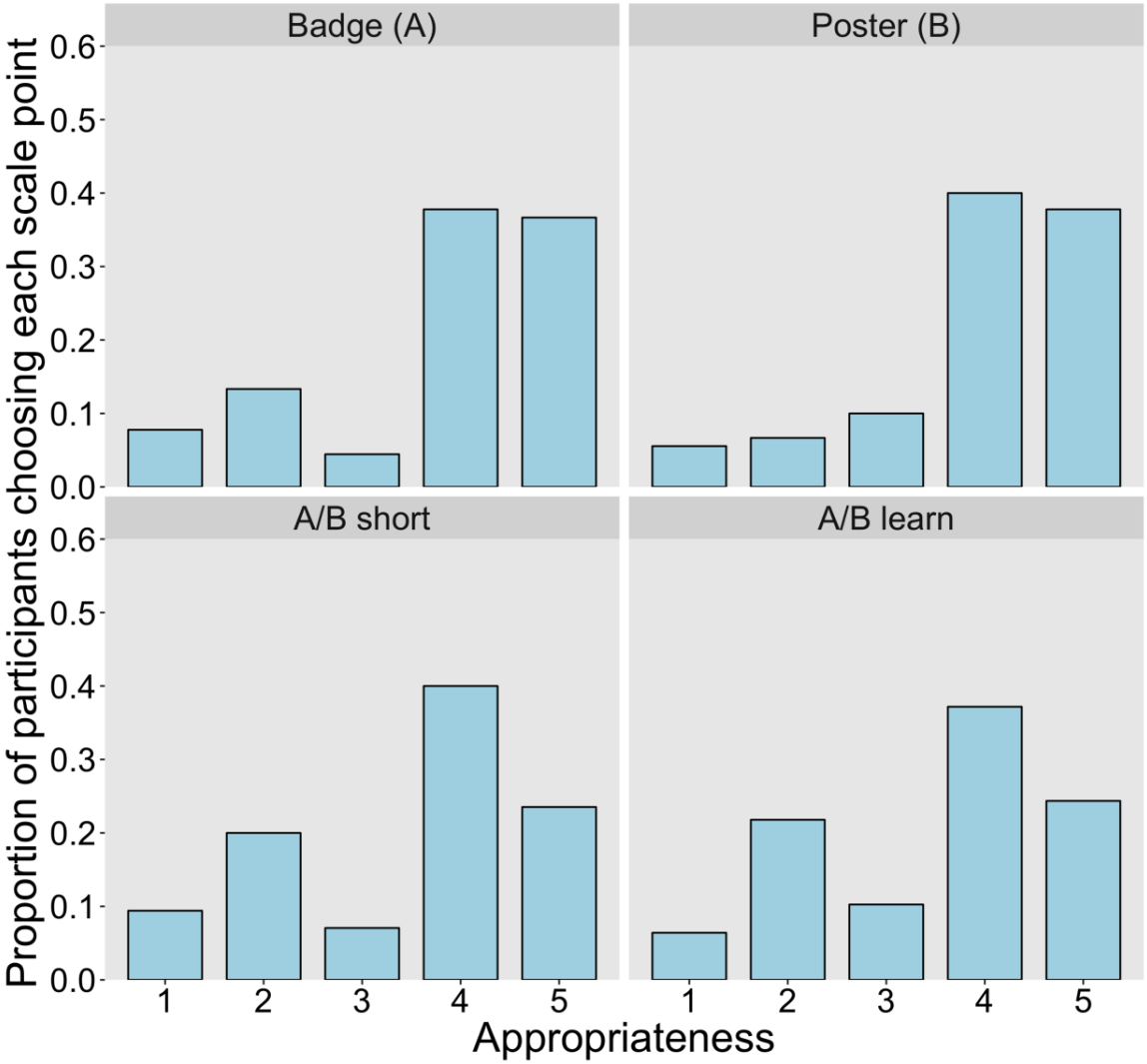


Fig. S5. Distributions of appropriateness responses within each condition (Study 2c).

Critical inferential tests

Table S20. Inferential Statistics (Study 2c)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.75	4.14	2.63	322	0.01	0.29
Experiment Equivalence	A/B short	A/B learn	2.87	2.91	0.27	353	0.79	0.03
A:A/B	A	A/B	3.75	2.89	6.53	530	0.001	0.60
B:A/B	B	A/B	4.14	2.89	9.15	500	0.001	0.90
A/B Effect	**A or B**	**A/B**	**3.92**	**2.89**	**9.64**	**677**	**0.001**	**0.74**

Additional preregistered analyses

Table S21. Additional Preregistered Analyses (Study 2c)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	3.75	2.87	5.61	333	0.001	0.61
B	A/B short	4.14	2.87	8.06	303	0.001	0.92
A	A/B learn	3.75	2.91	5.76	372	0.001	0.60
B	A/B learn	4.14	2.91	8.35	342	0.001	0.91

Italicized text copied from preregistration document.

Independent groups t-test for mean-level differences in appropriateness between A [badge] and B [poster]. This is just for descriptive purposes. **Policy equivalence** comparison reported in Table S20.

Independent groups t-test for differences in appropriateness between the two AB conditions [bp_long] and [bp_short]. This is just for descriptive purposes. **Experiment equivalence** comparison reported in Table S20.

Compare A [badge] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that A [badge] will be rated more appropriate than either AB case. See Table S21.

Compare B [poster] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups t-tests. The AB Illusion hypothesis predicts that B [poster] will be rated more appropriate than either AB case. See Table S21.

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]. **A/B Effect** comparison reported in Table S20.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B effect in replication of the venous catheterization scenario in a sample of mobile users. The effect size was medium (d = 0.74). This sample was determined to be noisy and revealed a substantial order effect (details below), which raises concerns about the quality of the sample collected using Pollfish.

Robustness check after hand-coded exclusions. Results are displayed below for the same analyses conducted again after excluding participants for preregistered reasons based on their free response answer text (reversing the scale pole order; providing a nonsense response; demonstrating a clear misunderstanding of the vignette). These exclusions (n = 104) did not meaningfully affect the direction or statistical significance of any critical comparison.

Table S22. Descriptive Statistics After Removing Hand-Coded Exclusions (Study 2c)

Condition	*N*	% Objecting	M Appropriateness	SD	SEM	95% CI +/–
Badge (A)	147	14.29	3.93	1.24	0.10	0.20
Poster (B)	111	14.41	4.12	1.24	0.12	0.23
A/B short	135	42.96	2.92	1.48	0.13	0.25
A/B learn	182	38.46	2.95	1.42	0.11	0.21

Table S23. Inferential Statistics After Removing Hand-Coded Exclusions (Study 2c)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.93	4.12	1.19	256	0.24	0.15
Experiment Equivalence	A/B short	A/B learn	2.92	2.95	0.20	315	0.85	0.02
A:A/B	A	A/B	3.93	2.94	7.21	462	0.001	0.72
B:A/B	B	A/B	4.12	2.94	7.67	426	0.001	0.85
A/B Effect	A or B	A/B	4.01	2.94	9.44	573	0.001	0.79

Order effect analysis. Because we suspected that many mobile participants may simply choose the first response available to them when presented with the survey, we manipulated the display order of the available responses to the primary dependent measure. In the *ascending* condition, participants were presented with a vertical scale where Very Inappropriate was the first option and Very Appropriate was the last option. In the *descending* condition, this order was reversed. Participants were randomly assigned to one of these two conditions. Collapsing over all treatment conditions (A, B, A/B short, and A/B learn), participants in the *descending* condition (n = 318) provided greater appropriateness ratings (M = 3.72, SD = 1.43) than participants in the *ascending* condition (n = 361) (M = 3.10, SD = 1.48), $t(677) = 5.35$, $p < .001$, d = 0.41, suggesting the presence of a substantial effect of scale order.

To determine whether the observed order effect meaningfully interacted with the treatment effect (i.e., the A/B Effect comparison), we entered each as an independent factor in a 2 (treatment condition: Policy or A/B Test) × 2 (order condition: Ascending or Descending) ANOVA model (see Fig. S6). A main effect of treatment condition revealed that participants in the Policy condition provided greater appropriateness ratings than participants in the A/B Test

condition, $F(1,675) = 96.25$, $p < .001$. A main effect of order condition revealed that participants in the descending condition provided greater appropriateness ratings than participants in the ascending condition, $F(1,675) = 25.98$, $p < .001$. There was no interaction effect between treatment condition and order condition, $F(1,675) = .01$, $p = .92$, suggesting that that result of the critical comparison for the A/B effect did not depend on which scale order was presented to participants.

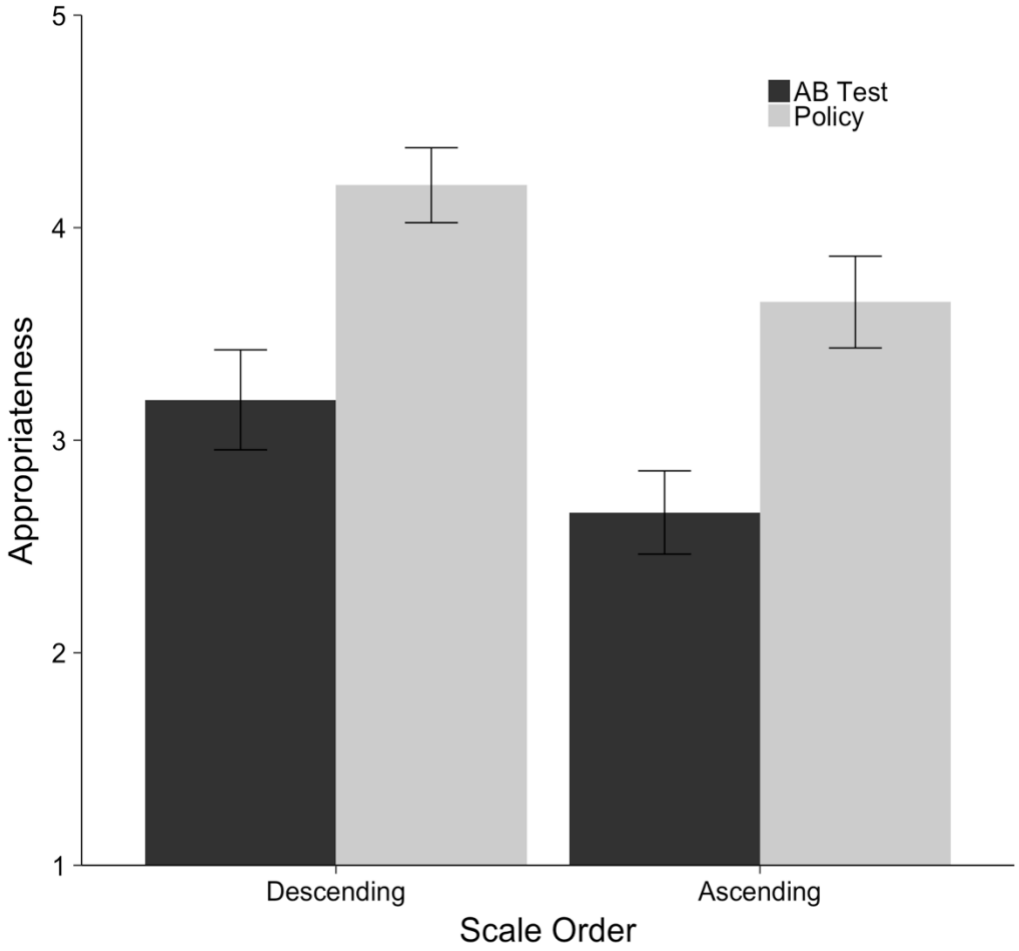


Fig. S6. Mean appropriateness ratings grouped by treatment condition and scale order condition.

***Study 3a**: Genetic Testing*

Descriptive statistics

Table S24. Descriptive Statistics (Study 3a)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Actionable (A)	97	15.46	4.13	1.17	0.12	0.24
All (B)	102	8.82	4.30	1.00	0.10	0.20
A/B short	91	19.78	3.58	1.17	0.12	0.24
A/B learn	88	22.73	3.63	1.34	0.14	0.28

