Moral dilemmas and trust in leaders during a global health crisis

Jim A. C. Everett†[1], Clara Colombatto†[2], Edmond Awad[3], Paulo Boggio[4], Björn Bos[5], William J. Brady[2], Megha Chawla[2], Vladimir Chituc[2], Dongil Chung[6], Moritz A. Drupp[5], Srishti Goel[2], Brit Grosskopf[3], Frederik Hjorth[7], Alissa Ji[2], Caleb Kealoha[2], Judy S. Kim[2], Yangfei Lin[3], Yina Ma[8,9], Michel André Maréchal[10], Federico Mancinelli[11], Christoph Mathys[11,12,13], Asmus L. Olsen[7], Graeme Pearce[3], Annayah M. B. Prosser[14], Niv Reggev[15], Nicholas Sabin[16], Julien Senn[10], Yeon Soon Shin[2], Walter Sinnott-Armstrong[17], Hallgeir Sjåstad[18], Madelijn Strick[19], Sunhae Sul[20], Lars Tummers[21], Monique Turner[22], Hongbo Yu[23], Yoonseo Zoh[2], Molly J. Crockett*[2]

†These authors contributed equally

*Correspondence to: molly.crockett@yale.edu

Affiliations

[1] School of Psychology, University of Kent, Canterbury, UK
[2] Department of Psychology, Yale University, New Haven, CT, USA
[3] Department of Economics, University of Exeter, Exeter, UK
[4] Social and Cognitive Neuroscience Laboratory, Mackenzie Presbyterian University, São Paulo, Brazil
[5] Department of Economics, University of Hamburg, Hamburg, Germany
[6] Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea
[7] Department of Political Science, University of Copenhagen, Copenhagen, Denmark
[8] State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China
[9] Chinese Institute for Brain Research, Beijing, China
[10] Department of Economics, University of Zurich, Zurich, Switzerland
[11] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[12] Interacting Minds Centre, Aarhus University, Aarhus, Denmark
[13] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland
[14] Department of Psychology, University of Bath, Bath, UK
[15] Department of Psychology and Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Be'er Sheva, Israel
[16] Department of Management, Faculty of Management and Economics, Universidad de Santiago de Chile, Santiago, Chile
[17] Department of Philosophy and Kenan Institute for Ethics, Duke University, Durham, NC, USA

[18] Department of Strategy and Management, Norwegian School of Economics, Bergen, Norway
[19] Department of Psychology, Utrecht University, Utrecht, The Netherlands
[20] Department of Psychology, Pusan National University, Busan, South Korea
[21] School of Governance, Utrecht University, Utrecht, The Netherlands
[22] Department of Communication, Michigan State University, East Lansing, MI, USA
[23] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, USA

Abstract

Trust in leaders is central to citizen compliance with public policies. One potential determinant of trust is how leaders resolve conflicts between utilitarian and non-utilitarian ethical principles in moral dilemmas. Past research suggests that utilitarian responses to dilemmas can both erode and enhance trust in leaders: sacrificing some people to save many others (‘instrumental harm’) reduces trust, while maximizing the welfare of everyone equally (‘impartial beneficence’) may increase trust. In a multi-site experiment spanning 22 countries on six continents, participants (N = 23,929) completed self-report (N = 17,591) and behavioral (N = 12,638) measures of trust in leaders who endorsed utilitarian or non-utilitarian principles in dilemmas concerning the COVID-19 pandemic. Across both the self-report and behavioral measures, endorsement of instrumental harm decreased trust, while endorsement of impartial beneficence increased trust. These results show how support for different ethical principles can impact trust in leaders, and inform effective public communication during times of global crisis.

Protocol Registration Statement

The Stage 1 protocol for this Registered Report was accepted in principle on November 13th 2020. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.13247315.v1.

During times of crisis, such as wars, natural disasters, or pandemics, citizens look to leaders for guidance. Successful crisis management often depends on mobilizing individual citizens to change their behaviors and make personal sacrifices for the public good [1]. Crucial to this endeavour is trust: citizens are more likely to follow official guidance when they trust their leaders [2]. Here, we investigate public trust in leaders in the context of the COVID-19 pandemic, which continues to threaten millions of lives around the globe at the time of writing [3,4].

Because the novel coronavirus is highly transmissible, a critical factor in limiting pandemic spread is compliance with public health recommendations such as social distancing, physical hygiene and mask wearing [5,6]. Trust in leaders is a strong predictor of citizen compliance with a variety of public health policies [7–12]. During pandemics, trust in experts issuing public health guidelines is a key predictor of compliance with those guidelines. For example, during the avian influenza pandemic of 2009 (H1N1), self-reported trust in medical organizations predicted self-reported compliance with protective health measures and vaccination rates [13,14]. During the COVID-19 pandemic, data from several countries show that public trust in scientists, doctors and the government is positively associated with self-reported compliance with public health recommendations [15,16,17,18]. These data suggest that trust in leaders is likely to be a key predictor of long-term success in containing the COVID-19 pandemic around the globe. However, the factors that *determine* trust in leaders during global crises remain understudied.

One possible determinant of trust in leaders during a crisis is how they resolve moral dilemmas that pit distinct ethical principles against one another. The COVID-19 pandemic has raised particularly stark dilemmas of this kind, for instance whether to prioritize young and otherwise healthy people over older people and those with chronic illnesses when allocating scarce medical treatments [19,20]. This dilemma and similar others highlight a tension between two major approaches to ethics. Consequentialist theories – of which utilitarianism is the most well-known exemplar [21] – posit that only consequences should matter when making moral decisions. Because younger, healthier people are more likely to recover and have longer lives ahead of them, utilitarians would argue that they should be prioritized for care because this is likely to produce the best overall consequences [22–24]. In contrast, non-utilitarian theories of morality, such as deontological theories [25–29], argue that morality should consider more than just consequences, including rights, duties, and obligations (see Supplementary Note 1 for further details). Non-utilitarians, on deontological grounds, could argue that everyone who is eligible (e.g., by being a citizen and/or contributing through taxes or private health insurance) has an equal right to receive medical care, and therefore it is wrong to prioritize some over others [30]. While it is unlikely that ordinary citizens *explicitly* think about moral issues in terms of specific ethical theories [21,31], past work shows that these philosophical concepts explain substantial variance in the moral judgments of ordinary citizens [32,33], including those in the context of the COVID-19 pandemic [34].

There is robust evidence that people who endorse utilitarian principles in sacrificial dilemmas – deeming it morally acceptable to sacrifice some lives to save many others – are seen as less moral and trustworthy, chosen less frequently as social partners, and trusted less in economic exchanges than people who take a non-utilitarian position and reject sacrificing some to save many [35–40]. This suggests that leaders who take a utilitarian approach to COVID-19 dilemmas will be trusted less than leaders who take a non-utilitarian approach. Anecdotally, some recent case studies of public communications are consistent with this hypothesis. In the United States, for example, public discussions around whether to reopen schools and the economy versus remain in lockdown highlighted tensions between utilitarian approaches and other ethical principles, with some leaders stressing an imperative to remain in lockdown to prevent deaths from COVID-19 (consistent with deontological principles), but others arguing that lockdown also has costs and these need to be weighed against the costs of pandemic-related deaths (consistent with utilitarian principles; see Supplementary Note 2). Those who appealed to utilitarian arguments – such as President Donald Trump, who argued "we cannot let the cure be worse than the problem itself" [41] and Texas Lieutenant Governor Dan Patrick, who suggested that older Americans might be "willing to take a chance" on their survival for the sake of their grandchildrens' economic prospects [42] – were met with widespread public outrage [43]. Likewise, when leaders in Italy suggested prioritizing young and healthy COVID-19 patients over older patients when ventilators became scarce, they were intensely criticized by the public [44]. Mandatory contact tracing policies, which have been proposed on utilitarian grounds, have also faced strong public criticisms about infringement of individual rights to privacy [45–47].

While past research and recent case studies suggest that utilitarian approaches to pandemic dilemmas are likely to erode trust in leaders, other evidence suggests this conclusion may be premature. First, some work shows that utilitarians are perceived as more competent than non-utilitarians [38], and to the extent that trust in leaders is related to perceptions of their competence [2], it is possible that utilitarian approaches to pandemic dilemmas will increase rather than decrease trust in leaders. Second, utilitarianism has at least two distinct dimensions: it permits harming innocent individuals to maximize aggregate utility ('instrumental harm'), and it treats the interests of all individuals as equally important ('impartial beneficence') [21,33]. Indeed, preliminary evidence suggests these two dimensions characterize the way ordinary people think about moral dilemmas in the context of the COVID-19 pandemic [34]. These two dimensions of utilitarianism not only are psychologically distinct in the general public [33] but also have distinct impacts on perception of leaders. Specifically, when people endorse (versus reject) utilitarian principles in the domain of instrumental harm they are seen as *worse* political leaders, but in some cases are seen as *better* political leaders when they endorse utilitarian principles in the domain of impartial beneficence [37].

Another dilemma that pits utilitarian principles against other non-utilitarian principles – this time in the domain of impartial beneficence – is whether leaders should prioritize their own citizens over people in other countries when allocating scarce resources. The utilitarian sole focus on consequences mandates a strict form of impartiality: the mere fact that someone is one's friend (or their mother or fellow citizen) does not imply that they have any obligations to such a person that they do not have to any and all persons [48]. Faced with a decision about whether to help a friend (or family member or fellow citizen) or instead provide an equal or slightly larger benefit to a stranger, this strict utilitarian impartiality means that one cannot morally justify favouring the person closer to them. In contrast, many non-utilitarian approaches explicitly incorporate these notions of special obligations, recognizing the relationships between people as morally significant. Here, President Trump went against utilitarian principles when he ordered a major company developing personal protective equipment (PPE) to stop distributing it to other countries who needed it[49], or when he ordered the US government to buy up all the global stocks of the COVID-19 treatment remdesivir [50]. His actions generated outrage across the world and stood in contrast to statements from many other Western leaders at the time. The Prime Minister of the UK, Boris Johnson, for example, endorsed impartial beneficence when he argued for the imperative to "ensure that the world's poorest countries have the support they need to slow the spread of the virus" (June 3rd, 2020) [51]. In a similar vein, the Dutch government donated 50 million euros to the Coalition for Epidemic Preparedness Innovations, an organization that aims to distribute vaccines equally across the world [52].

In sum, public trust in leaders is likely to be a crucial determinant of successful pandemic response, and may depend in part on how leaders approach the many moral dilemmas that arise during a pandemic. Utilitarian responses to such dilemmas may erode or enhance trust relative to non-utilitarian approaches, depending on whether they concern instrumental harm or impartial beneficence. Past research on trust and utilitarianism is insufficient to understand how utilitarian resolutions to moral dilemmas influence trust during the COVID-19 pandemic – and future crises – for several reasons. First, it has relied on highly artificial moral dilemmas, such as the 'trolley problem' [53,54], that most people have not encountered in their daily lives. Thus, the findings of past studies may not generalize to the context of a global health crisis, where everyone around the world is directly impacted by the moral dilemmas that arise during a pandemic. Second, because the vast majority of previous work on trust in utilitarians has focused on instrumental harm, we know little about how impartial beneficence impacts trust. Third, most previous work on this topic has focused on trust in ordinary people. However, there is evidence that utilitarianism differentially impacts perceptions of ordinary people and leaders [37,38,40], which means we cannot generalize from past research on trust in utilitarians to a leadership context. Because leaders have power to resolve moral dilemmas through policymaking, and therefore can have far more impact on the outcomes of public health crises than ordinary people can, it is especially important to understand how leaders' approaches to moral dilemmas impact trust. Finally, past work on inferring trust from moral decisions has been conducted in just a handful of Western populations – in the United States, Belgium, and Germany – and so may not generalize to other countries that are also affected by the COVID-19 pandemic. We need, therefore, to assess cross-cultural stability by testing this hypothesis in different countries around the world. Indeed, given observations of cultural variation in the willingness to endorse sacrificial harm [32], it is not a foregone conclusion that utilitarian decisions will impact trust in leaders universally. For further details of how the present work advances our understanding of moral dilemmas and trust in leaders, see Supplementary Notes 3-5.

