Supplementary Materials

Index

I. Supplementary Methods I: Stimuli selection

1. Stimuli selection procedure

To select the stimuli, we followed Pennycook and colleagues practical guide for behavioral research on misinformation (Pennycook, Binnendyk, et al., 2021). We first created a set of 210

brief news items (114 false news, 96 true news), each no longer than 140 characters, including spaces. The nature of these items was restricted to information with a cognitive utility, that is, presenting information relevant to concepts that individuals often think of or that could influence an individual's understanding of the state of the world. We avoided ego-relevant stimuli and those likely to elicit strong emotions or affect daily decision-making. The topics covered ecology, social justice, and democracy—key themes that were timely and relevant during the experiment but not directly related to participants' health, such as COVID-19. Some news items were sourced directly from French fake news debunking platforms (Les Décodeurs du Monde, AFP Factcheck, and Libération Checknews from 2017-2020), while others were fabricated based on content from these platforms.

We conducted a pretest to ensure the stimuli varied in content imprecision and were thematically consistent. After the pretest, we aimed to keep 96 counterbalanced news items, ensuring no statistical difference in content imprecision or thematic consensus between true and false news. The pretest, conducted on Testable.org and rewarded with $7, involved 55 independent raters (33 females, 22 males; mean age ± SD: 26.2 ± 4.78). Five groups of 11 French-speaking raters evaluated subsets of 42 news items each, assessing content imprecision, desirability, divisiveness, and themes.

Raters answered the following questions: 'On a scale of 0 to 10, to what extent would you say that the content of this information lacks precision? (0 = not at all imprecise - 10 = very imprecise)'; 'On a scale of 0 to 10, to what extent would you like to know more about the content of this information? (0 = I would not at all want to know more at all - 10 = I would very much want to know more); 'On a scale of 0 to 10, how divisive and how likely to divide opinion do you think the content of this information is? (0 = not at all likely to divide opinion - 10 = very likely to divide opinion); 'Which theme do you think best fits the content of this information: Ecology, Democracy, Social justice, Health, Economy, Education, Identity, Security, Travel, Freedom, None of the above'; 'What other theme do you think would best fit the content of this information?'.

Themes other than Ecology, Democracy, or Social Justice were distractors. We retained the 16 news items with the highest thematic agreement for each theme and veracity level, resulting in a final set of 96 counterbalanced true and false news items categorized as democracy-related, ecology-related, or social justice-related.

Inter-rater reliability (ICC2k) estimates and their 95% confidence intervals for content imprecision, consensuality, and desirability were calculated using the R package 'psych,' based on a mean-rating (k = 3), absolute-agreement, 2-way mixed-effects model. The average random raters' correlation coefficient was 0.54 for content imprecision (95% CI [0.4, 0.66]), 0.81 for consensuality (95% CI [0.75, 0.87]), and 0.59 for desirability (95% CI [0.46, 0.7]) (all $p<.001$). There was no significant difference between true (mean ± SD = 5.53 ± 1.24) and false (mean ± SD = 5.17 ± 1.25) news content imprecision distributions (ranksum, $p=0.09$), nor between true (mean ± SD = 6.24 ± 1.57) and false (mean ± SD = 6.55 ± 1.55) news divisiveness distributions (ranksum, $p=0.16$). Thus, content imprecision was balanced between false and true news, and both were equally divisive among raters. A (weak) correlation was found between stimuli desirability and participants' choices to seek more information ($Rhô=0.349$; $p<0.001$).

Sentiment analysis of each stimulus, using a multilingual XLM-roBERTa-base model trained by Cardiff NLP Group (Barbieri, Espinosa Anke, & Camacho-Collados, 2022) and performed with Python's 'transformers' package (Wolf et al., 2020), revealed that 93.75% of the news items were predominantly categorized as emotionally neutral. The model was trained on ~198M tweets and fine-tuned for sentiment analysis on eight languages, including French. The analysis was performed with the Python package transformers (Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, 2020). Out of the 96 stimuli, 90 had a probability score higher than 0.5 for the neutral dimension. In other words, 93.75% of the news were predominantly categorized as emotionally neutral.