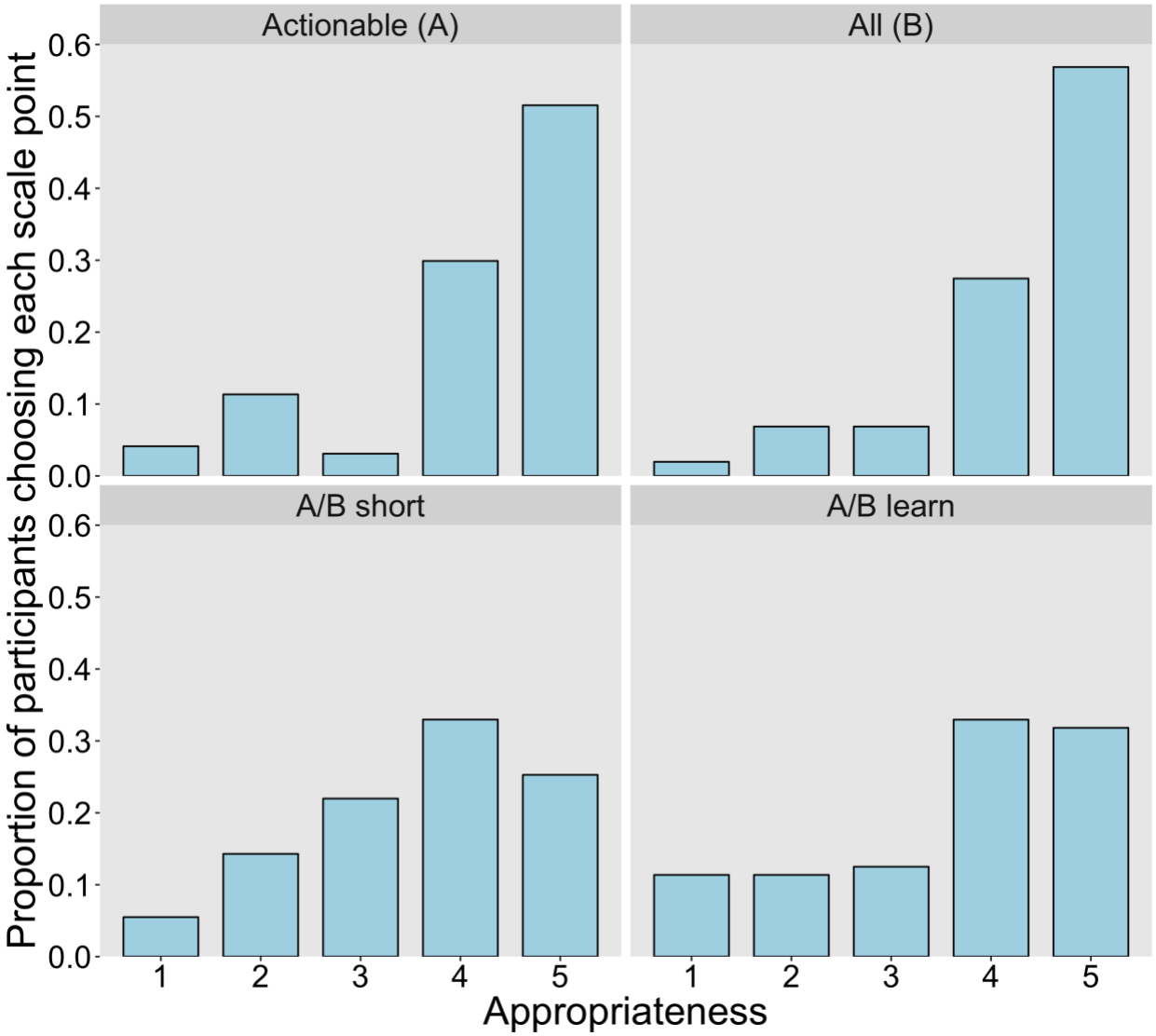


Fig. S7. Distributions of appropriateness responses within each condition (Study 3a).

Critical inferential tests

Table S25. Inferential Statistics (Study 3a)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.13	4.30	1.10	197	0.27	0.16
Experiment Equivalence	A/B short	A/B learn	3.58	3.63	0.23	177	0.82	0.03
A:A/B	A	A/B	4.13	3.60	3.43	274	0.001	0.43
B:A/B	B	A/B	4.30	3.60	4.82	279	0.001	0.60
A/B Effect	**A or B**	**A/B**	**4.22**	**3.60**	**5.12**	**376**	**0.001**	**0.53**

Additional preregistered analyses

Table S26. Additional Preregistered Analyses (Study 3a)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	4.13	3.58	3.23	186	0.002	0.47
B	A/B short	4.30	3.58	4.60	191	0.001	0.66
A	A/B learn	4.13	3.63	2.76	183	0.01	0.41
B	A/B learn	4.30	3.63	3.98	188	0.001	0.58

Italicized text copied from preregistration document.

Independent groups t-test for mean-level differences in appropriateness between A and B. This is just for descriptive purposes. **Policy equivalence** comparison reported in Table S25.

Independent groups t-test for differences in appropriateness between the two AB conditions [AB_long] and [AB_short]. This is just for descriptive purposes. **Experiment equivalence** comparison reported in Table S25.

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. See Table S26.

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. See Table S26.

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]. A/B Effect comparison reported in Table S26.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the effect in the genetic testing scenario. The effect size was medium (d = 0.53).

***Study 3b**: Autonomous Vehicles*

Descriptive statistics

Table S27. Descriptive Statistics (Study 3b)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Lever (A)	104	11.54	4.20	1.22	0.12	0.24
Automatic (B)	100	19.00	3.98	1.46	0.15	0.29
A/B short	97	25.77	3.61	1.35	0.14	0.27
A/B learn	96	30.21	3.41	1.30	0.13	0.26

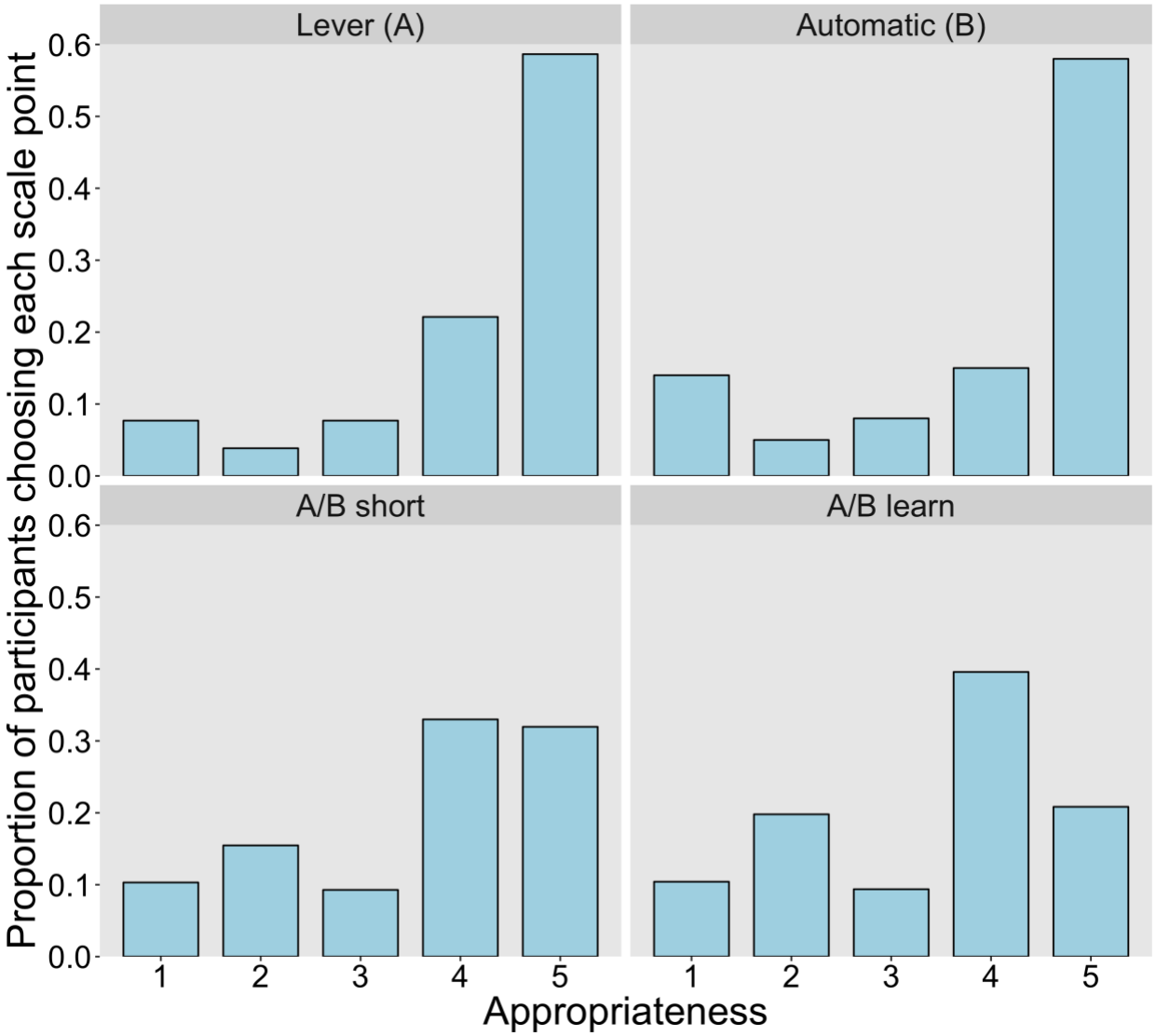


Fig. S8. Distributions of appropriateness responses within each condition (Study 3b).

Critical inferential tests

Table S28. Inferential Statistics (Study 3b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.20	3.98	1.18	202	0.24	0.17
Experiment Equivalence	A/B short	A/B learn	3.61	3.41	1.06	191	0.29	0.15
A:A/B	A	A/B	4.20	3.51	4.42	295	0.001	0.54
B:A/B	B	A/B	3.98	3.51	2.79	291	0.01	0.34
A/B Effect	**A or B**	**A/B**	**4.09**	**3.51**	**4.36**	**395**	**0.001**	**0.44**

Additional preregistered analyses

Table S29. Additional Preregistered Analyses (Study 3b)

Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
A	A/B short	4.20	3.61	3.28	199	0.001	0.46
B	A/B short	3.98	3.61	1.85	195	0.07	0.26
A	A/B learn	4.20	3.41	4.46	198	0.001	0.63
B	A/B learn	3.98	3.41	2.90	194	0.004	0.41

Italicized text copied from preregistration document.

Independent groups t-test for mean-level differences in appropriateness between A and B. This is just for descriptive purposes. **Policy equivalence** comparison reported in Table S28.

Independent groups t-test for differences in appropriateness between the two AB conditions [AB_long] and [AB_short]. This is just for descriptive purposes. **Experiment equivalence** comparison reported in Table S28.

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. See Table S29.

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. See Table S29.

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]. **A/B Effect** comparison reported in Table S28.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the effect in the autonomous vehicles scenario. The effect size was medium (d = 0.44).

***Study 3c**: Retirement Plans*

Descriptive statistics

Table S30. Descriptive Statistics (Study 3c)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
More Funds (A)	98	19.39	3.76	1.32	0.13	0.26
Popular (B)	103	19.42	3.90	1.35	0.13	0.26
A/B learn	95	36.84	3.27	1.32	0.14	0.27

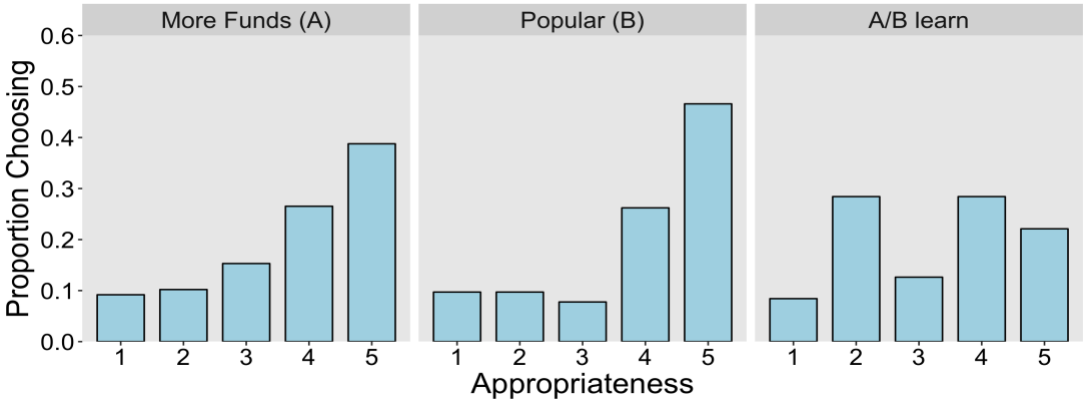


Fig. S9. Distributions of appropriateness responses within each condition (Study 3c).

Critical inferential tests

Table S31. Inferential Statistics (Study 3c)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.76	3.90	0.79	199	0.43	0.11
A:A/B	A	A/B learn	3.76	3.27	2.54	191	0.01	0.37
B:A/B	B	A/B learn	3.90	3.27	3.32	196	0.001	0.47
A/B Effect	**A or B**	**A/B learn**	**3.83**	**3.27**	**3.37**	**294**	**0.001**	**0.42**

Additional preregistered analyses. All preregistered analyses are reported in Tables S30 and S31.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the effect in the retirement plans scenario. The effect size was small to medium (d = 0.42).

***Study 3d**: Health Worker Recruitment*

Descriptive statistics

Table S32. Descriptive Statistics (Study 3d)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Social (A)	96	9.38	4.14	0.98	0.10	0.20
Career (B)	101	9.90	4.10	1.03	0.10	0.20
A/B learn	96	16.67	3.70	1.19	0.12	0.24

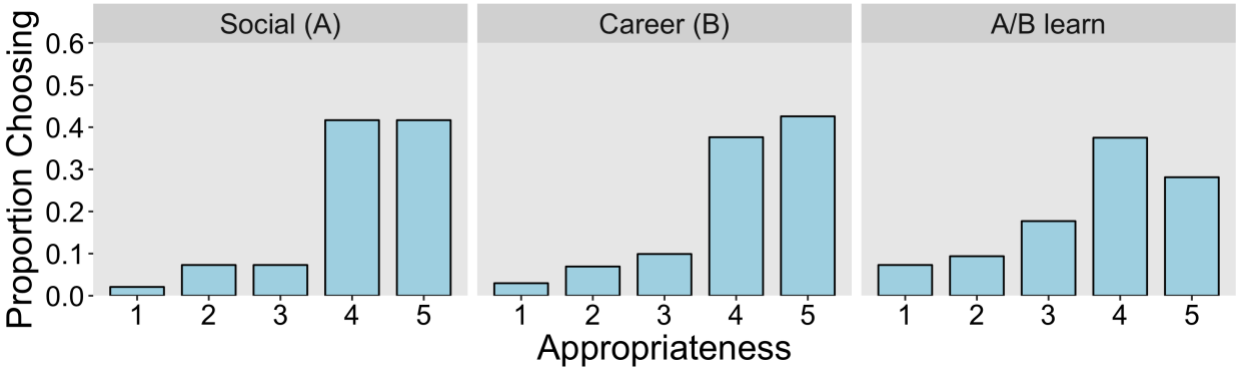


Fig. S10. Distributions of appropriateness responses within each condition (Study 3d).