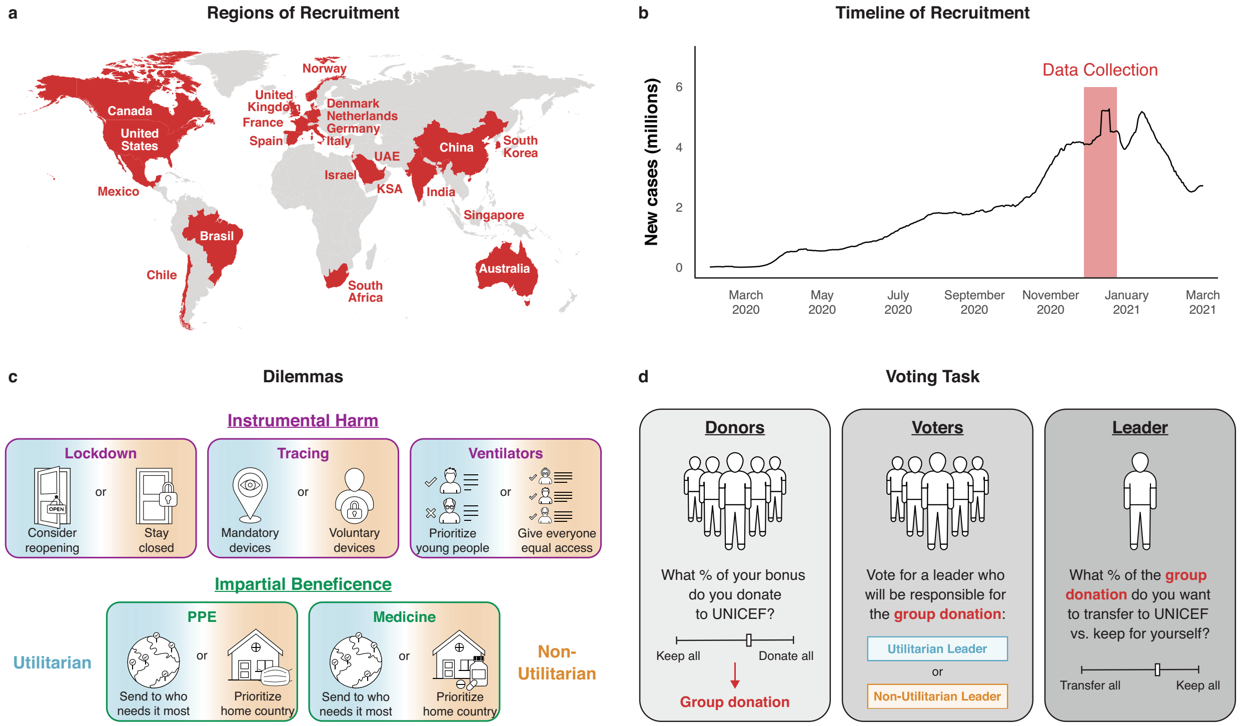


Figure 1. Overview of Experimental Methods. (A) Regions of recruitment for online samples broadly nationally representative with respect to age and gender. KSA, the Kingdom of Saudi Arabia. UAE, the United Arab Emirates. (B) Running 7-day average of new COVID-19 confirmed global infections from January 29th 2020 to March 14th 2021, with highlighted data collection window (red; from November 26th 2020 to December 22nd 2020). Number of COVID-19 confirmed infections were taken from the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University [71] (last update March 14th 2021). (C) Summary of the five COVID-19 dilemmas employed in the experimental tasks. (D) Voting task: participants were asked to vote for a leader who would later be entrusted with a group's charitable donation and be able to 'embezzle' some of the donation money for themselves.

The goal of the current research is to test the hypothesis that endorsement of instrumental harm would decrease trust in leaders while endorsement of impartial beneficence would increase trust in leaders, in the context of the COVID-19 pandemic. Testing this hypothesis across a diverse set of 22 countries spanning six continents (Figure 1A and Supplementary Figure 1) in November – December 2020, we aim to inform how leaders around the globe can communicate with their constituencies in ways that will preserve trust during global crises. Given the public health consequences of mistrust in leaders [7–9], if our hypothesis is confirmed, leaders may wish to carefully consider weighing in publicly on moral dilemmas that are unresolvable with policy, because their opinions might erode citizens' trust in other pronouncements that may be more pressing, such as advice to comply with public health guidelines.

To test our hypothesis empirically, we drew on case studies of public communications to identify five moral dilemmas that have been actively debated during the COVID-19 pandemic (Figure 1C). Three of these dilemmas involve instrumental harm: the *Ventilators* dilemma concerns whether younger individuals should be prioritized to receive intensive medical care over older individuals when medical resources such as ventilators are scarce [23,44], the *Lockdown* dilemma concerned whether to consider reopening schools and the economy or remain in lockdown [23,55], and the *Tracing* dilemma concerned whether it should be mandatory for residents to carry devices that continuously trace the wearer's movements, allowing the government to immediately identify people who have potentially been exposed to the coronavirus [45–47]. The other two dilemmas involved impartial beneficence: the *PPE* dilemma concerned whether PPE manufactured within a particular country should be reserved for that country's citizens under conditions of scarcity, or sent where it is most needed [23,56–58], and the *Medicine* dilemma concerned whether a novel COVID-19 treatment developed within a particular country should be delivered with priority to that country's citizens, or shared impartially around the world [56,59,60]. Participants in our studies read about leaders who endorsed either utilitarian or non-utilitarian solutions to the dilemmas (see Table 1), and subsequently completed behavioral and self-report measures of trust in the respective leaders (see Extended Data Figure 1). For example, some read about a leader who endorsed prioritizing younger over older people for scarce ventilators, and were then asked how much they trusted that leader. While there are many similar dilemmas potentially relevant to the COVID-19 crisis, we chose to focus on the five described above because they (1) have been publicly debated at time of writing; and (2) apply to all countries in our planned sample. For further details of why we chose these specific dilemmas and how they can test our theoretical predictions, see Supplementary Notes 2 and 6-9.

Table 1. Summary of Moral Arguments in COVID-19 Dilemmas

Dilemma	**Argument Type**	
Instrumental Harm (IH)	**Utilitarian**	**Non-Utilitarian**
Lockdown	"We need to think about all the consequences. Preventing deaths from COVID isn't all that matters, and continuing these prolonged restrictions will have a far worse effect on our overall well-being."	"As leaders, our primary duty is to protect our citizens. We must think of our responsibilities to one another, and we cannot sacrifice some of our most vulnerable people in pursuit of the greater good."
Ventilators	"We have to think about how we can do the most good with the resources we have, and that means prioritizing those people who have the best	"It's not our place to choose who lives. Everyone has the same right to receive equal access to treatment, and we cannot abandon our most

	chance of recovering and living a long and healthy life.”	vulnerable in an effort to save more lives.”
Tracing	“We need to control the pandemic, and sometimes you have to sacrifice the right to privacy for the greater good.”	“Everyone has a right to privacy, and we cannot sacrifice this right in an effort to control the pandemic.”
Impartial Beneficence (IB)	**Utilitarian**	**Non-Utilitarian**
Medicine	“COVID-19 is a global pandemic that affects all humans equally. We need to be impartial and send treatment where it can achieve the greatest good.”	“We have a right to use our own resources to help our own citizens before everyone else. Other countries can produce their own treatments for COVID-19.”
Personal Protective Equipment	“COVID-19 is a global pandemic that affects all humans. We need to be impartial in how we distribute resources like PPE and send it where it can achieve the greatest good.”	“We have a duty to protect our own citizens first, not everyone in the world. Other countries are responsible for protecting their own citizens from COVID-19.”

We measured trust in two complementary ways. First, we asked participants to self-report their general trust in the leaders, in terms of both an overall character judgment (“How trustworthy do you think this person is?”) and how likely they would be to trust this person on other issues not related to the dilemma (“How likely would you be to trust this person’s advice on other issues?”). Second, we used a novel, incentivized *voting task* designed to measure public trust in leaders (Figure 1D). Following past work, we define leaders as people who are responsible for making decisions on behalf of a group [61,62]. In the voting task, participants were invited to cast a vote to appoint a leader who would be responsible for making a charitable donation on behalf of a group. Crucially, the leader had the opportunity to ‘embezzle’ some of the donation money for themselves. Participants were asked to vote for either a person who endorsed a utilitarian or a non-utilitarian position on a COVID-19 dilemma; the person who received the most votes would have control over the group’s donation. By measuring preferences for a leader who was responsible for a group’s donations to help those in need, the voting task captures trust in leaders in a specific context that is highly relevant to our central research question: during a health crisis, effective leadership requires responsible stewardship of public resources to help those in need. For further details of why we designed our trust measures in this way, see Supplementary Notes 10-12.

Our analyses therefore tested two complementary hypotheses. First, we predicted that self-reported trust would be lower for leaders who endorse utilitarian over non-utilitarian approaches to dilemmas involving instrumental harm, while the reverse pattern would be observed for impartial beneficence, with greater trust for leaders who endorse utilitarian approaches to dilemmas involving impartial beneficence (hypothesis 1). Second, we predicted that participants would be less likely to vote for leaders who endorse utilitarian over non-utilitarian views on dilemmas involving instrumental harm, while the reverse pattern would be observed for dilemmas involving impartial beneficence (hypothesis 2). Pilot studies conducted in the United States and the United Kingdom in July 2020 provided initial support for these hypotheses (see Pilot Data in Supplementary Information and Supplementary Figures 2-6 for details). All analyses controlled for participants' demographics and own policy preferences in each dilemma (see Table 2).

Finally, we note that the framing of both the self-report and behavioral measures of trust are deliberately unrelated to the pandemic dilemmas we use to highlight the moral commitments of the leader. This crucial design choice allowed us to measure the impact of utilitarian versus non-utilitarian endorsements of pandemic dilemmas on *subsequent trust* in leaders. In this way, the current design illuminates an important real-life question: if a leader weighs in publicly on a moral dilemma during a crisis, how likely are they to be trusted later on other matters of public concern?

Table 2. Design Table

Question	Hypothesis	Sampling Plan	Analysis Plan	Interpretation
How do leaders' endorsements of utilitarian versus non-utilitarian approaches to dilemmas affect self-reported trust?	Participants will report higher trust in leaders who reject (versus endorse) instrumental harm, and higher trust in leaders who endorse (versus reject) impartial beneficence.	Power analyses suggested that our planned sample size (N = 12,600) would be sufficient to achieve 95% power to detect an effect size of *d* = 0.05.	We will conduct a linear mixed-effects model of the effect of argument type, dimension type and their interaction on the composite measure of self-reported trust, controlling for demographic variables and participants' own policy preferences.	A significant interaction (after following up with post hoc tests) will be interpreted as evidence that endorsement of instrumental harm decreases general trust in leaders while endorsement of impartial beneficence increases general trust in leaders. Should the observed effect size of the interaction be statistically equivalent (with the larger of the two p values in the TOST using equivalence bounds set by the SESOI smaller than alpha = .05), this will be interpreted as evidence for null effects.

How do leaders' endorsements of utilitarian versus non-utilitarian approaches to dilemmas affect trusting behavior?	Participants will be more likely to vote for a leader who rejects (versus endorses) instrumental harm, while they will be more likely to vote for a leader who endorses (versus rejects) impartial beneficence.	Power analyses suggested that our planned sample size (N = 12,600) would be sufficient to achieve 95% power to detect an odds ratio (*OR)* of 1.30.	We will conduct a generalized linear mixed-effects model with the logit link of the effect of dimension type on leader choice, controlling for demographic variables and participants' own policy preferences.	A significant main effect of dimension will be interpreted as evidence that endorsement of instrumental harm decreases trusting behavior toward leaders, while endorsement of impartial beneficence increases trusting behavior toward leaders. Should the observed effect size be statistically equivalent (with the larger of the two p values in the TOST using equivalence bounds set by the SESOI smaller than alpha = .05), this will be interpreted as evidence for null effects.

Results

Analysed dataset

Donations task. A few days prior to running the main experiment, we recruited a convenience sample of donor participants (total N = 100; 58 women, 40 men, 2 with another gender identity; mean age 33.95 years) in the United States via Prolific (www.prolific.co). The donor participants chose to contribute a total of US$87.89 to the United Nations Children's Fund (UNICEF). We displayed this amount to voter participants in the main experiment.

Participants. Following the preregistered sampling plan (see Methods), we recruited participants via several online survey platforms from November 26th 2020 to December 22nd 2020, as new cases of COVID-19 in 2020 were peaking globally (Figure 1B). In total, we recruited a sample of 24,809 participants across the following countries: Australia, Brazil, Canada, Chile, China, Denmark, France, Germany, India, Israel, Italy, the Kingdom of Saudi Arabia, Mexico, the Netherlands, Norway, Singapore, South Africa, South Korea, Spain, the United Arab Emirates, the United Kingdom, and the United States (Figure 1A and Supplementary Tables 1 and 2).

As specified in our preregistered sampling plan (see Methods), participants who did not pass the attention checks were screened out immediately prior to beginning the survey, but due to platform and institutional review board requirements, participants in the United States and the United Kingdom were able to complete the survey even if they failed such checks, and so they were excluded post hoc, after data collection (N = 101 for attention check 1, N = 118 for attention check 2). In addition, participants were excluded according to our exclusion criteria if they (1) took the survey more than once (N = 565); (2) reported living in a country different from that of intended recruitment (N = 96, of which 4 did not answer the question); or (3) failed to answer more than 50% of the questions (N = 0). The sample size after applying these exclusion criteria was 23,929; we then excluded participants from specific analyses if they (4) did not provide a response for one of our main dependent variables (N = 177 for self-report, N = 201 for voting); or (5) failed the comprehension check for the task being analyzed (see Design; N = 6,161 for self-report, N = 11,090 for voting). This resulted in a total final sample of N = 17,591 for the self-report task, and N = 12,638 for the voting task. Crucially, the comprehension check failure rates were balanced across experimental conditions for each task (failure rate for self-report task comprehension check: 25.30% after Instrumental Harm dilemmas, Utilitarian argument [final N = 4,499]; 26.08% after Instrumental Harm, Non-Utilitarian argument [final N = 4,299]; 25.25% after Impartial Beneficence, Utilitarian argument [final N = 4,461]; 27.13% after Impartial Beneficence, Non-Utilitarian argument [final N = 4,332]; fail rate for voting task comprehension check: 46.46% after Instrumental Harm dilemmas [final N = 6,373]; 47.02% after Impartial Beneficence dilemmas [final N = 6,265]).

Representativeness. As stated in the Stage 1 Report, while we aimed to recruit samples broadly representative for age and gender in all countries, we anticipated that it would be difficult to obtain fully representative quotas in all countries for some demographic categories. To evaluate the representativeness of our samples across age and gender categories, we examined the differences between our targeted quotas (based on available published population characteristics) and actual quotas in the data, separately for each country. We achieved broadly representative samples for gender, with most differences between the observed and targeted proportions being less than or equal to 5% in all but two countries (Singapore and the United Arab Emirates). Note that, because available population data across countries primarily report binary gender categories, our estimates of representativeness were not able to account for those identifying as non-binary, which is a limitation. Similarly, in 15 countries we obtained broadly representative samples for age, with the difference between targeted and actual proportions being less than or equal to 5%. In six countries (the Kingdom of Saudi Arabia, Singapore, South Korea, the United Arab Emirates, the United Kingdom, and the United States), older participants were underrepresented in our sample by 6-15%. In one country (Germany), older participants were overrepresented by 6% (for details, see Supplementary Results; for figures depicting expected versus obtained counts in each gender and age category, see Supplementary Figures 7-8).