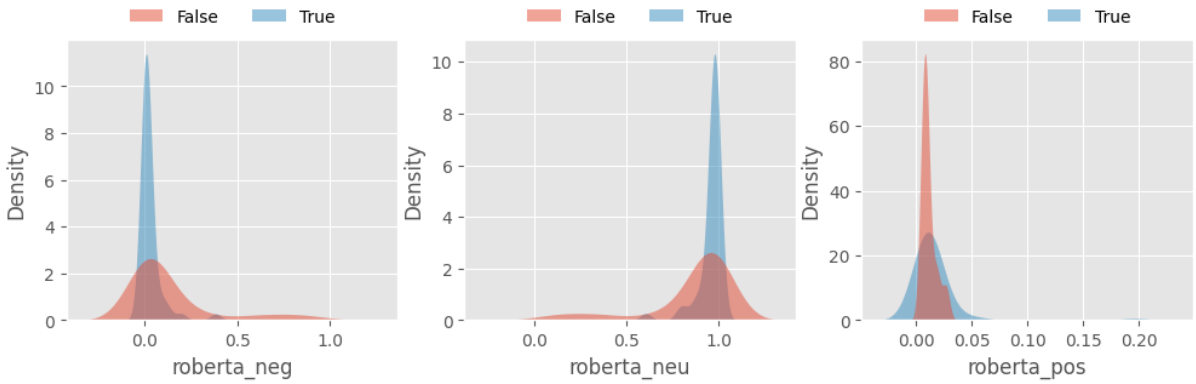


Figure S1: Most stimuli were classified as emotionally neutral. Density distributions show the sentiment-class probabilities assigned to each stimulus for the negative, neutral, and positive classes. Probabilities were obtained from the softmax output of a multilingual sentiment-analysis model implemented with the Python *transformers* package. Most stimuli received their highest probability for the neutral class, consistent with the stimulus-selection goal of avoiding strongly emotional material.

An MLM analysis was performed to investigate *receivers*' reception choices by modelling reception probability with *receivers*' behavior associated with the reception choice. Independent variables were the social distance to the trial's organization (i.e., the cue value provided to the participants). Only the social distance had a significant effect on reception probability (odds-ratio = .511, 95% CI [.297, .879], $p < .05$) which indicates that the more r*eceivers* were close to the organizations, the more they chose to receive.

2. Stimuli selection instructions (translated from French)

Welcome to this session.
We ask that you turn off your phone.

You will be participating in a session consisting of 42 periods.
This session will last approximately 40 minutes.
In each period, you will see a short piece of information. For each brief you will be asked to answer 5 questions. The first 4 questions will be, for each brief, mandatory.
The brief information will be different for each period.
Some of the news briefs are from the French written press for the period 2017-2020. Others have been fabricated.
Unless specified, their content is recent, topical and concerns the French territory.
Please read each brief and each question carefully, then take the time you need to answer.

II. Supplementary Methods II: Main experiment task instructions (translated from French)

Instructions of the experiment

Welcome to this experiment.

First, we ask you to turn off your phone.

Please read these instructions carefully. They give you all the information you need to participate in this experiment. If you don't understand something, don't hesitate to raise your hand. We will come and answer you.

At the end of this experiment, depending on your decisions, you can earn Experimental Currency Units (ECU). These ECUs will be converted into euros at a ratio of 100 ECUs = 1€.

At the end of the experiment, you will receive the monetary amount equivalent to the conversion of the ECUs you have earned plus the 60 euros initially provided for your participation.

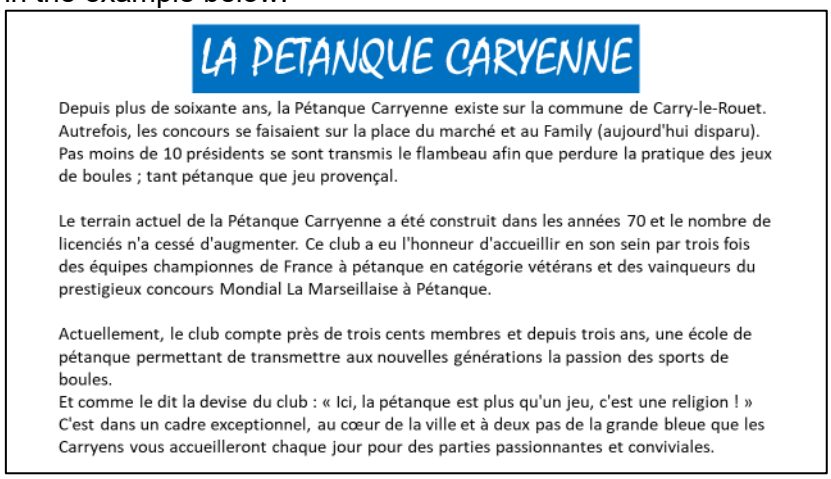
This session is composed of **2 parts.**

You will first receive instructions for the first part, then instructions for the second part. At the end of the task, we will ask you to fill in several questionnaires.

Your answers throughout the experiment will be and remain anonymous.