Critical inferential tests

Table S33. Inferential Statistics (Study 3d)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.14	4.10	0.25	195	0.80	0.04
A:A/B	A	A/B learn	4.14	3.70	2.78	190	0.01	0.40
B:A/B	B	A/B learn	4.10	3.70	2.53	195	0.01	0.36
A/B Effect	**A or B**	**A/B learn**	**4.12**	**3.70**	**3.15**	**291**	**0.002**	**0.39**

Additional preregistered analyses. All preregistered analyses are reported in Tables S32 and S33.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B Effect in the health worker scenario. The effect size was small to medium (d = 0.39).

***Study 3e**: Poverty Alleviation*

Descriptive statistics

Table S34. Descriptive Statistics (Study 3e)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Roof (A)	96	23.96	3.57	1.25	0.13	0.25
Training (B)	103	10.68	4.06	1.15	0.11	0.23
A/B learn	103	34.95	3.31	1.24	0.12	0.24

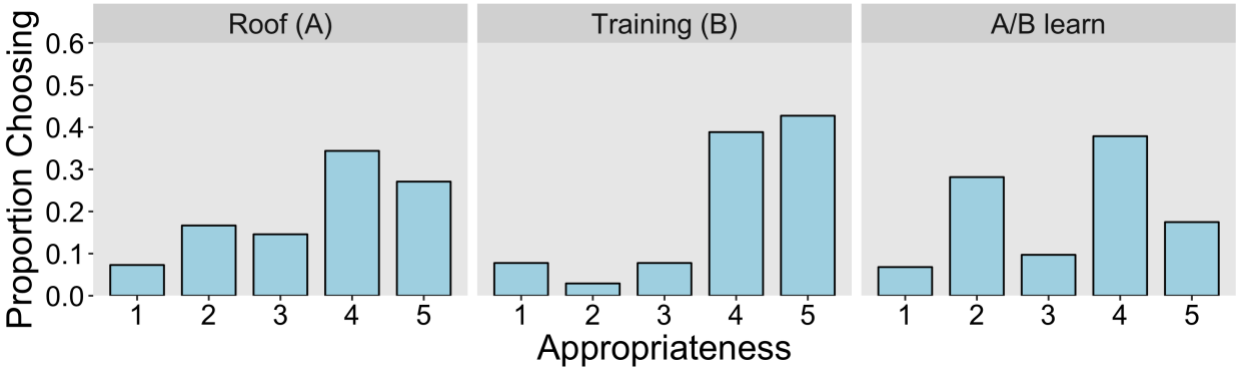


Fig. S11. Distributions of appropriateness responses within each condition (Study 3e).

Critical inferential tests

Table S35. Inferential Statistics (Study 3e)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.57	4.06	2.84	197	0.005	0.40
A:A/B	A	A/B learn	3.57	3.31	1.48	197	0.14	0.21
B:A/B	B	A/B learn	4.06	3.31	4.47	204	0.001	0.62
A/B Effect	**A or B**	**A/B learn**	**3.82**	**3.31**	**3.44**	**300**	**0.001**	**0.42**

Additional preregistered analyses. All preregistered analyses are reported in Tables S34 and S35.

Summary of main result. Two of the three critical comparisons (B:A/B; A/B Effect) provided evidence for the A/B Effect in the poverty alleviation scenario. The effect size was small to medium (d = 0.39).

The third critical comparison (A:A/B) did not yield a significant difference, though the pattern of results was in the predicted direction. Participants did not rate the A condition

(endowing a household with a new roof) as significantly less appropriate than the A/B test of this policy decision and another (providing one month of training in a trade). Additionally, the policy stimuli (the A and B conditions) in this study were not perceived as similar in appropriateness; participants viewed the B condition (training) as more appropriate than the A condition (new roof).

***Study 3f**: Teacher Wellbeing*

Descriptive statistics

Table S36. Descriptive Statistics (Study 3f)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Bonus (A)	99	12.12	4.04	1.10	0.11	0.22
Vacation (B)	97	27.84	3.64	1.36	0.14	0.27
A/B learn	104	31.73	3.34	1.25	0.12	0.24

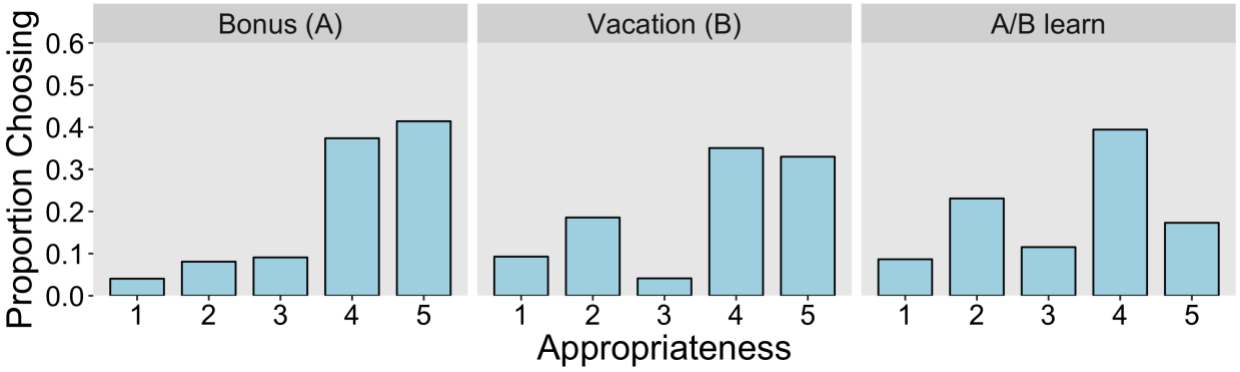


Fig. S12. Distributions of appropriateness responses within each condition (Study 3f).

Critical inferential tests

Table S37. Inferential Statistics (Study 3f)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.04	3.64	2.28	194	0.02	0.33
A:A/B	A	A/B learn	4.04	3.34	4.25	201	0.001	0.60
B:A/B	B	A/B learn	3.64	3.34	1.65	199	0.10	0.23
A/B Effect	**A or B**	**A/B learn**	**3.84**	**3.34**	**3.34**	**298**	**0.001**	**0.41**

Additional preregistered analyses. All preregistered analyses are reported in Tables S36 and S37.

Summary of main result. Two of the three critical comparisons (A:A/B; A/B Effect) provided evidence for the A/B Effect in the education scenario. The effect size was medium (d = 0.41). The third critical comparison (B:A/B) did not yield a significant difference, though the pattern of results was in the predicted direction. Participants did not rate the B condition (a policy rewarding teachers with additional vacation time) as significantly less appropriate than the

A/B test of this policy decision and another (rewarding teachers with a yearly bonus). Additionally, the policy stimuli (the A and B conditions) in this study were not perceived as similar in appropriateness; participants viewed the A condition (bonuses) as more appropriate than the B condition (vacation).

***Study 3g**: Basic Income*

Descriptive statistics

Table S38. Descriptive Statistics (Study 3g)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
$1000 (A)	102	20.59	3.73	1.15	0.11	0.23
$500 (B)	106	18.87	3.75	1.14	0.11	0.22
A/B learn	96	21.88	3.64	1.21	0.12	0.24

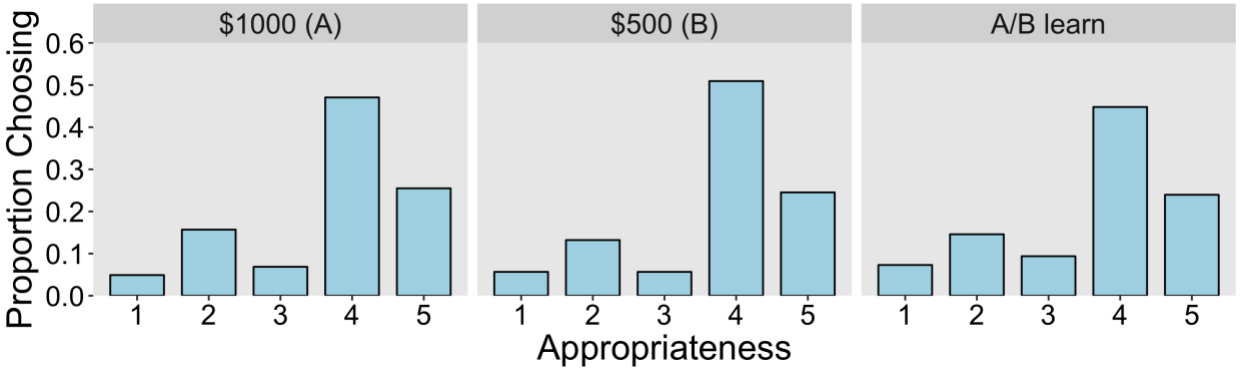


Fig. S13. Distributions of appropriateness responses within each condition (Study 3g).

Critical inferential tests

Table S39. Inferential Statistics (Study 3g)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.73	3.75	0.18	206	0.85	0.03
A:A/B	A	A/B learn	3.73	3.64	0.54	196	0.59	0.08
B:A/B	B	A/B learn	3.75	3.64	0.72	200	0.47	0.10
A/B Effect	**A or B**	**A/B learn**	**3.74**	**3.64**	**0.73**	**302**	**0.47**	**0.09**

Additional preregistered analyses. All preregistered analyses are reported in Tables S38 and S39.

Summary of main result. In the basic income scenario, no critical comparison was significant. This was the only experiment we ran that did not provide evidence for the A/B Effect. Here, participants did not differ significantly in appropriateness ratings when reading about the A condition (providing $1000 per month to unemployed individuals for one year until they find a job), B condition (providing $500 per month to unemployed individuals for one year even if they find a job) or A/B condition.

***Study 4**: Drug Effectiveness*

Descriptive statistics

Table S40. Descriptive Statistics (Study 4)

Condition	*N*	% Objecting	*M* Appropriateness	SD	SEM	95% CI +/–
Drug A	98	4.08	4.53	0.81	0.08	0.16
Drug B	102	7.84	4.33	1.05	0.10	0.21
A/B learn	102	33.33	3.33	1.46	0.14	0.29

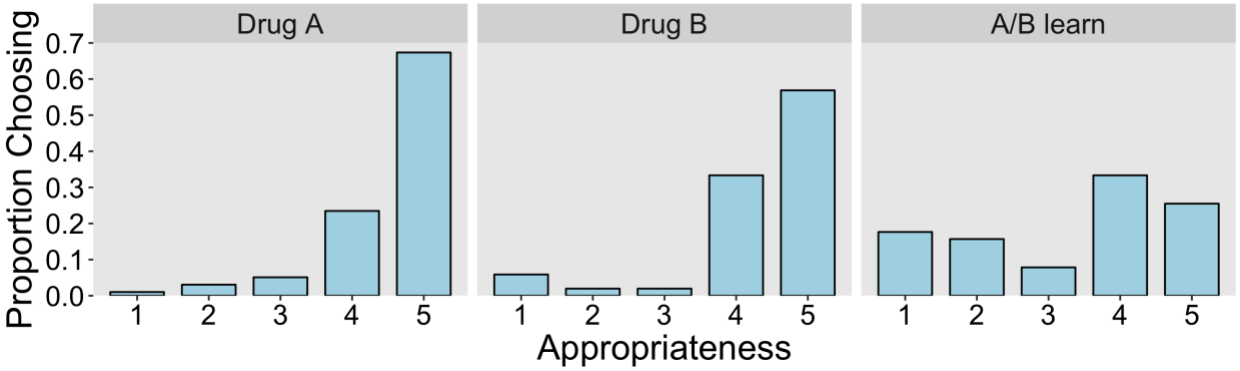


Fig. S14. Distributions of appropriateness responses within each condition (Study 4).

Critical inferential tests

Table S41. Inferential Statistics (Study 4)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.53	4.33	1.48	198	0.14	0.21
A:A/B	A	A/B learn	4.53	3.33	7.13	198	0.001	1.01
B:A/B	B	A/B learn	4.33	3.33	5.63	202	0.001	0.79
A/B Effect	**A or B**	**A/B learn**	**4.43**	**3.33**	**7.89**	**300**	**0.001**	**0.96**

Additional preregistered analyses. All preregistered analyses are reported in Tables S40 and S41.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B Effect in the drug effectiveness scenario. The effect size was large (d = 0.96).

Free response analyses. As hypothesized, Benefit was the most commonly coded response (72.41% of participants) among the 232 participants who rated the doctor's decision as

somewhat or very appropriate. The next most frequent codes were Good Intent (33.62%) followed by Learning (12.07%).

Contrary to our hypothesis, Ineffective was not the most frequently coded response among the 11 participants in the Policy conditions who rated the doctor's decision as somewhat or very inappropriate. The most frequently coded responses among these participants were Benefit and Good Intent (each 45.54%, or 5 participants), while 0% of participants mentioned Ineffective.

As hypothesized, Negative Research (now simply referred to as "Research") was the most frequently coded response (48.48%) among the 33 participants in the AB conditions who rated the doctor's decision as somewhat or very inappropriate. The next most frequent codes were Consent (45.45%) and Infallibility (24.24%).

As hypothesized, a greater proportion of participants mentioned consent in the AB conditions (16/98) than in the A and B conditions (0/191), $\chi^2 = 29.96$, $p < .001$.