Main Analyses

The main results are depicted in Figures 2-3, across both the self-report and behavioral measures, respectively. As predicted, participants showed *more* trust in leaders who endorsed utilitarian views in impartial beneficence dilemmas and *less* trust in leaders who endorsed utilitarian views in instrumental harm dilemmas. This pattern of results was observed for each dilemma (Figures 2B and 3C) and was robust across countries (Figures 4A and 4B). Following our preregistered analysis plan (see Analysis Plan for Hypothesis Testing), we examined self-report and behavioral measures of trust in two separate models, with results passing a corrected alpha of p ≤ .005 being interpreted as *'supportive* evidence' for our hypotheses, and results passing a corrected alpha of p < .05 being interpreted as *'suggestive* evidence' (all the CIs reported below are 97.5%).

Hypothesis 1: Self-reported trust. To examine participants' self-reported trust in the leaders, we fitted a linear mixed-effects model of the effect of argument type (Utilitarian vs. Non-Utilitarian), dimension type (Instrumental Harm vs. Impartial Beneficence), and their interaction, on the composite score of trust, adding demographic variables (gender, age, education, subjective socio-economic status (SES), political ideology, and religiosity) and policy support as fixed effects, and dilemmas and countries as random intercepts, with participants nested within countries (for details, see Analysis Plan for Hypothesis Testing). As specified in Analysis Plan for Hypothesis Testing, we also ran a model that included countries as random slopes of the two main effects and the interactive effect; the results were consistent with the simpler model, but due to convergence issues with the more complex model, we report the simpler model.

We observed a significant main effect of argument type (*B* = -0.53, SE = 0.02, *t*(17,562) = -24.81, *p* < .001, CI = [-0.58, -0.48]), no significant main effect of dimension type (*B* = 0.10, SE = 0.10, *t*(3) = 0.95, *p* = .408, CI = [-0.15, 0.35]) and, crucially, a significant interaction between argument and dimension type (*B* = 2.12, SE = 0.04, *t*(17,558) = 49.44, *p* < .001, CI = [2.03, 2.22]). Post hoc comparisons with Bonferroni corrections confirmed that, in Instrumental Harm dilemmas, utilitarian leaders were seen as less trustworthy than non-utilitarian leaders (mean trust for utilitarian leaders = 3.35, SE = 0.09, CI = [3.05, 3.65]; mean trust for non-utilitarian leaders = 4.95, SE = 0.09, CI = [4.64, 5.25]; *B* = -1.60, SE = 0.03, *t*(17,559) = -52.51, *p* < .001, CI = [-1.66, -1.53]), but in Impartial Beneficence dilemmas this effect was reversed, such that utilitarian leaders were seen as more trustworthy than non-utilitarian leaders (mean trust for utilitarian leaders = 4.51, SE = 0.10, CI = [4.14, 4.88]; mean trust for non-utilitarian leaders = 3.98, SE = 0.10, CI = [3.61, 4.35]; *B* = 0.53, SE = 0.03, *t*(17,560) = 17.41, *p* < .001, CI = [0.46, 0.60]; see Figure 2A; for results by dilemma, see Figure 2B; for results by country, see Figure 4A).

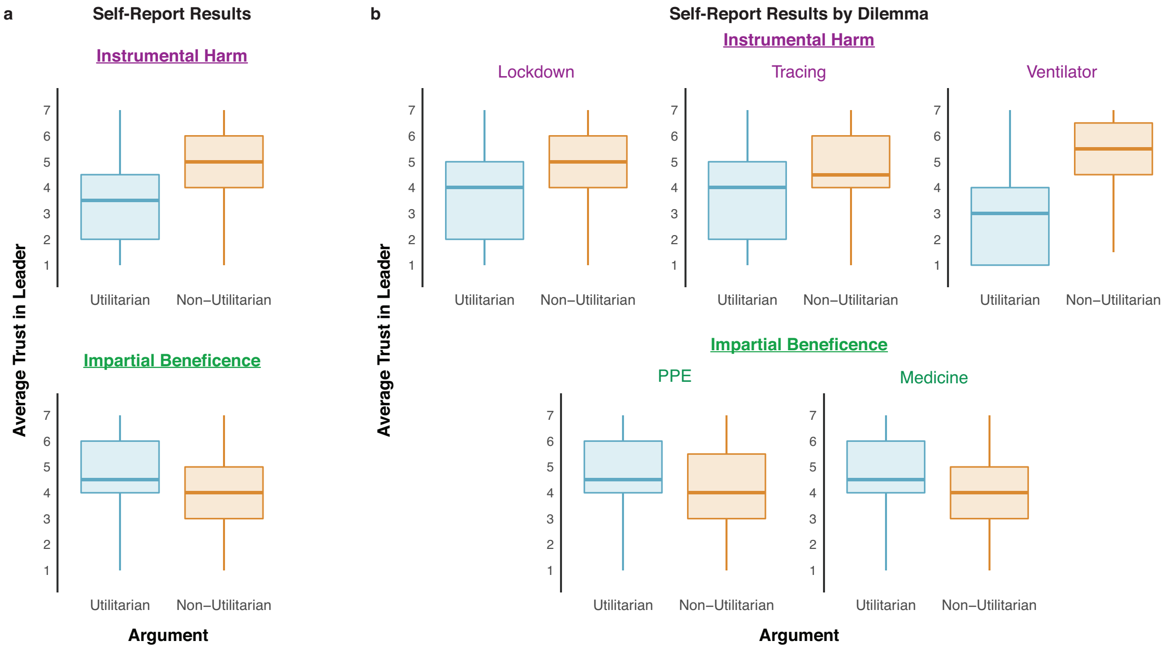


***Figure 2**. **Self-Reported Trust in Utilitarian and Non-Utilitarian Leaders**.* Average trust in utilitarian versus non-utilitarian leaders, (A) with results collapsed across Instrumental Harm and Impartial Beneficence dilemmas and (B) separately for each of the Instrumental Harm dilemmas (Lockdown, Tracing, and Ventilators) and Impartial Beneficence dilemmas (PPE and Medicine) in the self-report task (N = 17,591). Non-utilitarian leaders were seen as more trustworthy than utilitarian leaders for Instrumental Harm dilemmas, while the reverse was observed for Impartial Beneficence dilemmas. Bars correspond to median scores, lower and upper hinges correspond to the first and third quartiles, respectively, and whiskers ends correspond to the most extreme data points within 1.5 times the interquartile range.

Hypothesis 2: Voting measure. To examine participants' trust in the leaders as demonstrated by their voting behavior, we fitted a generalized linear mixed-effects model with the logit link of the effect of dimension type (Instrumental Harm vs. Impartial Beneficence) on leader choice in the voting task (Utilitarian vs. Non-Utilitarian), adding demographic variables (gender, age, education, subjective SES, political ideology, and religiosity) and policy support as fixed effects, and dilemmas and countries as random intercepts, with participants nested within countries (for details, see Analysis Plan for Hypothesis Testing). This yielded a singular fit, so following our analysis plan, we reduced the complexity of the random-effects structure by only including dilemmas and countries as random intercepts. As specified in Analysis Plan for Hypothesis Testing, we also ran a model that included countries as random slopes of the effect of dimension type; the results were consistent with the simpler model, but due to singularity issues (both with and without participants nested within countries), we report the simpler model.

We observed a significant main effect for dimension type (*B* = 1.37, SE = 0.32, *z* = 4.21, *p* < .001, CI = [0.41, 2.33], odds ratio [OR] = 3.93) such that participants were almost 4 times more likely to choose the utilitarian leader in Impartial Beneficence dilemmas compared with Instrumental Harm dilemmas. Post hoc comparisons with Bonferroni corrections confirmed that in Instrumental Harm dilemmas, participants were less likely to vote for utilitarian leaders than non-utilitarian leaders (probability of choosing utilitarian leader = 0.21, SE = 0.04, CI = [0.13, 0.31]), but in Impartial Beneficence dilemmas this effect vanished (probability of choosing utilitarian leader = 0.50, SE = 0.07, CI = [0.34, 0.67]; see Figure 3A; for model estimates, see Figure 3B; for results by dilemma, see Figure 3C; for results by country, see Figure 4B).

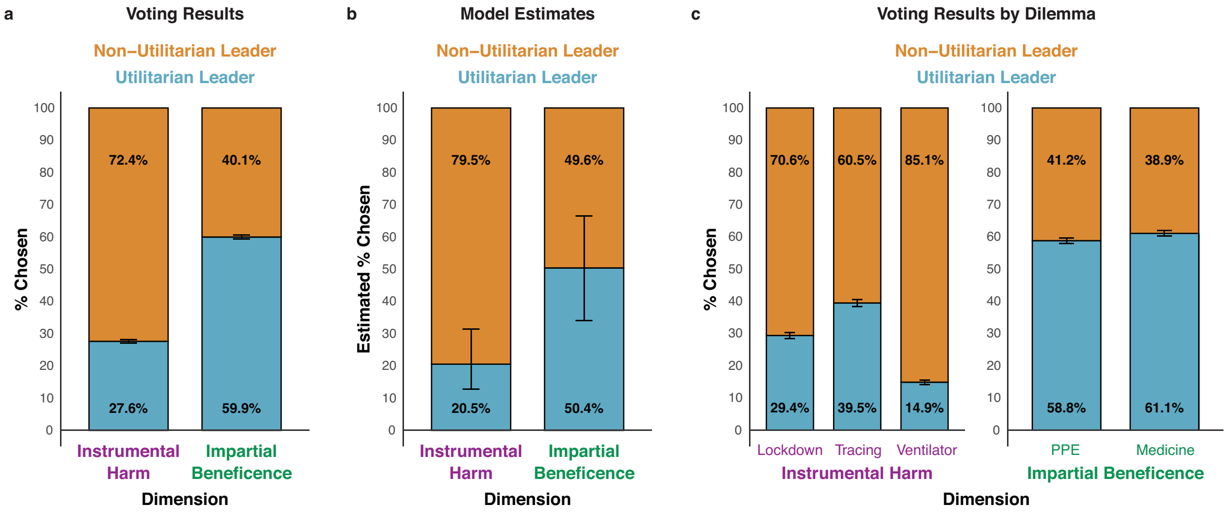


Figure 3*. *Voting Choices for Utilitarian and Non-Utilitarian Leaders. (A) Percentage of participants who chose to trust utilitarian versus non-utilitarian leaders, separately for Instrumental Harm and Impartial Beneficence dilemmas in the voting task (N = 12,638). (B) Choices for utilitarian versus non-utilitarian leaders as estimated from a logit model including demographic variables (gender, age, education, subjective SES, political ideology, and religiosity) and policy support as covariates, and dilemmas and countries as random intercepts (for details, see Hypothesis 2: Voting measure). (C) Percentage of participants who chose to trust utilitarian versus non-utilitarian leaders, separately for each of the Instrumental Harm dilemmas (Lockdown, Tracing, and Ventilators) and Impartial Beneficence dilemmas (PPE and Medicine). Non-utilitarian leaders were more likely to be voted in Instrumental Harm dilemmas, but not in Impartial Beneficence dilemmas. Error bars represent standard error of the percentages (A) and (C), and the 97.5% CIs of the model estimates (B).

Based on suggestions that logit and linear models should converge and that linear models can in some cases be preferable [63,64], we had also preregistered the same analysis using a linear model (instead of a model with the logit link) with the identical fixed- and random-effects structures. However, the linear model yielded non-significant results for the main effect of dimension type with our Bonferroni-corrected alpha (*B* = 0.18, SE = 0.05, *t*(3) = 3.73, *p* = .034, CI = [0.07, 0.30]; probability of choosing utilitarian leader in Instrumental Harm dilemmas = 0.30, SE = 0.03, CI = [0.16, 0.45], in Impartial Beneficence dilemmas = 0.49, SE = 0.04, CI = [0.31, 0.67]). This discrepancy was unusual, since binomial and linear approaches most often give converging results [65,66]. Following our preregistered analysis plan, we followed up on this non-significant result using the two one-sided tests (TOST) procedure to differentiate between insensitive versus null results. Given the equivalence bounds set by our smallest effect size of interest (SESOI) (Δ_L = -0.15 and Δ_U = 0.15; see Power Analysis section), the effect of dimension on leader choice (a 32% difference) was statistically not equivalent to zero (*z* = 20.77, *p* = 1.000 for the test with Δ_U). This analysis, however, does not take into account the covariates specified in the models.

To resolve the discrepancy between our preregistered binomial and linear models, we ran a number of additional exploratory models. These are described in Exploratory Analyses and summarized in Table 3.

Table 3. Results for Voting Task Models. Voting task results obtained with the preregistered models, as well as a variety of exploratory models, including binomial and linear models (‘Model Type’) with different specifications of fixed (‘Fixed Effects’) and random effects (‘Random Effects’). ‘B’ indicates model coefficients of the effect of moral dimension on voting preference, ‘p’ indicates significance value, and ‘Predicted Difference’ indicates the expected difference in votes for utilitarian leaders in Impartial Beneficence versus Instrumental Harm dilemmas.

Model Type	Fixed Effects	Random Effects	B	p	Predicted Difference
		Preregistered Models			
Binomial	Six demographics Support Dimension	Country Dilemma	1.37	<.001	0.32
Linear	Six demographics Support Dimension	Country Dilemma	0.18	.034	0.18
		Exploratory Models			
Binomial	Dimension	None	1.37	<.001	0.32

Linear	Dimension	None	0.32	<.001	0.32
Binomial	Six demographics Support Dimension	None	1.34	<.001	0.31
Linear	Six demographics Support Dimension	None	0.19	<.001	0.19
Linear with Robust SEs	Six demographics Support Dimension	Country	0.19	<.001	0.19
Binomial	Six demographics Support Country Dimension	None	1.34	<.001	0.32
Linear	Six demographics Support Country Dimension	None	0.19	<.001	0.19

Robustness checks. Following our analysis plan, we verified the robustness of our findings in several ways. First, due to the changes in country-specific lockdown policies that were implemented between preregistration and data collection, we ran a variation of our models which omitted the Lockdown dilemma. The results were substantially unchanged, both for the self-report task (interaction between argument and dimension type: *B* = 2.26, SE = 0.05, *t*(17,640) = 48.56, *p* < .001, CI = [2.16, 2.37]) and the voting task (main effect for dimension type in binomial model: *B* = 1.29, SE = 0.39, *z* = 3.33, *p* < .001, CI = [0.06, 2.52], OR = 3.63) tasks.

In addition, because some countries had already implemented mandatory contact tracing schemes at the time of data collection, we ran a variation of our models in those countries only (namely China, India, Israel, Singapore, and South Korea) with and without the Tracing dilemma. The results in those countries were similar when including and omitting the Tracing dilemma from the analysis, both for the self-report task (Tracing included: interaction between argument and dimension type: *B* = 1.13, SE = 0.10, *t*(3,267) = 11.62, *p* < .001, CI = [0.91, 1.35]; Tracing excluded: interaction between argument and dimension type: *B* = 1.55, SE = 0.10, *t*(3,266) = 14.86, *p* < .001, CI = [1.32, 1.78]) and voting task (Tracing included: main effect for dimension type in binomial model: *B* = 0.98, SE = 0.36, *z* = 2.70, *p* = .007, CI = [-0.09, 2.07], OR = 2.67; Tracing excluded: main effect for dimension type in binomial model: *B* = 1.32, SE = 0.14, *z* = 9.26, *p* < .001, CI = [0.88, 1.78], OR = 3.74). Finally, we also checked that the results in these countries were robust to order effects (i.e., regardless of whether participants had seen the tracing dilemma prior to other dilemmas). To do this, we analyzed participants' responses with an additional covariate indicating whether the participant had seen the tracing dilemma in the prior task. Again, the results were substantially unchanged both for the self-report task (interaction between argument and dimension type: $B = 1.13$, SE = 0.10, $t(3{,}266) = 11.62$, $p < .001$, CI = [0.91, 1.35]) and the voting task (main effect for dimension type in binomial model: $B = 1.11$, SE = 0.37, $z = 3.01$, $p = .003$, CI = [0.03, 2.20], OR = 3.03).

Exploratory Analyses

Additional models for voting task. As noted above, our main preregistered analysis for the voting task was a generalized linear mixed-effects model with the logit link of the effect of dimension type (Instrumental Harm vs. Impartial Beneficence) on the leader choice (Utilitarian vs. Non-Utilitarian), with demographics and participants' own policy preferences as fixed effects, and dilemmas and countries as random intercepts (Table 2). This analysis confirmed our predictions, but we had also preregistered the same analysis using a linear model (instead of logit link) with the identical fixed and random effects structure. As described above, the results from this model did not pass our preregistered Bonferroni-corrected significance threshold. This discrepancy was unusual, given prior reports that linear and binomial models yield identical results in the vast majority of cases [63,66]. As a first check on this discrepancy, we assessed the fits of the binomial and linear models by fitting each with half the data, and predicting the leader choices in the remaining half. The mean difference between the predicted and observed values was lower in the binomial model (mean error = 0.25) compared with the linear model (mean error = 0.27; $t(6{,}318) = -32.53$, $p < .001$), suggesting the binomial model is a better fit to our data.

Next, we ran a series of follow-up analyses to supplement our preregistered, theoretically informed models. There are a variety of opinions for how to best level complex nested binary data like ours. For example, while random effects aid generalizability [67], some advocate for modeling country variables as fixed rather than random effects to prevent increases in model bias [68,69] or overly complex random-effects structures [70]. Moreover, while controlling for demographic variables is important for generalizability of our findings, some advocate for minimal use of covariates to prevent type 1 error inflation [71]. Due to the discrepancy in the theoretically justified models we had preregistered, and ongoing debates over the specifications of modelling such complex data, we ran a variety of models (described in detail in the Supplementary Results and summarized in Table 3) with different link functions and different specifications of fixed and random effects, as well as robust random effects and randomization inference. Overall, all models led to the same conclusion: participants voted for the non-utilitarian leader more than the utilitarian leader in dilemmas about instrumental harm, but the reverse in impartial beneficence dilemmas, with the utilitarian leader trusted more than the non-utilitarian leader – suggesting that the discrepancy between our preregistered binomial and linear models was due to an overly complex random-effects structure.

Effects by country. To explore cross-cultural variation in trust in utilitarian versus non-utilitarian leaders, we ran additional models with country as a random slope and extracted the coefficients of interest (Figures 4A and 4B). For the self-report task, we conducted a linear mixed-effects model of the effect of argument type (Utilitarian vs. Non-Utilitarian), dimension type (Instrumental Harm vs. Impartial Beneficence), and their interaction, on the composite score of trust, adding demographic variables (gender, age, education, subjective SES, political ideology, and religiosity) and policy support as fixed effects, and countries as a random slope of the interactive effect of argument and dimension. First, we confirmed that there was a significant interaction between argument and dimension type (B = 2.08, SE = 0.16, t(21) = 13.08, p < .001, CI = [1.71, 2.45]), consistent with our preregistered model. Next, we extracted the interaction coefficients for each country, as well as the standard errors of the coefficients, with the estimates plotted in Figure 4A. While there were some variations in the effect sizes, the results were remarkably consistent across countries. The predicted pattern of results was observed in all 22 countries with Israel, South Korea, and China showing the smallest effects, and Brazil, the UAE, and Norway showing the largest effects.

For the voting task, we conducted a generalized linear mixed-effects model with the logit link of the effect of dimension type (Instrumental Harm versus Impartial Beneficence) on leader choice (Utilitarian versus Non-Utilitarian), adding demographic variables (gender, age, education, subjective SES, political ideology, and religiosity) and policy support as fixed effects, and countries as a random slope of dimension. First, we confirmed there was a significant main effect for dimension type (B = 1.34, SE = 0.07, z = 17.88, p < .001, CI = [1.16, 1.51], OR = 3.81), as in our preregistered model. Next, we extracted the coefficients for each country, as well as the standard errors of the coefficients, and exponentiated them to get the odds ratios, with the resulting estimates plotted in Figure 4B. Again, the results were remarkably consistent with the predicted pattern of results seen across all 22 countries with China, Israel, and Canada showing the smallest effects and Norway, the UAE, and the United States showing the largest effect size.

Correlations between self-report and behavioral measures across countries. The self-report and behavioral tasks employed in the current study are highly complementary in several ways: for example, the former is more generalizable across different situations, while the latter is incentivized and more concrete (see Supplementary Note 10 for further details). To ensure that despite their superficial differences the tasks targeted the same construct, i.e. trust in leaders, and measured robust preferences across countries, we checked that the effects of moral arguments and utilitarian dimensions on these measures were correlated across countries. Indeed, we found that the coefficients of the interaction between moral argument and moral dimension on trust in the self-report task were significantly correlated with the effect of moral dimension on leader choice in the voting task (r = 0.76, p < .001; see Figure 4C).

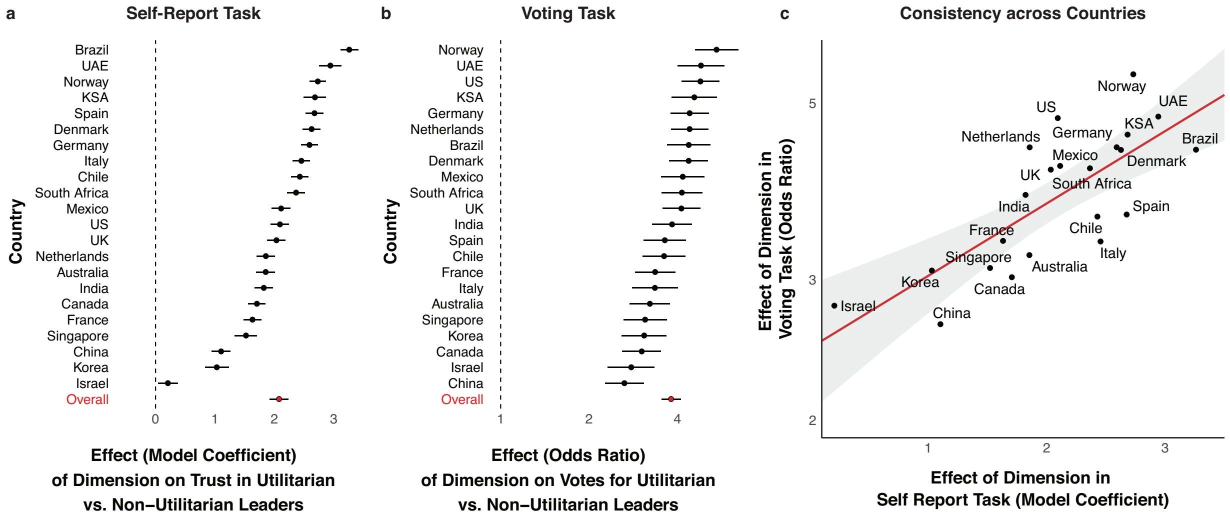


Figure 4. Trust in Leaders by Country as Measured by the Self-Report and Voting Tasks. (A) Predicted effect of moral dimension (Instrumental Harm versus Impartial Beneficence) and argument (Utilitarian versus Non-Utilitarian) on trust in the self-report task (N = 17,591) for each country and overall. Dots represent model coefficients extracted from a model including country as a random slope of the interactive effect of moral dimension and argument (see Exploratory Analyses); error bars represent standard errors of the model coefficients. (B) Odds ratio of the effect of moral dimension (Instrumental Harm versus Impartial Beneficence) on trust for the utilitarian versus non-utilitarian leader in the voting task (N = 12,638) for each country and overall. Dots represent odds ratios extracted from a model including country as a random slope of moral dimension (see Exploratory Analyses); error bars represent exponentiated standard errors of the model coefficients. (C) Correlation between the country-level effect size estimates in the self-report task (x axis; also depicted in [A] and voting task (y axis; also depicted in [B]). US, the United States; UK, the United Kingdom; UAE, the United Arab Emirates; KSA, the Kingdom of Saudi Arabia.

Effects of participant exclusions in voting task. The main analyses reported above were performed on a subset of participants who passed the comprehension checks, as per our preregistered sampling plan (criterion 5; see Sampling Plan). For the voting task, the observed pass rate (53.26%) was lower than the preregistered expected pass rate (60%), suggesting that the comprehension check may have been overly stringent. Therefore, we conducted additional analyses to explore whether this preregistered exclusion criterion might have affected the generalizability of our results across the study population in terms of education level.

Participants who failed the voting task comprehension check reported slightly lower educational attainment on average (mean = 5.32, SE = 1.39, CI = [5.30, 5.35]) than those who passed the comprehension check (mean = 5.42, SE = 1.37, CI = [5.40, 5.45]; t(23,224) = 5.51, p < .001, d = 0.07). However, we observed similar results in our preregistered models when including participants who failed the voting task comprehension check (main effect for dimension type in binomial model: $B = 1.26$, SE = 0.28, $z = 4.55$, $p < .001$, CI = [0.44, 2.08], OR = 3.53; main effect for dimension type in linear model: $B = 0.17$, SE = 0.04, $t(3) = 4.11$, $p = .026$, CI = [0.07, 0.27]).

Discussion

The COVID-19 pandemic has raised a number of moral dilemmas that engender conflicts between utilitarian and non-utilitarian ethical principles. Building on past work on utilitarianism and trust, we tested the hypothesis that endorsement of utilitarian solutions to pandemic dilemmas would impact trust in leaders. Specifically, in line with suggestions from previous work and case studies of public communications during the early stages of the pandemic, we predicted that endorsing instrumental harm would decrease trust in leaders, while endorsing impartial beneficence would increase trust. Experiments conducted during November – December 2020 in 22 countries across six continents (total N = 23,929; valid sample for self-report task = 17,591; valid sample for behavioral task = 12,638) provided robust support for our hypothesis. In the context of five realistic pandemic dilemmas, participants reported *lower* trust in leaders who endorsed instrumental sacrifices for the greater good, and *higher* trust in leaders who advocated for impartially maximizing the welfare of everyone equally. In a behavioral measure of trust, only 28% of participants preferred to vote for a utilitarian leader who endorsed instrumental harm, while 60% voted for an impartially beneficent utilitarian leader. These findings were robust to controlling for a variety of demographic characteristics as well as participants' own policy preferences regarding the dilemmas. Although we observed some variation in effect sizes across the countries we sampled, the overall pattern of results was highly robust across countries. Our results suggest that endorsing utilitarian approaches to moral dilemmas can both erode and enhance trust in leaders across the globe, depending on the *type* of utilitarian morality.

We designed our set of dilemmas to rule out several alternative explanations for our findings, such as a general preference for less restrictive leaders (see Supplementary Note 7), leaders who treat everyone equally (see Supplementary Note 8), and leaders who seek to minimize COVID-19 related deaths (see Supplementary Note 9). In addition, all of our results survived planned robustness checks to account for the possibility that local policies related to lockdowns or contact tracing could bias participants' responses. Post hoc analyses demonstrated that our findings were highly consistent across the different dilemmas for instrumental harm (Lockdown, Tracing, and Ventilators) and impartial beneficence (Medicine and PPE).