***Study 5a**: Drug Effectiveness Walk-in*

Descriptive statistics

Table S42. Descriptive Statistics (Study 5a)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Drug A	104	9.62	4.13	1.13	0.11	0.22
Drug B	100	9.00	4.19	1.10	0.11	0.22
A/B learn	99	33.33	3.38	1.36	0.14	0.27

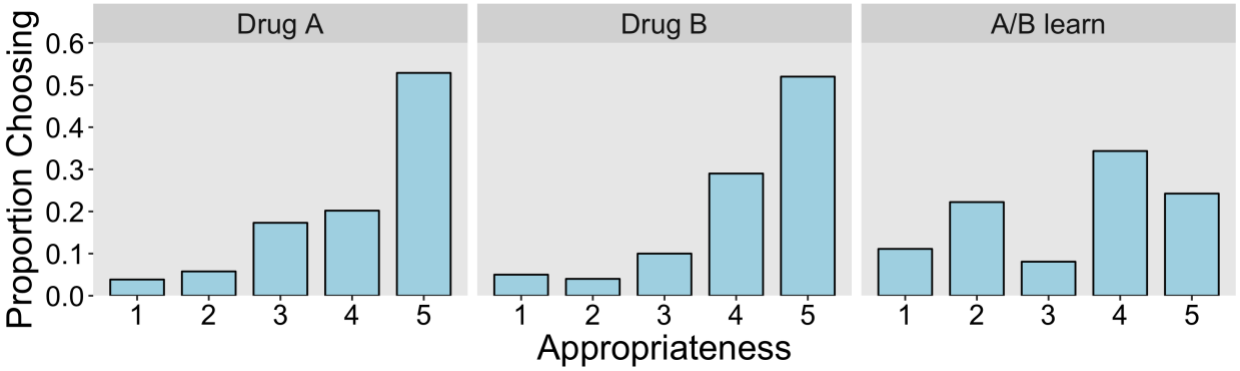


Fig. S15. Distributions of appropriateness responses within each condition (Study 5).

Critical inferential tests

Table S43. Inferential Statistics (Study 5a)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	4.13	4.19	0.42	202	0.68	0.06
A:A/B	A	A/B learn	4.13	3.38	4.23	201	0.001	0.59
B:A/B	B	A/B learn	4.19	3.38	4.60	197	0.001	0.65
A/B Effect	**A or B**	**A/B learn**	**4.16**	**3.38**	**5.27**	**301**	**0.001**	**0.64**

Additional preregistered analyses. All preregistered analyses are reported in Tables S42 and S43.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B Effect in the drug effectiveness (walk-in) scenario. The effect size was medium to large (d = 0.64).

Free response analyses. As hypothesized, Benefit was the most commonly coded response (51.16% of participants) among the 215 participants who rated the doctor's decision as

somewhat or very appropriate. The next most frequent codes were Equality (28.37%) followed by Agents Already Know (17.21%)

Contrary to our hypotheses, Ineffective was not the most frequently coded response among the 18 participants in the Policy conditions who rated the doctor's decision as somewhat or very inappropriate. The most frequently coded response among these participants was Participant Knows Best (55.56%, or 10 participants), while 0% mentioned Ineffective.

As hypothesized, Negative Research (now simply referred to as "Research") was the most frequently coded response (54.84%) among the 31 participants in the A/B condition who rated the doctor's decision as somewhat or very inappropriate. The next most frequent codes were Infallibility (48.39%) and Consent (35.48%).

As hypothesized, a greater proportion of participants mentioned Consent in the A/B condition (13/95) than in the A and B conditions (2/204), $\chi^2 = 19.37$, $p < .001$.

Summary of "Illusion of Knowledge by Proxy" codes across coded studies. As described in the main text in reference to the proxy form of the illusion of knowledge, we considered the number of participants who received this code within each condition across the four studies where participants' free response explanations were coded (see Table S44). Participants who used some form of proxy illusion of knowledge explanation (receiving the Infallibility code for agents/experts only) were more likely to approve of than object to the policy condition, and were more likely to object to than approve of the A/B test. As shown in the bottom row of Table S44, among participants whose comments indicated endorsement of the proxy illusion of knowledge, the odds of giving a rating of "inappropriate" are 2.9 times larger in the A/B than in the Policy conditions ([21/32] / [20/88]). Similarly, the odds of giving a rating of "appropriate" are 2.2 times larger in the Policy than in the A/B conditions ([68/88] / [11/32]).

Table S44. Distribution of "Illusion of Knowledge by Proxy" Codes by Condition and Rating

Study Name	Study #	Policy A	Policy B	Policy A	Policy B	A/B Learn	A/B Short	A/B Learn	A/B Short
		Rated Inappropriate		Rated Appropriate		Rated Inappropriate		Rated Appropriate	
Checklist	1	7	2	7	2	3	6	0	2
Checklist Replication	2a	10	0	11	2	1	4	3	4
Best Drug	4	1	0	5	2	1	n/a	0	n/a
Best Drug (Walk-in)	5a	0	0	19	20	6	n/a	2	n/a
	Subtotal	18	2	42	26	11	10	5	6
	Total	20 (17%)		68 (56%)		21 (18%)		11 (9%)	

Note: 9.5% (n = 120) of participants in the four coded experiments (N = 1,258) received Illusion of Knowledge by Proxy codes. The majority (74%) of these explanations were given by participants who rated a unilateral policy implementation as appropriate or rated an A/B test of two policies as inappropriate.

***Study 5b**: Drug Effectiveness Walk-In (Replication with Mobile Users using Pollfish)*

Note. In contrast with the other study conducted using Pollfish, no participants in this study participated in more than one survey. Analyses are therefore conducted on all 720 recruited participants.

Descriptive statistics

Table S45. Descriptive Statistics (Study 5b)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Drug A	240	20.00	3.63	1.34	0.09	0.17
Drug B	240	16.25	3.70	1.24	0.08	0.16
A/B learn	240	25.00	3.47	1.33	0.09	0.17

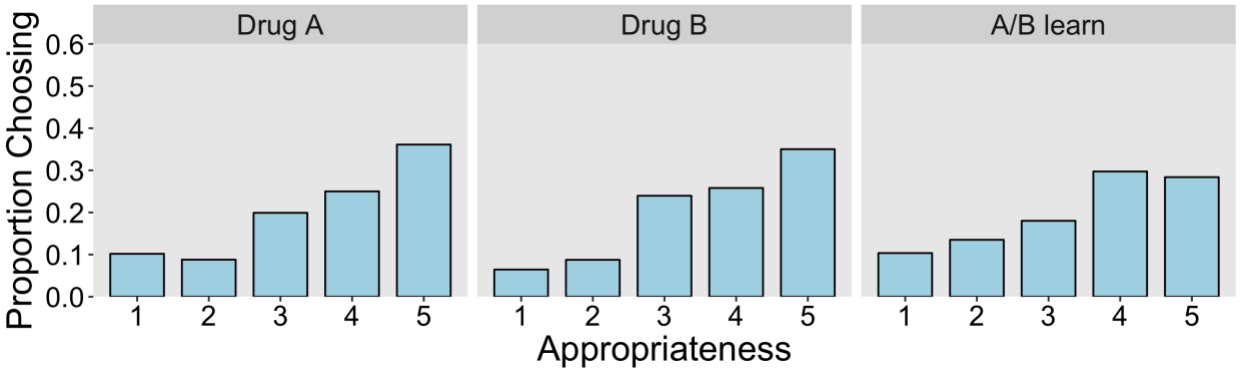


Fig. S16. Distributions of appropriateness responses within each condition (Study 5b).

Critical inferential tests

Table S46. Inferential Statistics (Study 5b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.63	3.70	0.60	478	0.548	0.05
A:A/B	A	A/B learn	3.63	3.47	1.33	478	0.183	0.12
B:A/B	B	A/B learn	3.70	3.47	1.98	478	0.048	0.18
A/B Effect	**A or B**	**A/B learn**	**3.66**	**3.47**	**1.92**	**718**	**0.055**	**0.15**

Additional preregistered analyses. All preregistered analyses are reported in Tables S45 and S46.

Summary of main result. The observed pattern of results was in the predicted pattern, but the statistical comparisons only approached or just achieved significance. The observed

effect size diminished substantially from what was observed using MTurk. This sample was determined to be quite noisy and revealed an extremely large order effect (details below).

Robustness check after hand-coded exclusions. Results are displayed below for the same analyses conducted again after excluding participants for preregistered reasons based on their free response answer text (reversing the scale pole order; providing a nonsense response; demonstrating a clear misunderstanding of the vignette). These exclusions (n = 65) slightly increased the significance level of the critical comparison (from p = .055 to p = .074), though it did not meaningfully change the pattern of results.

Table S47. Descriptive Statistics After Removing Hand-Coded Exclusions (Study 5b)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Drug A	216	18.98	3.68	1.32	0.09	0.18
Drug B	217	15.21	3.74	1.21	0.08	0.16
A/B learn	222	23.87	3.52	1.31	0.09	0.17

Table S48. Inferential Statistics After Removing Hand-Coded Exclusions (Study 5b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.68	3.74	0.51	431	0.614	0.05
A:A/B	A	A/B learn	3.68	3.52	1.26	436	0.209	0.12
B:A/B	B	A/B learn	3.74	3.52	1.82	437	0.069	0.17
A/B Effect	**A or B**	**A/B learn**	**3.71**	**3.52**	**1.79**	**653**	**0.074**	**0.15**

Order effect analysis. As in the Study 2c sample (recruited via Pollfish), we suspected that many mobile participants may simply choose the first response available to them when presented with the survey. We once again manipulated the display order of the available responses to the primary dependent measure. In the *ascending* condition, participants were presented with a vertical scale where Very Inappropriate was the first option and Very Appropriate was the last option. In the *descending* condition, this order was reversed. Participants were randomly assigned to one of these two conditions. Collapsing over all treatment conditions (A, B, and A/B learn), participants in the *descending* condition (n = 360) provided substantially greater appropriateness ratings (M = 3.98, SD = 1.11) than participants in the *ascending* condition (n = 360) (M = 3.21, SD = 1.24), $t(718) = 8.19$, $p < .001$, d = 0.61, suggesting the presence of a substantial effect of scale order. This order effect was even larger than the order effect observed in Study 2c (d = 0.41), suggesting that even more participants selected the first option available than in the previous Pollfish sample.

To determine whether the observed order effect meaningfully interacted with the treatment effect (i.e., the A/B Effect comparison), we entered each as an independent factor in a 2 (treatment condition: Policy or A/B Test) × 2 (order condition: Ascending or Descending) ANOVA model (see Fig. S17). A main effect of treatment condition revealed that participants in the Policy condition provided greater appropriateness ratings than participants in the A/B Test condition, $F(1,716) = 4.02$, p = .045. A main effect of order condition revealed that participants in the descending condition provided greater appropriateness ratings than participants in the ascending condition, $F(1,716) = 67.34$, $p < .001$. As in Study 2c, there was no interaction effect

between treatment condition and order condition, $F(1,716) = 0.55$, $p = .46$, suggesting that that result of the critical comparison for the A/B Effect did not depend on which scale order was presented to participants.

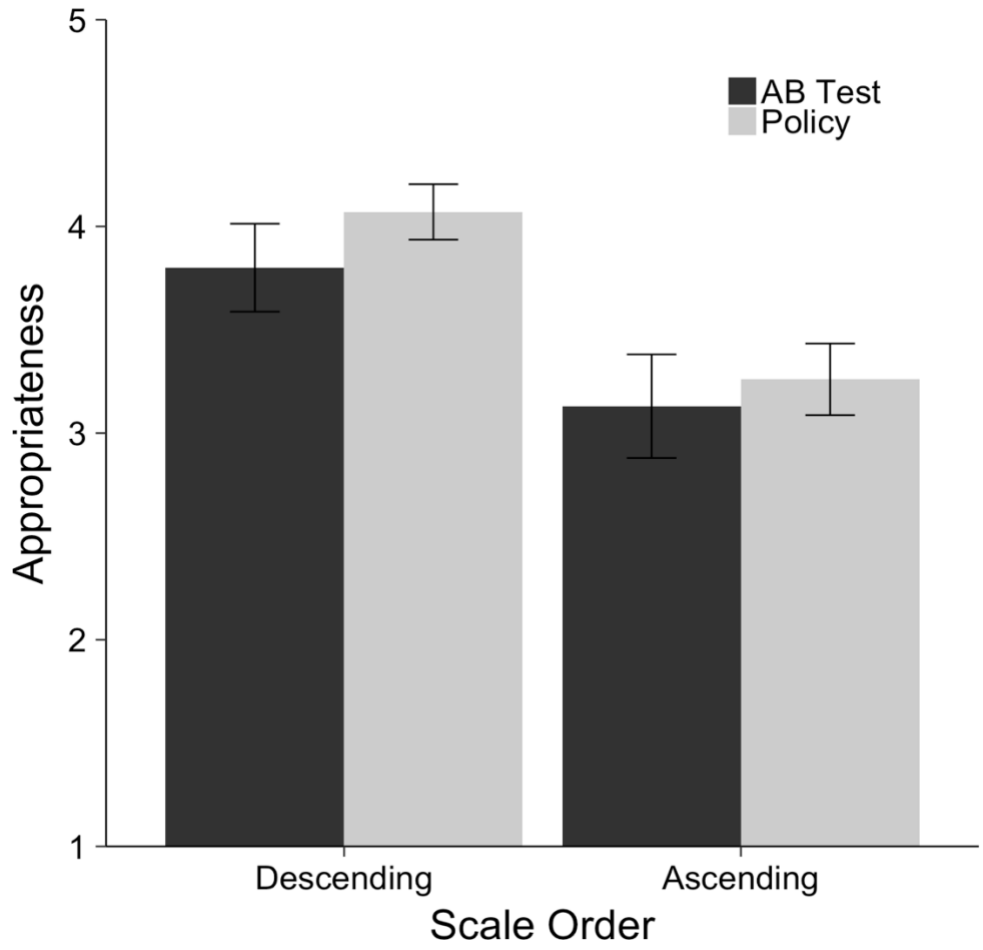


Fig. S17. Mean appropriateness ratings grouped by treatment condition and scale order condition.

***Study 6a**: Safety Checklist in Healthcare Provider Sample*

Descriptive statistics

Note. To remain consistent with previous analyses and reporting, even though each participant read and responded to two vignettes (one condition each of the Safety Checklist and Drug Effectiveness scenarios), these analyses are conducted between-subjects on the first vignette that each participant read.