While the robustness of our findings across countries speaks to their broad cultural generalizability, further work is needed to understand the observed variations in effect sizes across countries. It seems plausible that both economic (for example, gross domestic product or socio-economic inequality) and cultural (for example, social network structure) differences across countries could explain some of the observed variations. One possibility, for example, is that country-level variations in tightness-looseness [72], which have been associated with countries' success in limiting cases in the COVID-19 pandemic [73], might moderate the effects of moral arguments on trust in leaders. Another direction for future research could be to explore how country-level social network structure might influence our results. Individuals in countries with a higher kinship index [74] and a more family-oriented social network structure, for example, might be less likely to trust utilitarian leaders, especially when the utilitarian solution conflicts with more local moral obligations.

There are several important limitations to the generalizability of our findings. First, although our samples were broadly nationally representative for age and gender (with some exceptions; see Results), we did not assess representativeness of our samples on a number of other factors including education, income, and geographic location. Second, while our results do concord with the limited existing research examining the effects of endorsing instrumental harm and impartial beneficence on perceived suitability as a leader [37], and held across different examples of our pandemic-specific dilemmas, it of course remains possible that different results would be seen when judging leaders' responses in other types of crises (for example, violent conflicts, natural disasters, or economic crises), or at different stages of a crisis (for example, at the beginning versus later stages). Third, the reported experiments tested how responses to moral dilemmas influenced trust in anonymous, hypothetical political leaders. In the real world, however, people form and update impressions of known leaders with a history of political opinions and behaviors, and it is plausible that inferences of trustworthiness depend not just on a leader's recent decisions but also on their history of behavior, just as classic work on impression formation shows that the same information can lead to different impressions depending on prior knowledge about the target person [75]. Furthermore, we did not specify the gender of the leaders in our experiments (except in the voting task for China and for the Hebrew and Arabic translations, where it is not possible to indicate 'leader' without including a gendered pronoun; here it was translated in the masculine form). Past work conducted in the United States suggests that participants may default to an assumption that the leader is a man [76], but it will be important for future work to assess whether men and women leaders are judged differentially for their moral decisions. Because women are typically stereotyped as being warmer and more communal than men [77], it is plausible that women leaders would face more backlash for making 'cold' utilitarian decisions, especially in the domain of instrumental harm. Fourth, because the current work focused on trust in political leaders, it remains unclear how utilitarianism would impact trust in people who occupy other social roles, such as medical workers or ordinary citizens. Fifth, and finally, it could be interesting to explore further the connection between impartial beneficence and intergroup psychology, especially with regards to teasing apart 'impartiality' and 'beneficence'. For example, even holding beneficence constant, a leader who advocates for impartially sharing resources with a rival country may be perceived differently from one who impartially shares with an allied country (and, while speculative, this distinction might explain why Israel was an outlier in impartial beneficence, being a country in a region with ongoing local conflicts).

Our results have clear implications for how leaders' responses to moral dilemmas can impact how they are trusted. In times of global crisis, such as the COVID-19 pandemic, leaders will necessarily face real, urgent, and serious dilemmas. Faced with such dilemmas, decisions have to be made – and our findings suggest that how leaders make these judgments can have important consequences, not just for whether they are trusted on the issue in question but also more generally. Importantly, this will be the case even when the leader has little direct control over the resolution. While a national leader (for example, a president or prime minister) has the power and responsibility to resolve some moral dilemmas with policy decisions, not all political leaders (e.g., as in our study, local mayors) have that power. A leader with little ability to directly impact the resolution of a moral dilemma might consider that voicing an opinion on that dilemma could reduce their credibility on other issues that they have more power to control.

To conclude, we investigated how trust in leaders is sensitive to how they resolve conflicts between utilitarian and non-utilitarian ethical principles in moral dilemmas during a global pandemic. Our results provide robust evidence that utilitarian responses to dilemmas can both erode and enhance trust in leaders: advocating for sacrificing some people to save many others (i.e., instrumental harm) reduces trust, while arguing that we ought to impartially maximize the welfare of everyone equally (i.e., impartial beneficence) increases trust. Our work advances understanding of trust in political leaders, and shows that across a variety of cultures it depends not just on whether they make moral decisions, but also which specific moral principles they endorse.

Methods

Ethics Information

Our research complies with all relevant ethical regulations. The study was approved by the Yale Human Research Protection Program Institutional Review Board (protocol IDs 2000027892 and 2000022385), the Ben-Gurion University of the Negev Human Subjects Research Committee (request no. 20TrustCovR), the Centre for Experimental Social Sciences Ethics Committee (OE_0055), and the NHH Norwegian School of Economics Institutional Review Board (NHH-IRB 10/20). Informed consent was obtained from all participants.

Design

Overview. An overview of the experiment is depicted in Extended Data Figure 1. After selecting their language, providing their consent, and passing two attention checks, participants were told that they would “read about three different debates that are happening right now around the world”, that they would be given “some of the justifications that politicians and experts are giving for different policies”, and that they would be “ask[ed] some questions about [their] opinions”. They then completed two tasks measuring their trust in leaders expressing either utilitarian or non-utilitarian opinions (one using a behavioral measure and one using self-report measures, presented in a randomized order); these tasks were followed by questions about their impressions about the ongoing pandemic crisis, as well as individual difference and demographic measures, as detailed below. Data collection was performed blind to the conditions of the participants.

Both behavioral and self-report measures of trust involved five debates on the current pandemic crisis, three of which involved instrumental harm (IH), and two impartial beneficence (IB) (summarized in Figure 1C and Table 1; for full text, see Supplementary Methods). Each of these five dilemmas were based on real debates that have been occurring during the COVID-19 pandemic, and we developed the philosophical components of each argument in consultation with moral philosophers.

1) **Lockdown** (Instrumental Harm): whether the country should maintain severe restrictions on social gatherings until a vaccine is developed to prevent COVID-related deaths, or consider relaxing restrictions to maximize overall well-being
2) **Ventilators** (Instrumental Harm): whether doctors should give everyone equal access to COVID treatment, or prioritize younger and healthier people
3) **Tracing** (Instrumental Harm): whether the government should make it mandatory for residents to wear contact tracing devices to prevent pandemic spread, or make tracing devices optional to respect residents' right to privacy
4) **Medicine** (Impartial Beneficence): whether medicine developed in the home country should be reserved for treating the home country's citizens, or sent wherever it can do the most good, even if that means sending it to other countries
5) **PPE** (Impartial Beneficence): whether PPE manufactured in the home country should be reserved for protecting the home country's citizens, or sent wherever it can do the most good, even if that means sending it to other countries

See Supplementary Notes 2 and 6-9 for further details of why we chose these specific dilemmas and how they can test our theoretical predictions.

Translations. Where the survey was administered in a non-English-speaking country, study materials were translated following a standard forward- and back-translation procedure [78]. First, for forward translation, a native speaker translated materials from English to the target language. Second, for back translation, a second native translator (who had not seen the original English materials) translated the materials back into English. Results were then compared, and if there were any substantial discrepancies, a second forward- and back-translation was conducted with translators working in tandem to resolve issues. Finally, the finished translated and back-translated materials were checked by researchers coordinating the experiment for that country.

Experimental design. Participants were randomly and blindly assigned to one of 4 conditions in the beginning of the experiment. These conditions corresponded to a 2x2 between-subjects design: 2 (moral dimension in the voting task: Instrumental Harm or Impartial Beneficence) x 2 (argument in the self-report task: Utilitarian or Non-Utilitarian). In addition, we randomized the order of tasks (voting or self-report task first), the order of arguments in the voting task (Utilitarian or Non-Utilitarian first), the order of dilemmas in the self-report task (Lockdown, Ventilators, or Tracing first if Instrumental Harm, and PPE or Medicine first if Impartial Beneficence), and the dilemmas displayed (two in the self-report task and one in the voting task randomly chosen among Lockdown, Ventilators, and Tracing if Instrumental Harm, and PPE and Medicine if Impartial Beneficence). This design allowed us to minimize demand characteristics with between-subjects manipulations of key experimental factors, while at the same time maximizing efficiency of data collection.

Attention checks. We included two attention checks prior to the beginning of the experiment. Any participants who failed either of these were then screened out immediately. First, participants were told:

> "In studies like ours, there are sometimes a few people who do not carefully read the questions they are asked and just 'quickly click through the survey.' These random answers are problematic because they compromise the results of the studies. It is very important that you pay attention and read each question. In order to show that you read our questions carefully (and regardless of your own opinion), please answer 'TikTok' in the question on the next page"

Then, on the next page, participants were given a decoy question: "When an important event is happening or is about to happen, many people try to get informed about the development of the situation. In such situations, where do you get your information from?". Participants were asked to select among the following possible answers, displayed in a randomized order: TikTok, TV, Twitter, Radio, Reddit, Facebook, Youtube, Newspapers, Other. Participants who failed to follow our instructions and selected any answer other than the instructed one ("TikTok") were then screened out of the survey. Second, participants were asked to read a short paragraph about the history and geography of roses. On the following page, they were asked to indicate which of six topics was not discussed in the paragraph. Participants who answered incorrectly were then screened out of the survey (with the exception of those who participated via Prolific, who were instead allowed to continue due to platform requirements).

Dilemma introduction. Both the voting and self-report tasks began with an introduction to a specific dilemma. In the voting task, participants viewed a single dilemma, and in the self-report task, participants viewed two dilemmas in randomized order (see Extended Data Figure 1 for details). No participant saw the same dilemma in both the voting and self-report tasks.

The dilemma introduction consisted of a short description of the dilemma (for example, in the PPE dilemma: “Imagine that [...] there will soon be another global shortage of personal protective equipment [... and] political leaders are debating how personal protective equipment should be distributed around the globe.”), followed by a description of two potential policies (for example, in the PPE dilemma, US participants read: “[S]ome are arguing that PPE made in American factories should be sent wherever it can do the most good, even if that means sending it to other countries. Others are arguing that PPE made in American factories should be kept in the U.S., because the government should focus on protecting its own citizens.”).

After reading about the dilemma, participants were asked to provide their own opinion about the best course of action (“Which policy do you think should be adopted?”), answered on a 1-7 scale, with the endpoints (1 and 7) representing strong preferences for one of the policies (for example, in the PPE dilemma they were labeled “Strongly support U.S.-made PPE being reserved for protecting American citizens” and “Strongly support U.S.-made PPE being given to whoever needs it most”, respectively), and the midpoint (4) representing indifference (“Indifferent”). See Supplementary Note 13 for further details. As an exploratory measure that is not analyzed for the purposes of the current report, participants also indicated how morally wrong it would be for politicians to endorse the utilitarian approach in each dilemma.

For full text of dilemmas and introduction questions, see Supplementary Methods.

Voting task. Our behavioral measure of trust in the current studies is based on a novel task with two types of participants: *voters* and *donors.* Voters were asked to cast a vote for a leader who would be responsible for making a charitable donation to UNICEF on behalf of a group of donors, and would have the opportunity to ‘embezzle’ some of the donation money for themselves (Figure 1D).

We collected data from donors first. A few days before we ran our main experiment, a convenience sample of US participants (N = 100) was recruited from Prolific and was provided with a US$2 bonus endowment. They were given the opportunity to donate up to their full bonus to UNICEF. After making their donation decision, they read about the five COVID-19 dilemmas, in randomized order, and indicated which policy they thought should be adopted. Finally, they were instructed that they might be selected to be responsible for the entire group’s donations to UNICEF. Participants were told that if they were selected, they would have the opportunity to keep up to the full amount of total group donations for themselves, and were asked to indicate how much of the group’s donations they would keep for themselves if they were selected to be responsible.

Our main experiment focused on the behavior of voter participants. In the voting task, participants were randomly assigned to read about one dilemma, randomly selected amongst the five dilemmas summarised in Table 1. After completing the dilemma introduction, participants were asked to “make a choice that has real financial consequences” and told that “[a] few days ago, a group of 100 people were recruited via an international online marketplace and invited to make donations to the charitable organization UNICEF. In total, they donated an amount equivalent to $87.89". We instructed participants that we would like them to "vote for a leader to be responsible for the entire group's donations". Crucially, they were also told that "[t]he leader has two options: They can transfer the group's $87.89 donation to UNICEF in full, *or* [t]hey can take some of this money for themselves (up to the full amount) and transfer whatever amount is left to UNICEF". The exact donation amount was determined by the actual donation choices of the donor participants.

Following these details, participants were asked to cast a vote for the leadership position between two people who had also read about the same dilemma they had just read about. Participants were instructed that one person agreed with the utilitarian argument, while the other person agreed with the non-utilitarian argument. This information was displayed to participants on the same page, in a randomized order. Participants were then asked to vote for the person they wished to be responsible for the group's donations. We instructed participants that we would later identify the winner of the election, and implement their choice by distributing payments to the leader and UNICEF accordingly.

After completing the voting task, voter participants were asked the following comprehension question: "In the last page, you were asked to choose a leader that will be entrusted with the group's donation. Please select the option that best describes what the leader will be able to do with the donation". They were asked to select between three options, displayed in randomized order:

1. The leader can transfer the full donation to UNICEF or take some of the money for themselves.
2. The leader is not able to do anything with the donation.
3. The leader chooses how much of the group's donation to keep for themselves and how much to return to the people who donated the money.

We excluded voter participants who failed to select the correct answer (1), as per our exclusion criteria (see Exclusions). Note that in our stage 1 Registered Report the answer choices were slightly different, but we revised them after discovering in a soft launch that participants were systematically choosing one of the incorrect options, suggesting that the question was poorly worded. In consultation with the editor, we clarified the response options and began the data collection procedure anew. This was one of only three deviations from the stage 1 report (the others being that data collection took four weeks instead of the two weeks we had anticipated, and the use of Prolific instead of Lucid for recruitment in the United Kingdom and the United States).