Table S49. Descriptive Statistics (Study 6a)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Badge (A)	77	31.17	3.17	1.32	0.15	0.30
Poster (B)	75	17.33	4.11	1.44	0.17	0.33
A/B Learn	74	56.76	2.42	1.29	0.15	0.30

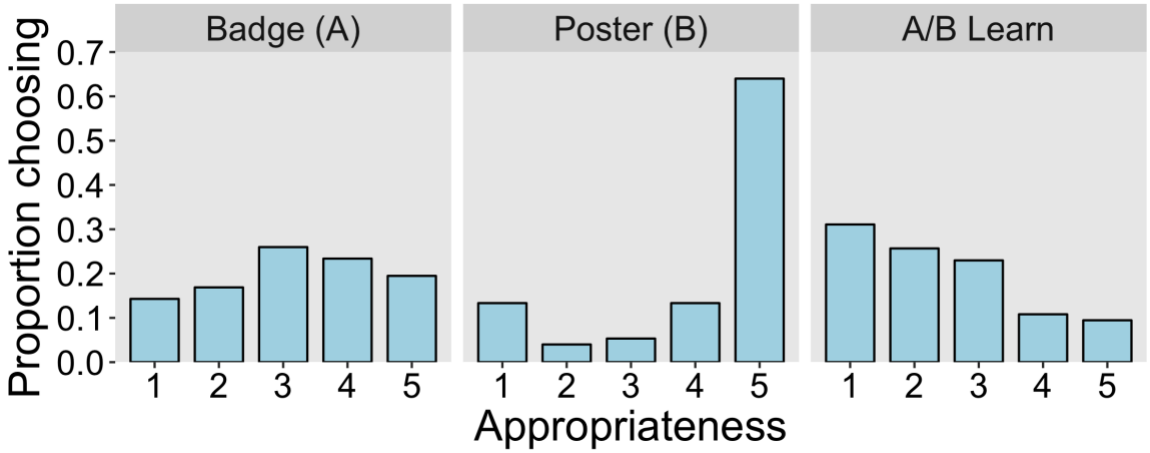


Fig. S18. Distributions of appropriateness responses within each condition (Study 6a).

Critical inferential tests

Table S50. Inferential Statistics (Study 6a)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.17	4.11	4.19	150	0.001	0.68
A:A/B	A	A/B Test	3.17	2.42	3.52	149	0.001	0.57
B:A/B	B	A/B Test	4.11	2.42	7.53	147	0.001	1.23
A/B Effect	**A or B**	**A/B Test**	**3.63**	**2.42**	**6.09**	**224**	**0.001**	**0.86**

Additional preregistered analyses. All preregistered analyses are reported in Tables S49 and S50.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B Effect in the safety checklist scenario, in this sample of expert healthcare providers. The effect size was large (d = 0.86), and comparable to the effect sizes found in studies of laypeople using the same scenario (Studies 1–2).

***Study 6b**: Drug Effectiveness Walk-In in Healthcare Provider Sample*

Descriptive statistics

Table S51. Descriptive Statistics (Study 6b)

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Drug A	76	22.37	3.46	1.40	0.16	0.32
Drug B	78	23.08	3.58	1.35	0.15	0.31
A/B Learn	77	57.14	2.31	1.40	0.16	0.32

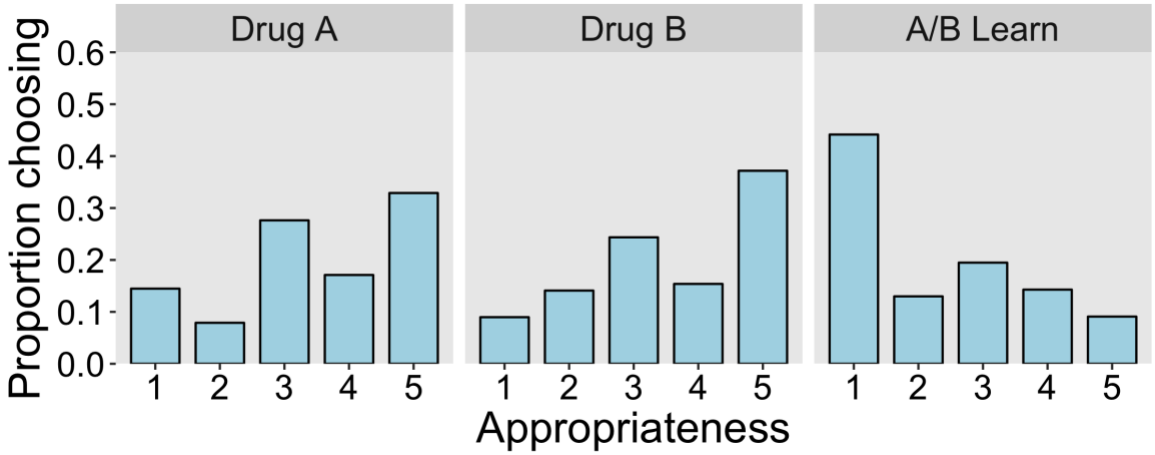


Fig. S19. Distributions of appropriateness responses within each condition (Study 6b).

Critical inferential tests

Table S52. Inferential Statistics (Study 6b)

Comparison	Group 1	Group 2	Mean 1	Mean 2	*t*	df	*p*	*d*
Policy Equivalence	A	B	3.46	3.58	0.52	152	0.601	0.08
A:A/B	A	A/B Test	3.46	2.31	5.08	151	0.001	0.82
B:A/B	B	A/B Test	3.58	2.31	5.73	153	0.001	0.92
A/B Effect	**A or B**	**A/B Test**	**3.52**	**2.31**	**6.26**	**229**	**0.001**	**0.87**

Additional preregistered analyses. All preregistered analyses are reported in Tables S51 and S52.

Summary of main result. All three critical comparisons (A:A/B; B:A/B; A/B Effect) provided evidence for the A/B Effect in the drug effectiveness walk-in scenario, in this sample of expert healthcare providers. The effect size was large (d = 0.87), and somewhat larger than the effect sizes found in our studies of laypeople using the same scenario (Study 5).

Summary Table (Studies 1–6)

Table S53. Summary of Critical Inferential Tests and Effect Sizes for all Studies.

Study		Scenario	A:A/B	B:A/B	A/B Effect	Effect Size (*d*)
Study 1	1	Safety Checklist	***	***	***	1.19
Study 2	2a	Safety Checklist Rep. (Exact)	***	***	***	0.86
	2b	Safety Checklist Rep. (Varied)	*n.s.*	**	**	0.33
	2c	Safety Checklist Rep. (Mobile)	***	***	***	0.74
Study 3	3a	Genetic Testing	***	**	***	0.53
	3b	Autonomous Vehicles	***	**	***	0.44
	3c	Retirement Plans	*	**	***	0.42
	3d	Health Worker Recruitment	**	*	**	0.39
	3e	Poverty Alleviation	*n.s.*	***	***	0.42
	3f	Teacher Wellbeing	***	*n.s.*	***	0.41
	3g	Basic Income	*n.s.*	*n.s.*	*n.s.*	0.09
Study 4	4	Drug Effectiveness	***	***	***	0.96
Study 5	5a	Drug Effectiveness Walk-in	***	***	***	0.64
	5a	Drug Effectiveness Walk-in (Mobile)	*n.s.*	*	*n.s.*	0.15
Study 6	6a	Safety Checklist (Healthcare)	***	***	***	0.86
	6b	Drug Effectiveness Walk-in (Healthcare)	***	***	***	0.87

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; *n.s.* not significant. Effect size was computed using the "A/B Effect" comparison (A and B conditions pooled together and tested against A/B learn and A/B short conditions pooled together, or A/B learn condition only if no A/B short condition was run).

Analyses Conducted Across Multiple Studies

Omnibus and Meta-Analyses on Studies 1–5

Table S54. Descriptive Results over all Studies

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
A/B	2237	33.53	3.26	1.36	0.03	0.06
Policy (A or B)	3179	15.29	3.96	1.22	0.02	0.04

Across Studies 1–5, the effect size estimate for the A/B Effect was $d = 0.55$, 95% CI: [0.50, 0.61]. More than twice as many participants (33.5%) objected to an A/B test of two policies (by rating this decision as very inappropriate or somewhat inappropriate) than to unilaterally implementing one of those policies (15.3%). Additionally, we conducted a mini meta-analysis of Studies 1–5 *(6)*, which yielded an unweighted average effect size estimate of $d = 0.54$, 95% CI: [0.48, 0.59]. *Note that we did not include Study 6 in these analyses because its participants were expert healthcare providers, not laypeople*.

Additional Preregistered and Exploratory Analyses

Science Literacy

The following analyses were conducted across all studies where scientific literacy was measured (Study 2b, Study 3a-3g, Study 4, Study 5a). For these analyses, we pooled together the A/B learn and A/B short condition (if one was run) to create the "A/B Test" condition. The variable "SciLit" was computed as the number of correct answers to nine questions, each of which had two possible answers (Cronbach's $\alpha = .49$). This low reliability coefficient may be due in part to most participants receiving close to the highest possible score on the science literacy scale (median score = 8/9 correct answers; see Table S55). There was no difference in science literacy score between the A and B conditions, $t(1986) = 1.24$, $p = .22$. Because of this, we pooled these conditions together to create the "Policy (A or B)" condition.

Science literacy scores

Table S55. Descriptive Statistics for Scientific Literacy Scores

Condition	*N*	Mean	Median	*SD*	SEM
Policy (A or B)	1988	7.79	8.00	1.31	0.03
A/B Test	1230	7.71	8.00	1.38	0.04

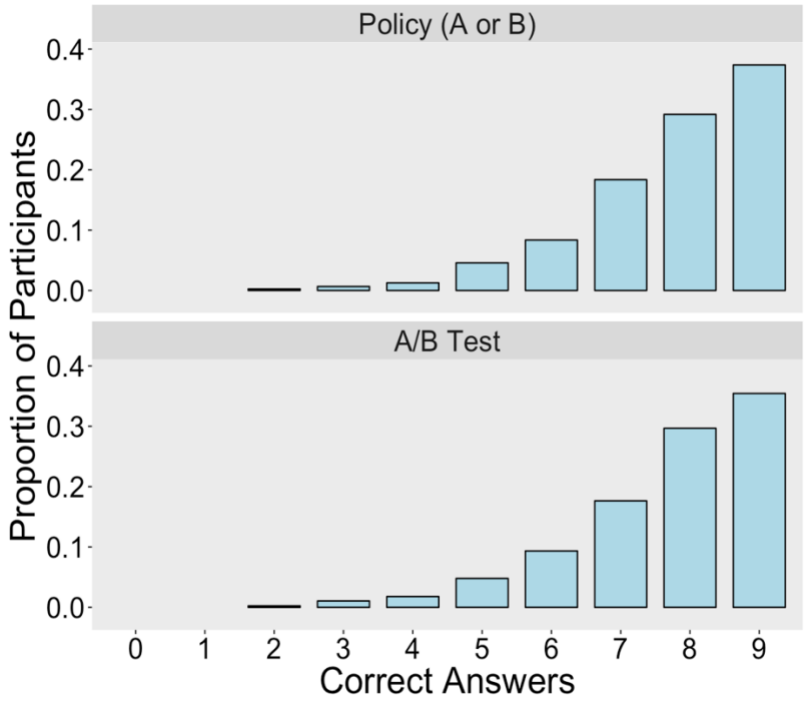


Fig. S20. Distributions of science literacy scores.

Preregistered analysis. Although we report the results of these preregistered analyses, we caution against their interpretation as conclusive evidence for lack of an effect, due to the low reliability and ceiling effect observed in the scale measure.

In the latest preregistration document that mentioned scientific literacy ("Full Experiment on Piloted Domains: Drug Prescriptions, Retirement Savings, Charity, Education, Community Healthcare, and Welfare Services"), we proposed to do the following: "*In each of the four [three] 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."*

Across all studies, there was no effect of scientific literacy on appropriateness for participants who were randomly assigned to an A condition, $b = .04$, $t(982) = 1.36$, $p = .17$, B condition, $b = .01$, $t(1002) = 0.23$, $p = .82$, or A/B Test condition, $b = –.04$, $t(1228) = 1.28$, $p = .20$. Table S56 displays the zero-order correlation between scientific literacy and appropriateness rating, grouped by condition, for each study. These correlations were uniformly low, though in a few cases they achieved statistical significance (see Table S56). We do not comment on whether these few observations of significance are meaningful, as there is a high likelihood that they are due to chance.

Table S56. Correlations Between Science Literacy and Appropriateness Rating

Scenario	Condition	*n*	*r*(SciLit,Appropriateness)
Safety Checklist 2	Policy	180	0.06
	A/B	163	0.04
Genetic Testing	Policy	199	0.18**
	A/B	179	–0.08
Autonomous Vehicles	Policy	204	0.10
	A/B	193	–0.01
Retirement Plans	Policy	201	0.02
	A/B	95	0.06
Health Worker Recruitment	Policy	197	0.01
	A/B	96	0.24*
Poverty Alleviation	Policy	199	–0.04
	A/B	103	–0.12
Teacher Wellbeing	Policy	196	–0.14
	A/B	104	–0.10
Basic Income	Policy	208	0.05
	A/B	96	0.03
Drug Effectiveness	Policy	200	–0.03
	A/B	102	–0.24*
Drug Effectiveness Walk-in	Policy	204	0.05
	A/B	99	–0.11

Note. * $p < .05$, ** $p < .01$

STEM Degree

A second approach to ask whether scientific knowledge or literacy can predict appropriateness ratings of a unilateral policy or randomized experiment is to analyze the effect of

having a STEM degree on the primary dependent measure (appropriateness) within each of the two conditions. Again, we largely find no effect of having a degree in science, technology, engineering, or mathematics on appropriateness ratings in either condition. Table S57 displays these results over all experiments, and Table S58 displays them within each scenario. There were no consistent differences observed between those who have a STEM degree and those who do not. This finding is also consistent with the outcome of Study 6, where healthcare providers, who have extensive science coursework and other training, produced similar A/B effects as we found in laypeople (Studies 1–5).

Table S57. Appropriateness Ratings by Participants With and Without a STEM Degree

Condition	Degree	*N*	Mean Appropriateness	*SD*	SEM
Policy (A or B)	STEM Degree	380	3.98	1.16	0.06
	No STEM Degree	1608	4.02	1.19	0.03
A/B	STEM Degree	188	3.45	1.31	0.10
	No STEM Degree	1042	3.48	1.29	0.04

Table S58. Appropriateness Ratings by STEM Degree Within Each Study

Scenario	Condition	Degree	*N*	Mean	*SD*	SEM
Safety Checklist 2	Policy (A or B)	STEM Degree	23	4.09	1.16	0.24
		No STEM Degree	157	3.87	1.21	0.10
	A/B	STEM Degree	25	3.52	1.39	0.28
		No STEM Degree	138	3.49	1.26	0.11
Genetic Testing	Policy (A or B)	STEM Degree	30	4.23	0.90	0.16
		No STEM Degree	169	4.22	1.12	0.09
	A/B	STEM Degree	22	3.73	1.32	0.28
		No STEM Degree	157	3.59	1.25	0.10
Autonomous Vehicles	Policy (A or B)	STEM Degree	39	3.85	1.53	0.25
		No STEM Degree	165	4.15	1.30	0.10
	A/B	STEM Degree	24	3.58	1.25	0.25
		No STEM Degree	169	3.50	1.34	0.10
Retirement Plans	Policy (A or B)	STEM Degree	39	3.90	1.21	0.19
		No STEM Degree	162	3.81	1.36	0.11
	A/B	STEM Degree	15	3.33	1.29	0.33
		No STEM Degree	80	3.26	1.33	0.15
Health Worker Recruitment	Policy (A or B)	STEM Degree	42	4.14	0.84	0.13
		No STEM Degree	155	4.11	1.05	0.08
	A/B	STEM Degree	14	3.29	1.49	0.40
		No STEM Degree	82	3.77	1.13	0.12
Poverty Alleviation	Policy (A or B)	STEM Degree	46	4.09	1.01	0.15
		No STEM Degree	153	3.75	1.27	0.10
	A/B	STEM Degree	13	3.23	1.42	0.39
		No STEM Degree	90	3.32	1.23	0.13
Teacher Wellbeing	Policy (A or B)	STEM Degree	42	3.50	1.37	0.21
		No STEM Degree	154	3.94	1.20	0.10
	A/B	STEM Degree	19	3.53	1.12	0.26
		No STEM Degree	85	3.29	1.28	0.14
Basic Income	Policy (A or B)	STEM Degree	34	3.76	1.02	0.17
		No STEM Degree	174	3.74	1.17	0.09
	A/B	STEM Degree	22	3.50	1.30	0.28
		No STEM Degree	74	3.68	1.18	0.14
Drug Effectiveness	Policy (A or B)	STEM Degree	42	4.24	1.19	0.18
		No STEM Degree	158	4.48	0.86	0.07
	A/B	STEM Degree	14	3.21	1.31	0.35
		No STEM Degree	88	3.35	1.49	0.16
Drug Effectiveness Walk-In	Policy (A or B)	STEM Degree	43	4.09	1.11	0.17
		No STEM Degree	161	4.17	1.12	0.09
	A/B	STEM Degree	20	3.30	1.42	0.32
		No STEM Degree	79	3.41	1.35	0.15

Sample Demographics and Appropriateness Ratings

The following analyses were conducted across all studies reported in the manuscript (Studies 1–5), except for Study 6, which was conducted with expert healthcare provider participants rather than laypeople. For these analyses, we pooled together the A/B learn and A/B short condition (if one was run) to create the "A/B" condition (see Table S59).

Table S59. Sample demographics

Study	Sample	Condition	*N*	% Male	% White	% with Degree	% >$60,000 / year	Age (*M*)	Age (*SD*)
1	MTurk	Policy (A or B)	193	51.8	75.6	48.7	31.1	34.21	10.92
		A/B	220	46.4	72.3	50.9	33.6	34.88	11.00
2a	MTurk	Policy (A or B)	194	53.1	74.2	58.2	39.2	33.63	11.13
		A/B	192	54.7	71.4	50.5	31.8	34.93	11.69
2b	MTurk	Policy (A or B)	180	48.3	78.3	53.9	40.6	36.49	11.18
		A/B	163	44.8	81.0	52.8	36.2	37.06	12.12
2c	Pollfish	Policy (A or B)	324	25.3	54.3	20.4	11.7	30.33	10.02
		A/B	355	18.6	58.0	19.4	7.6	30.17	10.10
3a	MTurk	Policy (A or B)	199	39.7	77.9	59.8	43.7	35.69	11.54
		A/B	179	44.7	71.5	53.1	43.0	37.26	11.55
3b	MTurk	Policy (A or B)	204	50.5	72.1	53.9	34.3	35.65	11.45
		A/B	193	51.3	75.1	56.0	40.4	38.49	13.44
3c	MTurk	Policy (A or B)	201	43.3	76.6	56.7	39.8	36.82	12.41
		A/B	95	42.1	71.6	52.6	48.4	37.59	11.50
3d	MTurk	Policy (A or B)	197	44.7	75.1	60.9	40.1	37.67	11.81
		A/B	96	37.5	80.2	60.4	40.6	37.15	12.87
3e	MTurk	Policy (A or B)	199	45.7	79.9	59.3	40.7	36.65	11.38
		A/B	103	42.7	77.7	47.6	32.0	36.61	13.93
3f	MTurk	Policy (A or B)	196	40.8	73.5	53.6	45.4	37.44	13.74
		A/B	104	39.4	79.8	56.7	27.9	35.53	13.22
3g	MTurk	Policy (A or B)	208	46.6	76.4	53.8	40.4	36.23	11.26
		A/B	96	45.8	77.1	53.1	42.7	37.78	11.87
4	MTurk	Policy (A or B)	200	51.0	73.0	59.0	40.0	34.81	10.46
		A/B	102	46.1	79.4	60.8	42.2	36.25	11.04
5a	MTurk	Policy (A or B)	204	51.5	74.5	57.8	43.6	35.22	12.57
		A/B	99	48.5	74.7	62.6	41.4	36.40	12.83
5b	Pollfish	Policy (A or B)	480	42.1	67.1	30.8	19.1	34.40	11.19
		A/B	240	41.7	67.7	28.7	16.0	35.49	11.68

Note. "% Male" categorizes into Male or Female/Other. "% White" categorizes into white or any other selection (or multiple selections); see materials. "% with Degree" categorizes into "four year college degree" or greater, or any lesser option. "% > $60,000 / year" categorizes into whether participants selected "Between $60,000 and $80,000" or greater per year or any lesser option. This question specified income as "total annual income from everyone in your household." In the Pollfish samples, this level was approximated from the income levels provided by the platform.

Tables S60 – S64 report the mean appropriateness ratings given by demographic subgroups (grouped by condition).

Table S60. Demographic Differences Between Conditions (Sex)

Condition	Sex	*n*	*M* Appropriateness	*SD*	SEM
A/B	Female	1312	3.22	1.36	0.04
	Male	925	3.31	1.35	0.04
Policy (A or B)	Female	1773	3.96	1.24	0.03
	Male	1406	3.97	1.20	0.03

Note. Here in subsequent demographic tables, we include all participants who provided demographic information about themselves. Demographic information was not collected in the sample of healthcare providers (Study 6).

Table S61. Demographic Differences Between Conditions (Race/Ethnicity)

Condition	Race / Ethnicity	*n*	*M* Appropriateness	*SD*	SEM
A/B	Nonwhite	629	3.14	1.39	0.06
	White	1603	3.30	1.35	0.03
Policy (A or B)	Nonwhite	880	3.92	1.26	0.04
	White	2281	3.98	1.20	0.03

Table S62. Demographic Differences Between Conditions (Education)

Condition	Education	*n*	*M* Appropriateness	*SD*	SEM
A/B	No Degree	1210	3.26	1.36	0.04
	College Degree	1027	3.24	1.37	0.04
Policy (A or B)	No Degree	1627	3.96	1.24	0.03
	College Degree	1552	3.96	1.21	0.03

Table S63. Demographic Differences Between Conditions (Income)

Condition	Household Income	*n*	*M* Appropriateness	*SD*	SEM
A/B	< $60,000 per year	1538	3.23	1.36	0.03
	≥ $60,000 per year	684	3.30	1.36	0.05
Policy (A or B)	< $60,000 per year	2057	3.98	1.20	0.03
	≥ $60,000 per year	1067	3.94	1.25	0.04

Table S64. Demographic Differences Between Conditions (Age)

Condition	Age	*n*	*M* Appropriateness	*SD*	SEM
A/B	≤ 33 Years	1217	3.23	1.36	0.04
	> 33 Years	1020	3.28	1.36	0.04
Policy (A or B)	≤ 33 Years	1742	4.00	1.20	0.03
	> 33 Years	1437	3.92	1.25	0.03

God, Intuition, and Science (GIS) scale

Descriptive statistics and scale

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

1 – Strongly disagree
2 – Somewhat disagree
3 – Neither disagree nor agree
4 – Somewhat agree
5 – Strongly agree

Table S65. Intercorrelations Among Items

	Intuition	God
God	0.64	
Science	–0.44	–0.34

After reverse-scoring the item "Science," these three items were summed into a scale score (Cronbach's alpha = .74) with a minimum of 3 and a maximum of 15, where higher numbers indicate greater belief in intuitive knowledge.

Table S66. Descriptive Scale Scores

Condition	*N*	Mean Score	Median	*SD*	sem
Policy (A or B)	2777	7.41	8	3.13	0.06
A/B	2065	7.74	8	3.05	0.07

Preregistered analyses. Taken from the Study 2a preregistration. In future preregistrations, we included these questions but did not preregister this analysis or a prediction. *"In each of the four conditions, regress appropriateness on the intuitive beliefs variable (answers to the three questions querying beliefs in God, intuition, and science)."*

Across all studies, there was no effect of GIS scale score on appropriateness for participants who were randomly assigned to an A condition, $b = –.019$, $t(1353) = 1.89$, $p = .07$, or A/B condition, $b = –.011$, $t(1228) = 1.28$, $p = .20$. However, participants assigned to rate a B condition did show a negative relationship between their GIS scale score and appropriateness rating of the decision, $b = –.024$, $t(1420) = 2.39$, $p = .02$. A similar negative relationship emerged when pooling together participants in an A or a B condition, $b = –.022$, $t(2775) = 2.95$, $p = .003$, suggesting that participants who scored higher on this scale gave lower appropriateness ratings to unilateral policy implementations, but not to randomized experiments.

Pilot Study Analyses

Summary table of pretest and pilot results

Table S67. Descriptive Results Over all Pretest and Pilot Studies

Condition	*N*	% Objecting	*M* Appropriateness	*SD*	SEM	95% CI +/–
Policy (A or B)	1575	20.76	3.76	1.28	0.03	0.06
A/B Test	504	36.11	3.19	1.30	0.06	0.11

Note. One series of pilot studies (all pilots in "abE10") was run using a within-subjects design. Because of this, the overall N in the Policy (A or B) conditions reports number of observations, and not number of unique participants.

Across all scenarios that were pretest and pilot tested, the effect size estimate for the A/B Effect was $d = 0.45$, 95% CI: [0.35, 0.55]. Although this effect is smaller than the effect size reported on Studies 1-5, it remained statistically significant, $t(2077) = 8.81$, $p = 2.2 \times 10^{-16}$. Tables S68 and S69 report descriptive statistics for each pilot study.

Expanded results of all pretest and pilot studies

(See tables S68 and S69)