After collecting the votes from the voter participants, we randomly selected ten donor participants to be considered for the leadership position: one who endorsed the utilitarian position for each of the five dilemmas and one who endorsed the non-utilitarian position for each of the five dilemmas. After tallying the votes from voter participants, we implemented the choices of each of the elected leaders and made the payments accordingly. For full text of instructions and questions for both the donor and the voting task, see Supplementary Methods.

Self-reported trust. Participants read about two dilemmas on the dimension of utilitarianism that they did *not* encounter in the voting task. That is, participants assigned to an Instrumental Harm dilemma (Lockdown, Ventilators, or Tracing) for the voting task read both Impartial Beneficence dilemmas (PPE and Medicine) for the self-report task, while participants assigned to an Impartial Beneficence dilemma (PPE or Medicine) for the voting task read a randomly assigned two out of three Instrumental Harm dilemmas (Lockdown, Ventilators, and Tracing) for the self-report task. The structure of the introduction to the dilemmas was identical to that in the voting task: they read a short description of the issue, followed by a description of two potential policies. On separate screens, they were asked which policy they themselves support.

After providing their own opinions, participants were asked to imagine that the mayor of a major city in their region was arguing for one of the two policies, providing either a utilitarian or non-utilitarian argument. Each participant was randomly assigned to read about leaders making either utilitarian or non-utilitarian arguments in both dilemmas presented in the self-report task. After reading about the leader's opinion and argument, they were then be asked to report their general trust in the leader ("How trustworthy do you think this person is?"), to be answered on a 1-7 scale, with labels "Not at all trustworthy", "Somewhat trustworthy", and "Extremely trustworthy" at points 1, 4, and 7, respectively. On a separate page they were then asked to report their trust in the leader's advice on other issues ("How likely would you be to trust this person's advice on other issues?"), to be answered on a 1-7 scale, with labels "Not at all likely", "Somewhat likely", and "Extremely likely" at points 1, 4, and 7, respectively.

After completing the self-report task, participants were asked the following comprehension question: "In the last page, you read about a mayor in a city in your region, and were asked about them. Please select the option that best describes the questions you were asked". Their options, displayed in a randomized order, were: (1) "How much I agreed with the mayor"; (2) "How much I trusted the mayor"; and (3) "How much I admired the mayor". This allowed us to exclude participants who failed to select the correct answer (2), as per our exclusion criteria (see Exclusions).

For full text of instructions and questions for the self-report task, see Supplementary Methods.

COVID concern. To assess their attitudes toward and experience with the pandemic, participants were asked three questions. Two measured how concerned participants currently felt about the pandemic, on both health-related and economic grounds ("How concerned are you about the health-related consequences of the COVID-19 pandemic?" and "How concerned are you about the financial and economic consequences of the COVID-19 pandemic?", both to be answered on a 1-7 scale, with labels "Not at all" and "Very much" at points 1 and 7, respectively). The third question measured their personal involvement ("Have you or anyone else you know personally suffered significant health consequences as a result of COVID-19?", to be answered by selecting one of three options: "Yes", "No", and "Unsure").

Oxford Utilitarianism Scale. All participants then completed the Oxford Utilitarianism Scale [33]. The scale consists of nine items in two subscales: instrumental harm (OUS-IH) and impartial beneficence (OUS-IB). The OUS-IB subscale consists of five items that measure endorsement of impartial maximization of the greater good, even at great personal cost (e.g., "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal"). The OUS-IH subscale consists of four items relating to willingness to cause harm so as to bring about the greater good (e.g., "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people"). Participants viewed all questions in a randomized order, and answered on a 1-7 scale, with labels "Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", and "Strongly agree".

Demographics. All participants were asked to report their gender, age, years spent in education, subjective SES, education (on the same scale, but with minor changes in the scale labels across countries), political ideology (using an item from the World Values Survey), and religiosity. These questions were the same across countries and represent the demographics used as covariates in the main analyses. Additionally, participants were asked to indicate their region of residence (for example, for the United States, "Which US State do you currently live in?"), and ethnicity/race, with the specific wording and response options depending on the local context (in France and Germany, this was not collected due to local regulations). In addition, participants were asked to confirm their country of residence, which allowed us to exclude participants who reported living in a country different from that of intended recruitment, as per our exclusion criteria (see Exclusions).

Debriefing questions. Finally, participants were asked a series of debriefing questions. Two of these assessed their participation in other COVID-related studies ("Approximately how many COVID-related studies have you participated in before this one?", answered by selecting one of the following options: "0", "1-5", "6-10", "11-20", "21-50", "More than 50", and "I don't remember"; and "If you have participated in any other COVID-related studies, how similar were they to this one?", to be answered by selecting one of the following options: "Extremely similar", "Very similar", "Moderately similar", "Slightly similar", "Not at all similar", and "Not applicable").

An additional question assessed participants' attitudes towards the charity involved in the voting task ("How reliable do you think UNICEF is as an organization in using donations for helping people?", answered on a 1-5 scale, with labels "Not reliable at all", "Somewhat reliable", and "Very reliable" at points 1, 3, and 5, respectively).

Analysis Plan: Pre-processing

Exclusions. We planned to exclude data either at the participant level as outlined in the Sampling Plan section, based on criteria 1 (duplicate response), 2 (different residence), and 3 (partial completion), or on an analysis-by-analysis basis as outlined in criteria 4 (missing variables) and 5 (failed comprehension checks).

Outliers. All participants' responses were analyzed, regardless of whether they were statistical outliers.

Computation of composite measures. Composite measures of self-reported trust were created by averaging responses to the two trust questions (trustworthiness of the leader, and trust in the leader's advice on other issues), separately for each participant and dilemma. In addition, we created composite OUS scores for each participant by averaging their responses on the scale items, separately for the Instrumental Harm (4 items) and Impartial Beneficence subscales (5 items).

Analysis Plan for Hypothesis Testing

We planned to examine behavioral measures and self-report measures of trust in two separate models. For testing our hypotheses across all countries, we set a significance threshold of alpha = .0025 (Bonferroni corrected for two tests). All analyses were conducted in R using the packages lme4 [79], lmerTest [80], estimatr [81], emmeans [82], ggeffects [83], ri2 [84], and glmnet [85]. We planned that, in the event of convergence or singularity issues, we would supplement the theoretically appropriate models described below with simplified models by reducing the complexity of the random-effects structure [86].

Hypothesis 1: Self-reported trust. To examine participants' self-reported trust in the leaders, we planned to examine the composite measure of their trust in each leader (i.e. the average of the two trust questions, computed separately for each participant and dilemma). We hypothesized that participants would report higher trust in non-utilitarian leaders compared with utilitarian leaders in the context of dilemmas involving instrumental harm, while the opposite pattern would be observed for impartial beneficence. To test this hypothesis, we planned to conduct a linear mixed-effects model of the effect of argument type (Utilitarian vs. Non-Utilitarian), dimension type (Instrumental Harm vs. Impartial Beneficence), and their interaction, on the composite score of trust, adding demographic variables (namely gender, age, education, subjective SES, political ideology, and religiosity) and policy support as fixed effects, and dilemmas and countries as random intercepts, with participants nested within countries. In addition, we planned to run a model that included countries as random slopes of the two main effects and the interactive effect. We said that, should the model converge, and should the results differ from the simpler model proposed above, we would compare model fits using the Akaike information criterion (AIC), and retain the model that better fits the data – while still reporting the other in supplementary materials. We planned to follow up on significant effects with post hoc comparisons using Bonferroni corrections. For the purposes of the analysis, we used effect coding such that for argument type, the Non-Utilitarian condition was coded as -0.5 and the Utilitarian condition as 0.5, and for the dimension type, Instrumental Harm was coded as -0.5, and Impartial Beneficence as 0.5. The demographic covariates were grand-mean-centered; the gender variable was dummy coded with "woman" as baseline. P values were computed using Satterthwaite's approximation for degrees of freedom as implemented in lmerTest. For analysis code, see https://osf.io/m9tpu/.

Hypothesis 2: Voting measure. To examine participants' trust in the leaders as demonstrated by their behavior, we planned to examine their choices in the voting task, where they were asked to select which of two leaders (one making a utilitarian argument, and the other a non-utilitarian one) to entrust with a group charity donation. We hypothesized that participants would be more likely to select the non-utilitarian leader over the utilitarian leader when reading about their arguments for dilemmas involving instrumental harm, while the opposite pattern would be observed for impartial beneficence. To test this hypothesis, we planned to conduct a generalized linear mixed-effects model with the logit link of the effect of dimension type (Instrumental Harm vs. Impartial Beneficence) on the leader choice (Utilitarian vs. Non-Utilitarian), adding demographic variables (namely gender, age, education, subjective SES, political ideology, and religiosity) and policy support as fixed effects, and dilemmas and countries as random intercepts, with participants nested within countries. In addition, we said we would also run a model that includes countries as random slopes of the effect of dimension type. Should the model converge, and should the results differ from the simpler model proposed above, we planned to compare model fits using the Akaike information criterion (AIC), and retain the model that better fits the data – while still reporting the other in supplementary materials. Based on recent reports that linear models might be preferable to logistic models in treatment designs [63,64], we said we would run the same analysis using a linear model (instead of logit link) with the identical fixed and random effects and again adjudicate between the models using the AIC. We planned to follow up on any significant effects observed with post hoc comparisons using Bonferroni corrections. For the purposes of this analysis, we planned to use effect coding such that for the binary response variable of argument type, the Non-Utilitarian trust response was coded as 0 and the Utilitarian trust response as 1, and for the dimension type, Instrumental Harm was coded as -0.5, and Impartial Beneficence as 0.5. Again, the demographic covariates were grand-mean-centered; the gender variable was dummy coded with "woman" as baseline. P values were computed using Satterthwaite's approximation for degrees of freedom as implemented in lmerTest. For analysis code, see https://osf.io/m9tpu/.

Robustness checks. Because there was evidence that public perceptions of lockdowns at the time of data collection were changing relative to July 2020 when we ran our pilots [87,88], which may affect responses to the Lockdown dilemma, we planned to examine the robustness of our findings using two variations of the models described above, one that includes the Lockdown dilemma and another that omits it.

As some of the countries in our sample already implement mandatory and/or invasive contact tracing schemes at the time of writing (China, India, Israel, Singapore, and South Korea), which may affect responses to the Tracing dilemma, we also planned to examine the robustness of our findings in these countries using two variations of the models described above, one that includes the Tracing dilemma and another that omits it. Furthermore, in this subset of countries we planned to examine an order effect to test whether completing the Tracing dilemma in the first task affects behavior on the subsequent task.

Null hypothesis testing. In the event of non-significant results from the approaches outlined above, we planned to employ the TOST procedure [89] to differentiate between insensitive versus null results. In particular, we planned to specify lower and upper equivalence bounds based on standardized effect sizes set by our SESOI (see Power Analysis and Table 2). For each of our two tasks, should the larger of the two *p* values from the two t tests be smaller than alpha = .05, we would conclude statistical equivalence. For example, the minimum guaranteed sample size (N = 12,600; see Sample Size for details) would give us over 95% power to detect an effect size of *d* = 0.05 in the self-report task, yielding standardized Δ_L = -0.05 and Δ_U = 0.05, and *OR* = 1.30 in the voting task, yielding standardized Δ_L = -0.15 and Δ_U = 0.15.

Sampling Plan

Participants. We planned to complete the study online with participants in the following countries: Australia, Brazil, Canada, Chile, China, Denmark, France, Germany, India, Israel, Italy, the Kingdom of Saudi Arabia, Mexico, the Netherlands, Norway, Singapore, South Africa, South Korea, Spain, the United Arab Emirates, the United Kingdom, and the United States (Figure 1A). We sampled on every inhabited continent and included countries that have been more or less severely affected by COVID-19 on a variety of metrics (Supplementary Figure 1). Country selection was determined primarily on a convenience basis. In April 2020, the senior author put out a call for collaborators via social media and email. Potential collaborators were asked whether they had the capacity to recruit up to 1,000 participants representative for age and gender within their home country. After the initial set of collaborators was established, we added additional countries to diversify our sample with respect to geographic location and pandemic severity.

We planned to recruit participants via online survey platforms (see Supplementary Table 1) and compensate them financially for their participation in accordance with local standard rates. We aimed to recruit samples that were nationally representative with respect to age and gender where feasible. We anticipated that this would be feasible for many, but not all countries in our study (see Supplementary Table 1 for details). We originally anticipated sampling to take place over a 14-day period, but to allow for more representative sampling (after discussion with the editor), we collected data over a period of 27 days (November 26th 2020 to December 22nd 2020). All survey materials were translated into the local language (see Translations for details). Prior to the survey, all participants read and approved a consent form outlining their risks and benefits, confirmed they agreed to participate in the experiment, and completed two attention checks. Participants who failed to agree to the consent or failed to pass the attention checks were not permitted to complete the survey (with the exception of participants in the United States and the United Kingdom, who due to recruitment platform requirements were instead allowed to continue the survey, and were only excluded after data collection).

Expected effect sizes. We informed our expected effect sizes by examining the published literature on utilitarianism and trust. Previous studies of social impressions of utilitarians reveal effect sizes in the range of $d = 0.19 – 0.78$ (mean $d = 0.78$ for the effect of instrumental harm on self-reported moral impressions; mean $d = 0.19$ for the effect of impartial beneficence on self-reported moral impressions; mean $d = 0.55$ for interactive effects of instrumental harm and impartial beneficence on self-reported moral impressions)[35–39]. However, there are several important caveats with using these past studies to inform expected effect sizes for the current study. First, past studies have measured trust in ordinary people, while we study trust in leaders, and there is evidence that instrumental harm and impartial beneficence differentially impact attitudes about leaders versus ordinary people[37]. Second, past studies have investigated artificial moral dilemmas, while we study real moral dilemmas in the context of an ongoing pandemic. Third, past studies have been conducted in a small number of Western countries (US, UK, Germany) while we sample across a much wider range of countries on 6 continents. Finally, for the voting task, it is more challenging to estimate an expected effect size because no previous studies to our knowledge have used such a task.

Because of the caveats described above, we also informed our expectations of effect sizes with data from Pilot 2, which was identical to the proposed studies in design apart from using The Red Cross instead of UNICEF in the voting task and the omission of the Tracing dilemma (see Pilot Data in Supplementary Information for a full description of the pilot experiments). Pilot 2 revealed a conventionally medium effect size for the interaction between argument and moral dimension in the self-report task ($B = 2.88$, SE = 0.24, $t(452) = 11.80$, $p < .001$, CI = [2.41, 3.35], $d = 0.55$), and a conventionally large effect size for the effect of moral dimension in the voting task ($B = 2.41$, SE = 0.33, $z = 7.30$, $p < .001$, CI = [1.77, 3.13], OR = 11.13, $d = 1.33$).

Sample size. Sample size was determined based on a cost-benefit analysis considering available resources and expected effect sizes that would be theoretically informative [89] (see Expected effect sizes). We aimed to collect the largest sample possible with resources available and verified with power analyses that our planned sample would be able to detect effect sizes that are theoretically informative and at least as large as expected based on prior literature (see Power Analysis). We expected to collect a sample of 21,000 participants in total, which conservatively accounting for exclusion rates up to 40% (see Exclusions) would lead to a final guaranteed minimum sample of 12,600 participants.

Power analysis. We conducted a series of power analyses to determine the smallest effect sizes that our minimum guaranteed sample of 12,600 participants would be able to detect with 95% power and an alpha level of .005, separately for each main model (see Analysis Plan for Hypothesis Testing for further details). To account for these two hypothesis tests, for all power analyses we applied Bonferroni corrections for two tests, thus yielding an alpha of .0025. Following recent suggestions [90,91], results passing a corrected alpha of $p \leq .005$ are interpreted as *'supportive* evidence' for our hypotheses, while results passing a corrected alpha of $p < .05$ are interpreted as *'suggestive* evidence'. Power analyses were conducted using Monte Carlo simulations [92] via the R package simr [93], with 1,000 simulations, using estimates of means and variances from pilot 2 (see Pilot Data in Supplementary Information for a full description of the pilot experiments; note that, for the purposes of the current simulations, the race variable was omitted from data analysis because this variable is not readily comparable across countries). Data and code for power analyses can be found at https://osf.io/m9tpu/.

First, we considered the interactive effect of moral dimension (Instrumental Harm vs. Impartial Beneficence) and argument (Utilitarian vs. Non-Utilitarian) on trust in the self-report task. We estimated that a sample of 12,600 participants would provide over 95% power to detect an effect size of $d = 0.05$ (power = 99.3%, CI = [98.56, 99.72]). This effect size is 9% of what we observed in pilot 2 and is the SESOI for the self-report task.

Next, we considered the effect of moral dimension (Instrumental Harm vs. Impartial Beneficence) on leader choice in the voting task. We estimated that a sample of 12,600 participants would provide over 95% power to detect an odds ratio of 1.30 (power = 95.8%, CI = [94.36, 96.96]). This effect size is 9% of what we observed in pilot 2 and is the SESOI for the voting task.

Given that these SESOI values are detectable at 95% power with our guaranteed sample (total N = 12,600), are theoretically informative, and are lower than our expected effect sizes (see Expected effect sizes), we concluded that our sample is sufficient to provide over 95% power for testing our hypotheses, and that our study is highly powered to detect useful effects.

At the time of submission, online survey platform representatives indicated that, while it is normally feasible to recruit samples nationally representative for age and gender in most of our target countries, due to the ongoing pandemic, final sample sizes may be unpredictable and in some countries it would not be possible to achieve fully representative quotas for some demographic categories, including women and older people (see Supplementary Table 1 for details). We planned that, if this issue arose, we would prioritize statistical power over representativeness. If we were unable to achieve representativeness for age and/or gender in particular countries, we planned to note this explicitly in the Results section.

Exclusions. We planned to exclude participants from all further analyses if they met at least one of the following criteria: (1) they had taken the survey more than once (as indicated by IP address or worker ID); (2) they reported in a question about their residence (further described in Design) that they lived in a country different from that of intended recruitment; (3) they did not answer more than 50% of the questions. In addition, participants would be selectively excluded from specific analyses if they (4) did not provide a response and are thus missing variables involved in the analysis; or (5) failed the comprehension check (further described in Design) for the task involved in the specific analysis.

Data Availability

All data and materials are openly available on the Open Science Framework (OSF) website at this link: https://osf.io/m9tpu/.

Code Availability

All analysis code (completed in R) are openly available on the Open Science Framework (OSF) website at this link: https://osf.io/m9tpu/.

References

1. Wilson, S. Pandemic leadership: lessons from New Zealand's approach to COVID-19. *Leadership* **16**, 279–293 (2020).
2. Levi, M. & Stoker, L. Political trust and trustworthiness. *Annu. Rev. Polit. Sci.* **3**, 475–507 (2000).
3. Ferguson, N. et al. Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *Imperial College London* http://spiral.imperial.ac.uk/handle/10044/1/77482 (2020).
4. Fink, S. Worst-case estimates for U.S. coronavirus deaths. *The New York Times* (13 March 2020).
5. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
6. Hsiang, S. et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262–267 (2020).
7. Alsan, M. & Wanamaker, M. Tuskegee and the health of Black men. *Q. J. Econ.* **133**, 407–455 (2018).
8. Christensen, D., Dube, O., Haushofer, J., Siddiqi, B. & Voors, M. J. Building resilient health systems: experimental evidence from Sierra Leone and the 2014 Ebola outbreak. *National Bureau of Economic Research* https://www.nber. org/papers/w27364 (2020).

9. Lowes, S. & Montero, E. The legacy of colonial medicine in Central Africa. *Am. Econ. Rev.* **111**, 1284–1314 (2021).

10. Udow-Phillips, M. & Lantz, P. Trust in public health is essential amid the COVID-19 Pandemic. *J. Hosp. Med.* **15**, 431–433 (2020).

11. Blair, R. A., Morse, B. S. & Tsai, L. L. Public health and public trust: survey evidence from the Ebola virus disease epidemic in Liberia. *Soc. Sci. Med. 1982* **172**, 89–97 (2017).

12. Rubin, G. J., Amlôt, R., Page, L. & Wessely, S. Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: cross sectional telephone survey. *BMJ* **339**, b2651 (2009).

13. Gilles, I. et al. Trust in medical organizations predicts pandemic (H1N1) 2009 vaccination behavior and perceived efficacy of protection measures in the Swiss public. *Eur. J. Epidemiol.* **26**, 203–210 (2011).

14. Prati, G., Pietrantoni, L. & Zani, B. Compliance with recommendations for pandemic influenza H1N1 2009: the role of trust and personal beliefs. *Health Educ. Res.* **26**, 761–769 (2011).

15. Maher, P. J., MacCarron, P. & Quayle, M. Mapping public health responses with attitude networks: the emergence of opinion-based groups in the UK's early COVID-19 response phase. *Br. J. Soc. Psychol.* **59**, 641–652 (2020).

16. Plohl, N. & Musil, B. Modeling compliance with COVID-19 prevention guidelines: the critical role of trust in science. *Psychol. Health Med*. 1–12 (2020).

17. Dohle, S., Wingen, T. & Schreiber, M. Acceptance and adoption of protective measures during the COVID-19 pandemic: the role of trust in politics and trust in science. Preprint at *OSF* https://osf.io/w52nv (2020).

18. Han, Q. et al. (2021). Trust in government regarding COVID-19 and its associations with preventive health behaviour and prosocial behaviour during the pandemic: A cross-sectional and longitudinal study. Psychological Medicine, 1-32. https://doi.org/10.1017/S0033291721001306

19. Bramble, B. *Pandemic Ethics: 8 Big Questions of COVID-19* (Bartleby Books, 2020).

20. Emanuel, E. J. et al. Fair allocation of scarce medical resources in the time of Covid-19. *N. Engl. J. Med.* **382**, 2049–2055 (2020).

21. Everett, J. A. C. & Kahane, G. Switching tracks? towards a multidimensional model of utilitarian psychology. *Trends Cogn. Sci.* **24**, 124–134 (2020).

22. Giubilini, A., Savulescu, J. & Wilkinson, D. COVID-19 vaccine: vaccinate the young to protect the old?. *J. Law. Biosci.* **7**, lsaa050 (2020).

23. Savulescu, J., Persson, I. & Wilkinson, D. Utilitarianism and the pandemic. *Bioethics* **34**, 620–632 (2020).

24. Savulescu, J. & Cameron, J. Why lockdown of the elderly is not ageist and why levelling down equality is wrong. *J. Med. Ethics* **46**, 717–721 (2020).

25. Fried, C. *Right and Wrong* (Harvard Univ. Press, 1978).

26. Kant, I. *Groundwork for the Metaphysics of Morals* (Yale Univ. Press, 2002).

27. Rawls, J. *A Theory of Justice* (Belknap Press of Harvard Univ. Press, 1971).

28. Ross, W. D. *The Right and the Good* (Oxford Univ. Press, 1930).

29. Scanlon, T. *What We Owe to Each Other* (Belknap Press, 1998).

30. Liddell, K., Martin, S. & Palmer, S. Allocating medical resources in the time of Covid-19. *N. Engl. J. Med.* **382**, e79 (2020).

31. Conway, P., Goldstein-Greenwood, J., Polacek, D. & Greene, J. D. Sacrificial utilitarian judgments do reflect concern for the greater good: clarification via process dissociation and the judgments of philosophers. *Cognition* **179**, 241–265 (2018).

32. Awad, E., Dsouza, S., Shariff, A., Rahwan, I. & Bonnefon, J.-F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc. Natl Acad. Sci. USA* **117**, 2332–2337 (2020).

33. Kahane, Everett,G. et al. Beyond sacrificial harm: a two-dimensional model of utilitarian psychology. *Psychol. Rev.* **125**, 131–164 (2018).

34. Navajas, J. et al. Utilitarian reasoning about moral problems of the COVID-19 crisis. Preprint at *OSF* https://osf.io/ktv6z (2020).

35. Bostyn, D. H. & Roets, A. Trust, trolleys and social dilemmas: a replication study. *J. Exp. Psychol. Gen.* **146**, e1–e7 (2017).

36. Everett, J. A. C., Pizarro, D. A. & Crockett, M. J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **145**, 772–787 (2016).

37. Everett, J. A. C., Faber, N. S., Savulescu, J. & Crockett, M. J. The costs of being consequentialist: social inference from instrumental harm and impartial beneficence. *J. Exp. Soc. Psychol.* **79**, 200–216 (2018).

38. Rom, S. C., Weiss, A. & Conway, P. Judging those who judge: perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *J. Exp. Soc. Psychol.* **69**, 44–58 (2017).

39. Sacco, D. F., Brown, M., Lustgraaf, C. J. N. & Hugenberg, K. The adaptive utility of deontology: deontological moral decision-making fosters perceptions of trust and likeability. *Evol. Psychol. Sci.* **3**, 125–132 (2017).

40. Uhlmann, E. L., Zhu, L. (Lei). & Tannenbaum, D. When it takes a bad person to do the right thing. *Cognition* **126**, 326–334 (2013).

41. Trump, D. J. WE CANNOT LET THE CURE BE WORSE THAN THE PROBLEM ITSELF. AT THE END OF THE 15 DAY PERIOD, WE WILL MAKE A DECISION AS TO WHICH WAY WE WANT TO GO! *Twitter* https://twitter.com/realDonaldTrump/ status/1241935285916782593?s=20 (2020).

42. Patrick, D. Tucker. *Carlson Tonight* (2020).

43. Burke, D. Reopening the country: the dangerous moral arguments behind this movement. *CNN* https://edition.cnn.com/2020/04/23/us/ reopening-country-coronavirus-utilitarianism/index.html (2020).

44. Rosenbaum, L. Facing Covid-19 in Italy — ethics, logistics, and therapeutics on the epidemic's front line. *N. Engl. J. Med.* **382**, 1873–1875 (2020).

45. Fahey, R. A. & Hino, A. COVID-19, digital privacy, and the social limits on data-focused public health responses. *Int. J. Inf. Manag.* https://doi. org/10.1016/j.ijinfomgt.2020.102181 (2020).

46. Asher, S. TraceTogether: Singapore turns to wearable contact-tracing Covid tech. *BBC News* (5 July 2020).

47. From India to Cyprus, understanding the global debate over virus contact tracing apps. *The Week* https://www.theweek.in/news/world/ 2020/05/07/ understanding-the-global-privacy-debate-over-coronavirus-contact-tracing-apps.html (2020).

48. Jeske, D in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford Univ., 2014).

49. Breuninger, K. & Wilkie, C. Trump bans export of coronavirus protection gear, says he's 'not happy with 3M'. *CNBC* https://www.cnbc.com/2020/04/03/ coronavirus-trump-to-ban-export-of-protective-gear-after-slamming-3m.html (2020).

50. Trump administration secures new supplies of remdesivir for the United States. *US Department of Health and Human Services* https://www.hhs.gov/ about/news/2020/06/29/trump-administration-secures-new-supplies-remdesivir-united-states.html (2020).

51. Boris, J. Prime minister's statement on coronavirus (COVID-19): 3 June 2020. *GOV.UK*https://www.gov.uk/government/speeches/pm-statement-at-the-coronavirus-press-conference-3-june-2020 (2020).