Table S68. Pilot Test Descriptive Statistics for A, B, and A/B Conditions

Scenario (see Materials)	R File	Condition	Vignette Keyword	*N*	*M* Rating	*SD*	SEM	Run as Full Experiment?
Online Dating Pilot 1	abE4	A	Low %	20	3.15	1.27	0.28	No
	abE4	B	High %	19	3.68	1.29	0.30	No
	abE4	A/B	A/B	36	3.53	1.11	0.18	No
Autonomous Vehicles Pilot 1	abE6	A	Accelerate	20	1.90	1.17	0.26	No
	abE6	B	Brake	22	3.73	1.24	0.26	No
	abE6	A/B	A/B Test	33	2.94	1.39	0.24	No
Autonomous Vehicles Pilot 4	abE8	A	Lever	29	4.59	0.95	0.18	Study 3b
	abE8	B	Automatic	32	4.16	1.25	0.22	Study 3b
	abE8	A/B	A/B	28	3.39	1.47	0.28	Study 3b
Genetic Testing Pilot 3	abE8	A	Cancer	30	3.07	1.36	0.25	No
	abE8	B	Dementia	31	3.45	1.48	0.27	No
	abE8	A/B	A/B	29	3.21	1.24	0.23	No
Genetic Testing Pilot 4	abE8	A	Actionable	29	4.41	1.02	0.19	Study 3a
	abE8	B	All Results	31	3.84	1.16	0.21	Study 3a
	abE8	A/B	A/B Test	27	2.56	1.42	0.27	Study 3a
Online Dating Pilot 4	abE8	A	Very	28	3.86	1.01	0.19	No
	abE8	B	Somewhat	31	3.42	1.15	0.21	No
	abE8	A/B	A/B	32	3.44	1.19	0.21	No
Drug Effectiveness Pilot 1	abE10	A	Drug A	26	4.38	1.06	0.21	Study 4
	abE10	B	Drug B	28	4.50	0.64	0.12	Study 4
	abE11	A/B	A/B	29	2.72	1.46	0.27	Study 4
Drug Effectiveness Walk-in Pilot 1	abE10	A	Drug A	30	3.97	1.03	0.19	Study 5a/b
	abE10	B	Drug B	30	4.10	0.96	0.18	Study 5a/b
	abE11	A/B	A/B	30	3.03	1.45	0.26	Study 5a/b
Resident Hours Pilot 1	abE10	A	16 Hours	26	4.31	1.12	0.22	No
	abE10	B	24 Hours	28	3.86	1.41	0.27	No
	abE11	A/B	A/B	28	3.82	1.16	0.22	No
Basic Income Pilot 2	abE10	A	$1,000	32	3.75	1.39	0.25	Study 3g
	abE10	B	$500	36	3.89	1.17	0.19	Study 3g
	abE11	A/B	A/B	27	3.04	1.22	0.24	Study 3g
Health Worker Recruitment Pilot 1	abE10	A	Social	30	4.53	0.94	0.17	Study 3d
	abE10	B	Career	30	4.23	1.07	0.20	Study 3d
	abE11	A/B	A/B	27	3.63	1.11	0.21	Study 3d
Colonoscopies Pilot 1	abE10	A	Humor	30	3.73	1.17	0.21	No
	abE10	B	Prescheduled	29	3.41	1.38	0.26	No
	abE11	A/B	A/B	28	3.36	1.10	0.21	No
Simple SUPPORT Pilot	abE10	A	Low Oxygen	32	3.91	1.15	0.20	No
	abE10	B	High Oxygen	36	3.39	1.38	0.23	No
	abE11	A/B	A/B	28	3.18	1.36	0.26	No
Teacher Wellbeing Pilot 1	abE10	A	Bonus	26	4.04	1.11	0.22	Study 3f
	abE10	B	Vacation	28	4.04	1.14	0.22	Study 3f
	abE11	A/B	A/B	31	3.10	1.11	0.20	Study 3f
Basic Income Pilot 1	abE10	A	Poorest	26	3.77	1.37	0.27	No
	abE10	B	All	28	3.57	1.35	0.25	No
	abE11	A/B	A/B	30	3.27	1.36	0.25	No
Poverty Alleviation	abE10	1B	Roof	31	3.87	1.28	0.23	Study 3e
	abE10	2B	Training	36	4.19	1.17	0.19	Study 3e
	abE11	A/B	A/B	30	3.20	1.13	0.21	Study 3e
Retirement Plans	abE10	1A	Increase	26	4.19	1.10	0.21	Study 3c
	abE10	2A	Highlight	32	4.06	1.24	0.22	Study 3c
	abE11	A/B	A/B	31	2.74	1.34	0.24	Study 3c

Note. All pilot studies in "abE10" were run within-subjects. *N* reports observations in these cases.

Table S69. Descriptive Statistics for A and B Condition Pretesting

Scenario (see Materials)	R File	Condition	Keyword	*N*	*M* Rating	SD	SEM	Run as Full Experiment?
Autonomous Vehicles Pilot 3	abE7	A	Lever	29	4.21	1.24	0.23	No
	abE7	B	Automatic	28	4.68	0.55	0.10	No
Autonomous Vehicles Pilot 2	abE7	A	Match Speed	27	2.93	1.33	0.26	No
	abE7	B	Brake	28	4.18	0.94	0.18	No
Genetic Testing Pilot 1	abE7	A	Cancer	29	3.79	1.18	0.22	No
	abE7	B	Dementia	27	3.52	1.22	0.23	No
Genetic Testing Pilot 2	abE7	A	Actionable	32	4.03	1.00	0.18	No
	abE7	B	All Results	25	3.96	1.10	0.22	No
Online Dating Pilot 3	abE7	A	Common Friends	29	3.21	1.01	0.19	No
	abE7	B	Profile Questions	29	4.07	0.96	0.18	No
Online Dating Pilot 2	abE7	A	Very Similar	30	3.87	1.36	0.25	No
	abE7	B	Somewhat Similar	30	3.57	1.04	0.19	No
Music Streaming Pilot 1	abE10	A	More Ads	30	3.37	1.25	0.23	No
	abE10	B	Remove Songs	30	2.97	1.35	0.25	No
Music Streaming Pilot 2	abE10	A	Shorter Ads	32	3.03	1.26	0.22	No
	abE10	B	Longer Ads	36	2.86	1.36	0.23	No
Poverty Alleviation	abE10	1A	Livestock	30	3.83	1.21	0.22	No
	abE10	2A	Cash	32	3.50	1.46	0.26	No
Retirement Plans	abE10	1B	Decrease	28	3.00	1.33	0.25	No
	abE10	2B	Automatic	36	3.44	1.46	0.24	No

Note. All pilot studies in "abE10" were run within-subjects. Thus, *N* for each condition reports observations, and not individuals, in these cases.

Robustness Checks on Selection Bias and Multiple Comparisons

Strategy and Approach

Our strategy of first conducting many pilot studies and then selecting a subset of these pilots for full studies was intended to demonstrate the broad applicability of the A/B effect while conserving scarce resources; however, it raises two potential concerns about the validity of our results. First, it creates the possibility for a type of "file drawer" effect. As reported in the main text, we took several steps to avoid the obvious file drawer problem: first, by articulating clear criteria for promoting pilot studies to full studies; second, by employing power analyses to help ensure that our full studies were adequately powered; and third by pre-registering our hypotheses and analysis plans. Nevertheless, it could still be that the A/B effect either arises only in special circumstances or is highly fragile with respect to specific wording and/or framing. By conducting many variations in the pilot stage and only reporting results for the vignettes that graduate to full studies we could create the impression that the effect is more widespread and/or more robust than it really is. A second, related concern is that the large total number of studies—pilots and full studies—creates the possibility for a multiple comparisons problem. Even if the A/B effect is not real, that is, we may have found significant differences in some of our experiments simply because of random fluctuations in subject responses across vignettes. *As with the other omnibus analyses reported above, we do not include Study 6, because its participants were expert healthcare providers, not laypeople as in Studies 1–5.*

P-Curve Analysis

To confirm that our main finding is not an artifact of selection effects and/or multiple comparisons, we report here two additional robustness checks. First, to check that we have not inadvertently created our own file drawer problem, we perform a p-curve analysis [1] on just our main study results with $p \leq 0.05$. Per Simonsohn, Nelson, and Simmons *(7)*, we included all and only those *p*-values for main effects (here, the A/B effect comparison) for which $p < 0.05$. Following *(7)*, Fig. S21 compares our observed p-curve with two null models: a null in which no effect exists (corresponding to a uniform, or flat, p-curve); and a null model in which an effect exists but where the experiment has extremely low power (33%), equivalent to a very small, albeit real, effect. Fig. S21 shows that our observed p-curve is strongly right-skewed, as one would expect when a set of studies contains evidentiary value *(7)*, and that we can easily reject both null models (see Fig. S22).

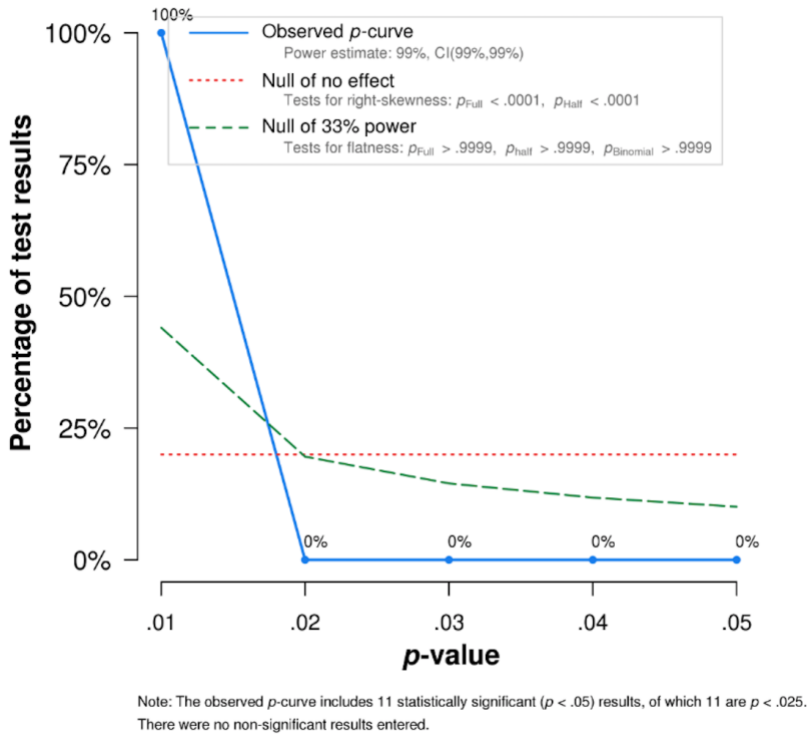


Fig S21. P-curve analysis for all studies with $p < 0.05$.

	Binomial Test (*Share of results p<.025*)	Continuous Test (*Aggregate with Stouffer Method*)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. *(Right skew)*	p=.0005	Z=-16.08, p<.0001	Z=-15.46, p<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p>.9999	Z=12.01, p>.9999	Z=12.85, p>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Fig S22. P-curve analysis for all studies with $p < 0.05$.

Bonferroni-corrected *p*-values

Across all of our full experiments, we conducted 16 total A/B Effect comparisons as preregistered focal tests of our research question. Because traditional statistical significance testing yields increasingly inflated false positive rates with each additional comparison *(e.g., 8)*, we sought to correct our 16 A/B Effect tests to account for this possibility. One approach is to use Hierarchical Linear Model (HLM) analysis to control for selection effects and multiple comparisons (see HLM section below). A simpler and in some cases more interpretable correction technique is the Bonferroni familywise alpha adjustment procedure. We note that this

analysis is highly conservative, and in all cases, generates corrected *p*-values that are greater (i.e., more conservative) than alternative correction approaches *(9)*.

We therefore recomputed *p*-values for the A/B effect in each experiment after adjusting the alpha level threshold to account for the total number of full experiments run throughout this project (.05 / 16 = .003125). Table S70 reports the original and corrected *p*-values for each experiment.

Table S70. A/B Effect *p*-values After Bonferroni Correction

Study		Scenario	Unadjusted A/B Effect *p*-value	Bonferroni-adjusted *p*-value	Passes Correction?
Study 1	1	Safety Checklist	< .00001	< .00016	Yes
Study 2	2a	Safety Checklist Rep. (Exact)	< .00001	< .00016	Yes
	2b	Safety Checklist Rep. (Varied)	.00281	.00450	Yes
	2c	Safety Checklist Rep. (Mobile)	< .00001	< .00016	Yes
Study 3	3a	Genetic Testing	< .00001	< .00016	Yes
	3b	Autonomous Vehicles	.00002	.00032	Yes
	3c	Retirement Plans	.00084	.00128	Yes
	3d	Health Worker Recruitment	.00182	.02912	Yes
	3e	Poverty Alleviation	.00068	.01088	Yes
	3f	Teacher Wellbeing	.00094	.01504	Yes
	3g	Basic Income	*n.s.*	*n.s.*	N/A
Study 4	4	Drug Effectiveness	< .00001	< .00016	Yes
Study 5	5a	Drug Effectiveness Walk-in	< .00001	< .00016	Yes
	5a	Drug Effectiveness Walk-in (Mobile)	*n.s.*	*n.s.*	N/A
Study 6	6a	Safety Checklist (Healthcare)	< .00001	< .00016	Yes
	6b	Drug Effectiveness Walk-in (Healthcare)	< .00001	< .00016	Yes

Note. n.s. not significant. Bonferroni-corrected *p*-values that remain below the alpha threshold of .05 are said to pass the correction.

To summarize, we found that using the most conservative possible *p*-value correction did not change any observation of statistical significance: all 14 of the 16 tests that were significant before correction remained significant after correction.

Hierarchical Linear Model: Selection Effects and Pilot Data

We next checked and corrected for selection bias and multiple comparisons by fitting a hierarchical linear model (HLM) for each of our outcome variables using all data from all studies, including pilot studies that we did not carry forward to a full-sample study, where we ran an A, B, and A/B condition. HLM assumes that effects (in our case the treatment effects of each vignette) are randomly sampled from a normal distribution that is centered on the mean of the effects for each of the fixed effects, which in our case comprise the reference condition (the A or

B condition that has the lower approval rating of the two) and the A/B condition (that is, participants who were assigned to rate any A/B test). By separately estimating the difference between the reference condition (A or B) and the A/B condition, we effectively control for conducting multiple comparisons *(10)*. In addition, by including all pilots and full experiments, the overall treatment effect (the estimated difference between all reference and all A/B conditions) can be said to account for experimenter bias in selecting which piloted scenarios to run as a full sample experiment.

Specifically, we fit a model of the form

$$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2), \text{ for } i=1,...,n_{j[i]}, j=1,...,J$$
$$\alpha_j \sim N(\gamma_0, \sigma_\alpha^2), \text{ for } j = 1,...,J$$
$$\beta_j \sim N(\xi_0, \sigma_\beta^2), \text{ for } j = 1,...,J$$

where y_i is subject i's response, j[i] is an index function for the vignette to which subject i is assigned, and x_i is an indicator for the AB treatment (in all cases, the reference condition is the policy condition with the lower appropriateness rating), σ_y^2 is the within-vignette variation, σ_α^2 is the between-vignette variation for the policy condition, and σ_β^2 is the between-vignette variation for the A/B Effect. In model 1, y_i corresponds to subjects continuous appropriateness rating for the vignette (1–5), and in model 2, y_i corresponds to the subjects rating the vignette as either inappropriate or very inappropriate (dichotomized into 1 or 0). In both models, the main coefficients of interest are ξ_0, the mean of the distribution that the β_j are drawn from, and γ_0 is the mean of the reference conditions. Figure S23 clarifies the interpretation of σ_α, σ_β, σ_y, and ξ_0, and γ_0.