52. Kerris, M. Onze missie: de hele wereld een vaccin. *NRC Handelsblad* https://www.nrc.nl/nieuws/2020/05/14/onze-missie-de-hele-wereld-een-vaccin-a3999818 (2020).

53. Foot, P. The problem of abortion and the doctrine of the double effect. *Oxf. Rev.* **5**, 5–15 (1967).

54. Thomson, J. J. The trolley problem. *Yale Law J.* **94**, 1395–1415 (1985).

55. Kupferschmidt, K. The lockdowns worked—but what comes next? *Science* **368**, 218–219 (2020).

56. Gertz, G. in *Reopening the World: How to Save Lives and Livelihoods* (eds. Allen, J. R. & West, D. M.) 12–15 (The Brookings Institution, 2020).

57. Mehrotra, P., Malani, P. & Yadav, P. Personal protective equipment shortages during COVID-19—supply chain-related causes and mitigation strategies. *JAMA Health Forum* **1**, e200553–e200553 (2020).

58. Zhou, Y. R. The global effort to tackle the coronavirus face mask shortage. *The Conversation* http://theconversation.com/the-global-effort-to-tackle-the-coronavirus-face-mask-shortage-133656 (2020).

59. Bollyky, T. J., Gostin, L. O. & Hamburg, M. A. The equitable distribution of COVID-19 therapeutics and vaccines. *JAMA* **323**, 2462–2463 (2020).

60. Liu, Y., Salwi, S. & Drolet, B. C. Multivalue ethical framework for fair global allocation of a COVID-19 vaccine. *J. Med. Ethics* **46**, 499–501 (2020).

61. Edelson, M. G., Polania, R., Ruff, C. C., Fehr, E. & Hare, T. A. Computational and neurobiological foundations of leadership decisions. *Science* **361**, eaat0036 (2018).

62. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).

63. Gomila, R. Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *J. Exp. Psychol. Gen.* https://doi.org/10.1037/xge0000920 (2020).

64. Angrist, J. D. & Pischke, J. S. *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton Univ. Press, 2009).

65. Gomila, R. Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *J. Exp. Psychol. Gen.* **150**, 700–709 (2021).

66. Hellevik, O. Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quant.* **43**, 59–74 (2009).

67. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).

68. Clark, T. S. & Linzer, D. A. Should I use fixed or random effects. *Polit. Sci. Res. Methods* **3**, 399–408 (2015).

69. McNeish, D. & Kelley, K. Fixed effects models versus mixed effects models for clustered data: reviewing the approaches, disentangling the differences, and making recommendations. *Psychol. Methods* **24**, 20 (2019).

70. Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious mixed models. Preprint at a*rXiv* https://arxiv.org/abs/1506.04967v2 (2018).

71. Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D. & Ledgerwood, A. Using independent covariates in experimental designs: quantifying the trade-off between power boost and Type I error inflation. *J. Exp. Soc. Psychol.* **72**, 118–124 (2017).

72. Gelfand, M. J., Nishii, L. H. & Raver, J. L. On the nature and importance of cultural tightness–looseness. *J. Appl. Psychol.* **91**, 1225–1244 (2006).

73. Gelfand, M. J. et al. The relationship between cultural tightness–looseness and COVID-19 cases and deaths: a global analysis. *Lancet. Planet. Health* **5**, E135–E144 (2021).

74. Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P. & Henrich, J. The church, intensive kinship, and global psychological variation. *Science* **366**, eaau5141 (2019).

75. Asch, S. E. Forming impressions of personality. *J. Abnorm. Soc. Psychol.* **41**, 258 (1946).

76. Bailey, A. H., LaFrance, M. & Dovidio, J. F. Is man the measure of all things? A social cognitive account of androcentrism. *Personal. Soc. Psychol. Rev.* **23**, 307–331 (2019).

77. Cuddy, A. J. C., Glick, P. & Beninger, A. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Res. Organ. Behav.* **31**, 73–98 (2011).

78. Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *Am. Econ. Rev.* **81**, 1068–1095 (1991).

79. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. Preprint at *arXiv* https://arxiv.org/abs/1406.5823v1 (2014).

80. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).

81. Blair, G., Cooper, J., Coppock, A., Humphreys, M. & Sonnet, L. estimatr: fast estimators for design-based inference, R package version 0.30.2 (2021).

82. Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. Emmeans: estimated marginal means, aka least-squares means, R package version 1.3 (2018).

83. Lüdecke, D. ggeffects: tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772 (2018).

84. Coppock, A. ri2: randomization inference for randomized experiments. https://cran.r-project.org/web/packages/ri2/ (2020)

85. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

86. Barr, D. J. Random effects structure for testing interactions in linear mixed-effects models. *Front. Psychol.* **4**, 328 (2013).

87. Bosman, J., Mervosh, S. & Santora, M. As the coronavirus surges, a new culprit emerges: pandemic fatigue. *The New York Times* (18 October 2020).

88. Santora, M. & Kwai, I. As virus surges in Europe, resistance to new restrictions also grows. *The New York Times* (9 October 2020).

89. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).

90. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).

91. Lakens, D. et al. Justify your alpha. *Nat. Hum. Behav.* **2**, 168–171 (2018).

92. Brysbaert, M. & Stevens, M. Power analysis and effect size in mixed effects models: a tutorial. *J. Cogn.* **1**, 1–20 (2018).

93. Green, P. & MacLeod, C. J. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**, 493–498 (2016).

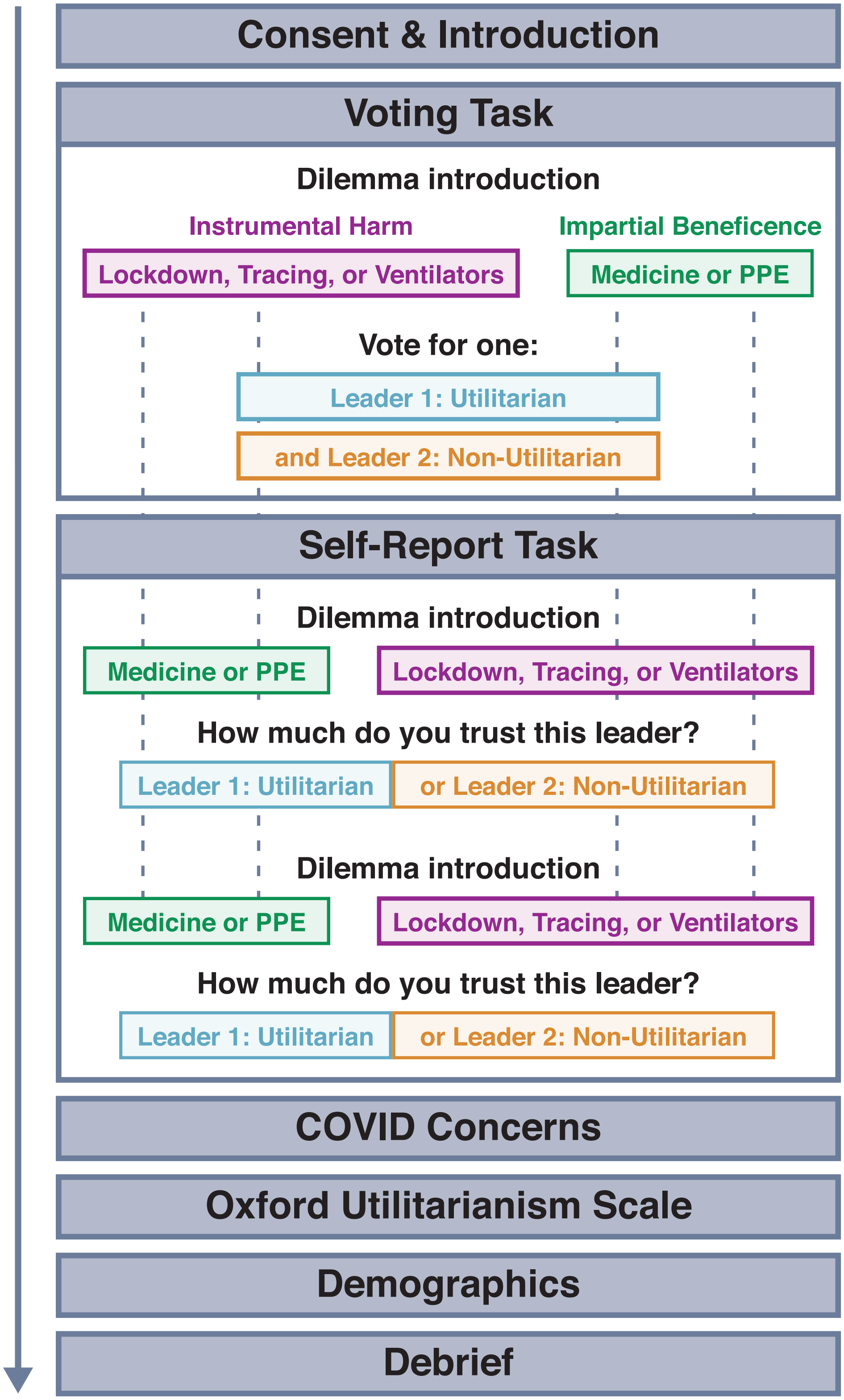
Acknowledgments

Pilot data collection was supported by Prolific Academic. Data collection for the main study was supported by grants from the Yale Tobin Center for Economic Policy (M.C.); the British Academy, Leverhulme Trust, and the Department for Business, Energy, and Industrial Strategy (SRG19\190050; J.A.C.E.); the Institutions for Open Science at Utrecht University (L.T. and M.S.), central internationalization funds of the Universität Hamburg and the Graduate School of its Faculty of Business, Economics and Social Sciences (B.B. and M.A.D.); and CAPES PRINT (88887.310255/2018 – 00; P.B.) and CAPES PROEX (1133/2019; P.B.). L.T. furthermore acknowledges funding from NWO grant (016.VIDI.185.017) and the National Research Foundation of Korea Grant, funded by the Korean Government (NRF-2017S1A3A2067636). H.S. was partly supported by the Research Council of Norway through its Centres of Excellence Scheme, FAIR project (262675). D.C. was partly supported by the National Research Foundation of Korea (NRF-2018R1D1A1B0704358). E.A., B.G., Y.L. and G.P. thank the University of Exeter Business School for funding their contribution to this research. N.S. gratefully acknowledges funding support provided by the Department of Management, Faculty of Management and Economics, Universidad de Santiago de Chile, and ANID FONDECYT de Iniciación en Investigación 2020 (Folio 11200781). A.L.O. and F.H. gratefully acknowledge support from the Independent Research Fund Denmark (0213-00052B and 8046-00034A) and the Faculty of the Social Sciences at the University of Copenhagen. N.R. was partly supported by the Israel Science Foundation (540/20). A.M.B.P. was supported by the ESRC. W.J.B. was supported by a postdoc fellowship from the National Science Foundation (#1808868). V.C. was supported by the National Science Foundation graduate research fellowship under grant no. DGE1752134. S.S. was partly supported by the National Research Foundation of Korea (NRF-2018R1C1B6007059). P.B. gratefully acknowledges support from CNPq (researcher fellowship 309905/2019-2). Y.M. gratefully acknowledges support from the National Natural Science Foundation of China (no. 31771204) and Major Project of National Social Science Foundation (19ZDA363). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

We thank members of the Crockett laboratory for feedback on previous drafts of this manuscript; J. Monrad for advice on scenario design; R. Gomila for statistical advice; J. Okoroafor, D. Shao, and X. Wang for assistance; and J. Apel, A. Bidani, N. Breedveld, R. Calcott, R. Carlson, L. Alfaro Cui cui, A. A. Gálvez, N. Kim, F. Michelsen, M. Meinert Pedersen, A. Mokady, A. Oline Ervik, J. Yang, X. Zeng, and M. Zoccali for assistance with survey translations.

Author Contributions Statement

M.J.C., J.A.C.E., C.C., V.C., and W.J.B. conceived the research. M.J.C., J.A.C.E., C.C., E.A., P.B., B.B., W.J.B., M.C., V.C., D.C., M.A.D., S.G., F.H., Y.M., M.A.M., C.M., A.L.O., A.M.B.P., N.R., N.S., J.S., W.S.-A., H.S., M.S., S.S., L.T., M.T., H.Y., and Y.Z. designed the research. M.A.M., J.S., M.J.C., J.A.C.E., C.C., H.S., L.T., N.S., and E.A. developed the voting task. J.A.C.E., V.C., M.J.C., C.C., and W.S.-A. wrote the moral dilemmas. C.C. conducted the power analysis in consultation with M.J.C., W.J.B., C.M., and N.R. C.C., J.A.C.E., M.J.C., W.J.B., C.M., and N.R. developed the analysis plan. C.C. analyzed the data in consultation with M.J.C., J.A.C.E., W.J.B., C.M., N.R., M.A.M., J.S., N.S., E.A., A.J., Y.S.S., and J.S.K. J.A.C.E., C.C., and M.J.C. prepared the manuscript with feedback from all co-authors. M.J.C., J.A.C.E., C.C., and C.K. coordinated the implementation of the project. M.J.C., J.A.C.E., C.C., E.A., P.B., B.B., M.C., D.C., M.A.D., S.G., B.G., F.H., C.K., J.S.K., Y.L., Y.M., M.A.M., F.M., C.M., A.L.O., G.P., N.R., N.S., J.S., Y.S.S., H.S., M.S., S.S., L.T., H.Y., and Y.Z. contributed to data collection and/or translation. All co-authors reviewed and approved the final manuscript.



Extended Data Figure 1. Overview of Experimental Design. Across subjects, we randomized the order of the voting and self-report tasks, the order of dilemmas in the self-report task, and the order of leaders in the voting task.