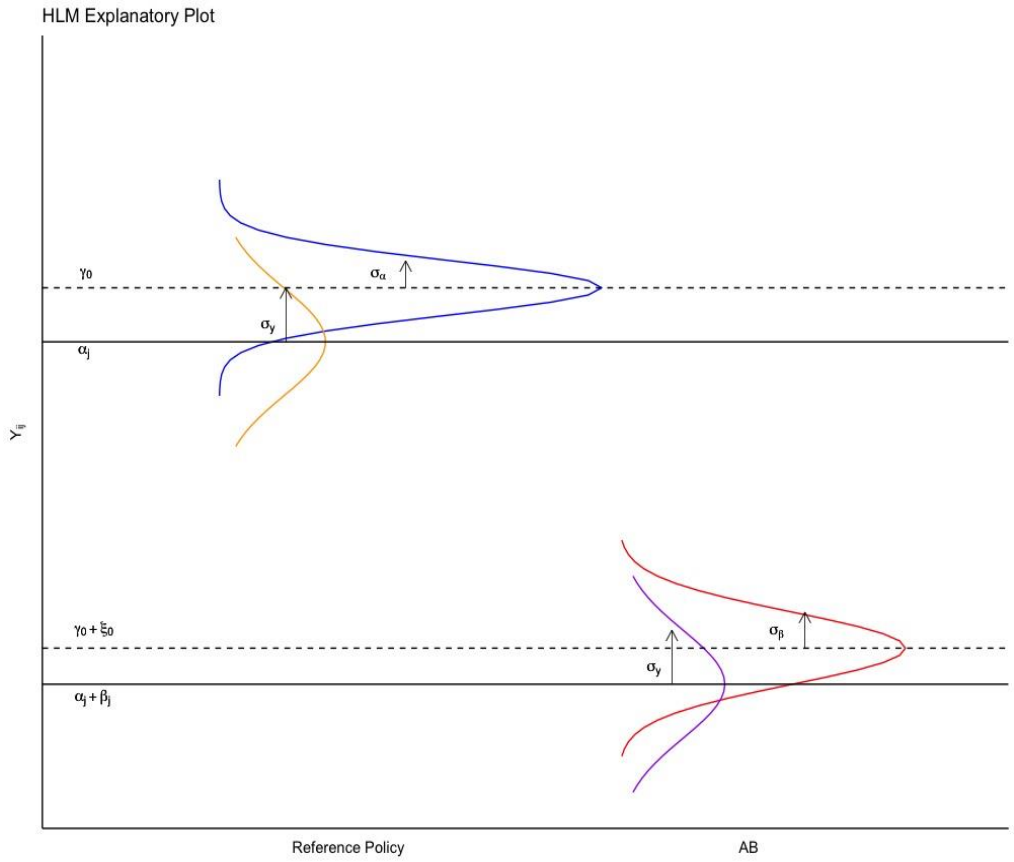


Figure S23. Schematic illustration of the hierarchical linear model (HLM). The coefficients for each vignette for the reference condition, the α_j, are drawn from the distribution in blue centered at γ_0 (the average rating for all vignettes in the reference condition). One example of α_j is shown, and the response variable for subjects in the Policy condition with lower average rating of vignette j would be drawn from the distribution in yellow. Analogously, the coefficients of the A/B Effect, the β_j, are drawn from a distribution centered at ξ_0 – the distribution in red shows this distribution shifted by γ_0. Responses of the subjects in the AB condition of vignette j are drawn from the purple distribution, centered at $\alpha_j + \beta_j$.

Table S71 shows our coefficient estimates with standard errors for both HLMs. The main effects of the A/B condition (ξ_0) are in the predicted direction and significant at the $p<0.05$ level in both models. This indicates that the treatment effect of being assigned to rate an A/B test (as opposed to a unilateral policy) caused greater inappropriateness ratings and a greater likelihood to object by labeling the decision as inappropriate. We are therefore confident that the A/B Effect remains significant and meaningful when considering all available data from Studies 1–5 and all pilot experiments that were not run as a full experiment.

Table S71. Coefficients for HLM models.

	Model 1 (Appropriateness, 1–5)			Model 2 (Inappropriateness 1/0)		
Variable	Estimate	Standard Error	P Value	Estimate	Standard Error	P Value
γ_0	3.7813259	0.07414971		0.19671802	0.01709903	
ξ_0	–0.4423172	0.08310293	< 0.0001	0.11300648	0.02407846	< 0.0001
σ_y	1.2980975			0.43468171		
σ_α	0.2724621			0.05373279		
σ_β	0.2857679			0.07701072		

Note. Model 1 is for the 5-point appropriateness rating dependent variable; Model 2 is for the binary "inappropriate" judgment dependent variable. In both cases, the main effect is in the predicted direction and significant at the $p < 0.0001$ level.

Hierarchical Linear Model: Treatment Effect Heterogeneity

To what extent was our treatment effect heterogeneous with regard to the average treatment effect discussed above? In considering possible corrections to standard errors for repeated treatment effect estimates, Gelman and Hill *(10)* wrote:

> "We do not recommend classical methods that alter p values or (equivalently) make confidence intervals wider. Instead, we prefer multilevel modeling, which shifts point estimates and their corresponding intervals closer to each other (that is, performs partial pooling) where necessary—especially when much of the variation in the data can be explained by noise."

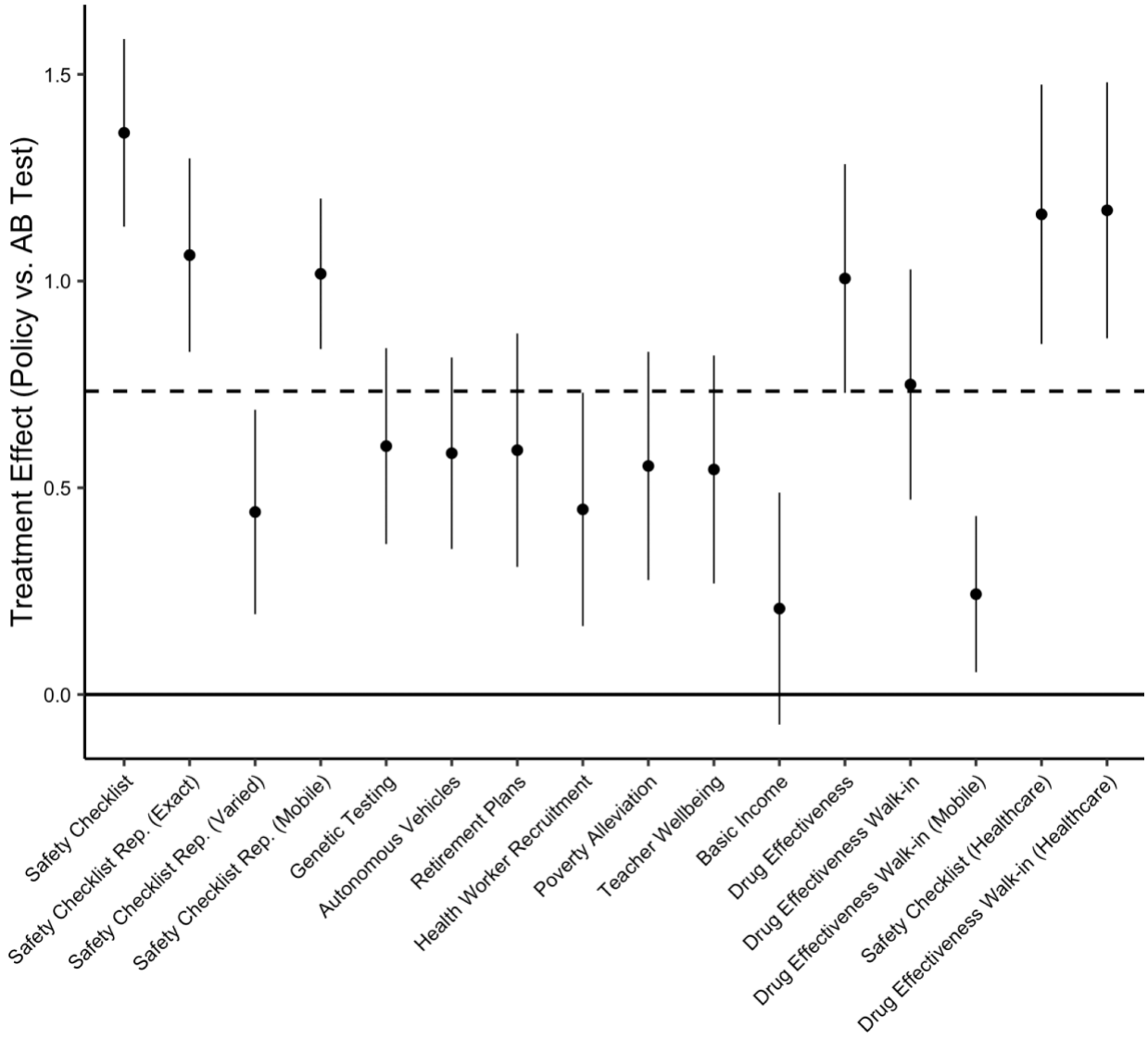
To ensure that the A/B Effect was robust to random effect variation in study domain, and to visualize treatment effect variation over domains, we ran a final HLM analysis on participants' continuous appropriateness ratings for all complete studies run during this project (i.e., excluding pretest and pilot study data). This analysis entered two variables as predictors:

1. Participants' assignment to a policy (A or B) or experiment (A/B Test) condition entered as a fixed effect grouping variable (i.e., the treatment).
2. The unique study that each participant completed (ranging from 1 (Safety Checklist) to 16 (Drug Effectiveness Walk-In (Healthcare)) entered as a random effect.

The resulting model yielded an overall intercept of 3.21 (standard error = 0.11) with an overall treatment effect estimate of 0.73 (standard error = 0.10). Figure S24 shows this average treatment effect of being assigned to rate an A/B test (assigned a 0) or a unilateral policy (assigned 1), and the treatment effect within each study. 95% CI are computed using within-study standard error estimates generated by the HLM *(10)*.

Using HLM analysis to control for random effect variation within the 16 studies run during this project, we observed no notable differences in effect size or statistical significance. The only study that didn't achieve traditional statistical significance (Basic Income) also yielded a 95% CI that includes zero (see Figure S24); all other studies yielded an A/B Effect.

Figure S24. Average and within-study treatment effect estimates using HLM.



Prior Beliefs and Post-Study Probability (PSP)

How much should we update our belief about the existence of a so-called A/B Effect? We explored this question by considering the post-study probability (PSP) of the existence of an effect given the following inputs:

- α is our preregistered false positive rate and *criterion* for statistical significance (.05 in all cases).

- *Prior belief* π is the probability from the outset of the study that an effect exists. We consider the PSP across a range of reasonable prior beliefs.

- $1 - \beta$ is the *power* of each study to detect an A/B Effect, if one truly exists. As described earlier in the supplementary information, our experiments were preregistered to be run with statistical power of .80, or an *a priori* 80% chance to detect as effect as small as d = 0.28, if such an effect exists. We refrain from calculating observed power *post hoc* on the basis of any single study.

As proposed by Maniadis, Tufano, and List *(12)*, the PSP can be calculated using the formula:

$$PSP = \frac{(1 - \beta)\pi}{(1 - \beta)\pi + \alpha(1 - \pi)}$$

Table S72 displays the resulting PSP, or post-study-probability, that represents the new belief we should hold about the existence of an A/B Effect given a range of prior beliefs and our preregistered estimate for statistical power. These values suggest that regardless of the chosen prior, our new prior belief (PSP) in an A/B Effect should be greater than what it was at the study's outset. Even the most extreme skeptics—those with a prior belief of .01—should update their prior belief in the existence of an A/B Effect in a positive direction.

Table S72. Post-Study-Probabilities Given Prior Belief and Power

Prior Belief π	PSP for power = 0.8 (*preregistered*)
.01	.14
.10	.64
.25	.84
.50	**.94**
.75	.98
.90	.99

Note: PSP computed using Formula 1 in *(12)*. The bolded row corresponds to our subjective prior belief in an A/B Effect *before* conducting studies 1–6 (.50) and the PSP following these studies (.94).

What value should we take as the prior probability that the A/B Effect exists? Given previous discussion of the topic *(13)*, anecdotal observations of public outrage to experimentation (summarized in Main Text), but a complete lack of previous empirical work, we think a reasonable prior belief to have held at the outset of the project is .50. Given this prior, and estimates of statistical power computed *a priori*, we consider a PSP of .94 be a reasonable estimate for our new prior belief in the A/B Effect as measured using our experimental paradigm and framing of an A/B experiment.

References

1. Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Meth.* 41:1149–1160.
2. National Science Board (2016) *Science and Engineering Indicators 2016* (Arlington, VA: National Science Foundation).
3. Heck PR, Meyer MN (2019). Information avoidance in genetic health: Perceptions, norms, and preferences. *Soc. Cogn.,* forthcoming.
4. Heck PR, Meyer MN (2019). Population whole exome screening: Primary care provider attitudes about preparedness, information avoidance, and nudging. *Med. Clin.*, forthcoming.
5. MacCallum RC, Zhang S, Preacher K, Rucker DD (2002) On the practice of dichotomization of quantitative variables. *Psychol. Meth.* 7:19–40.
6. Goh JX, Hall JA, Rosenthal R (2016) Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Soc. & Pers. Psychol. Compass* 10:535–549.
7. Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: a key to the file-drawer. *J Exp. Psychol.: General* 143:534–547.
8. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22:1359–1366.
9. List JA, Shaikh AM, Xu Y. (2019) Multiple hypothesis testing in experimental economics. *Exp. Econ.*
10. Gelman A, Hill J, Yajima N (2012) Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Effectiveness* 5:189–211.
11. Gelman A, Hill J (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (New York, Cambridge University Press).
12. Maniadis Z, Tufano F, List JA (2014) One swallow doesn't make a summer: new evidence on anchoring effects. *Am. Econ. Rev*. 104:277-90.
13. Meyer MN (2015) Two cheers for corporate experimentation: the A/B illusion and the virtues of data-driven innovation. *Colorado Tech. Law J.* 13:273–331.