1. Part 1 of the experiment

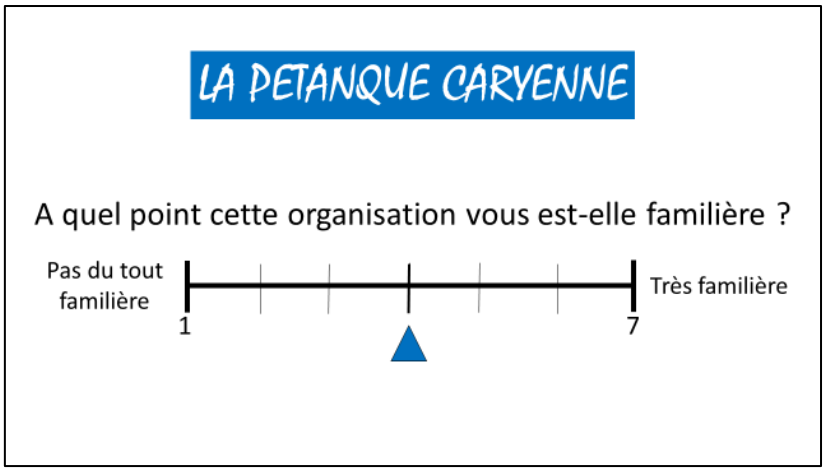
The first part of this experiment consists of evaluating 12 different organizations based on 6 criteria. You will first see the organization to be evaluated and a summary of its purpose from its website, as in the example below:



Then you will rate it on the following 6 criteria using a 7-point scale:

- How familiar is this organization to you?
- How close do you feel to this organization?
- How much do you like this organization?
- How familiar is this organization to those closest to you?
- How close do or would your family feel to this organization?
- How much do or would your relatives like this organization?

For example, below, 1 means that the organization is 'Not at all familiar' to you and 7 means that the organization is 'Very familiar' to you:



These evaluations will be conducted in private. They will not be visible to any other participant or experimenter. They will not be associated with any element that could identify you.

2. Part 2 of the experiment

There are two types of participants in the second part of the experiment: Participant A and Participant B. **You are a Participant A**. You will keep this role throughout this part. Participant B has completed their task in a **previous experiment**.

You will play with answers from these B participants.

There are 20 B participants. Participants B are all-comers. There is no way for you to recognize these B participants.

The answers you will play with have been drawn from all their answers. There is no way for you to know which participant B the answers you will play with come from.

There are 96 periods. E**ach period, you will be matched with a different B Participant**. You are likely to be paired with the same person several times. However, it is very unlikely that you will be associated with the same person several periods in a row. You will not be able to identify which participants you are associated with.

A. <u>Evaluation of the brief news:</u>

At the beginning of each period, you will see a short piece of information.

Each brief will be different and will come from different media. **For each news brief, there is a supplement**. The supplementary information consists of an investigative file. Their content is related to the content of the brief and can be related to other information around the content of the brief.

Warning: news items may be true or false. Some of the stories you will be exposed to and manipulate will be false information.

False information has been fabricated.

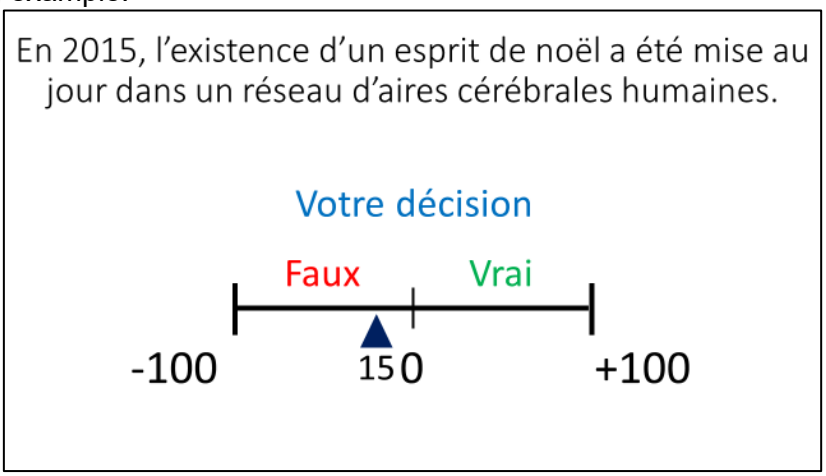
The non-fabricated news items are taken from the French written press over the period 2017 - 1st quarter 2020. The non-fabricated news items are all true.

Unless specified, their content is therefore recent, current and concerns the French territory at the time of the 1st quarter 2020.

At the end of the experiment, you will be debriefed on the information you have been exposed to.

After the brief information has been displayed, you will be asked to evaluate whether its content is true or false by specifying, “in your opinion, what is the number of chances out of 100 that this brief is true or false? “

To answer that a news item is true or false, you will drag the slider below the question, as in the following example:



- To answer that the content of the brief is false, you will drag the cursor between the **0 and -100 bounds: these bounds delimit the 'False' answer.**
- To answer that the content of the brief is true, you will drag the cursor between the **0 and +100 bounds: these bounds delimit the answer 'True'.**

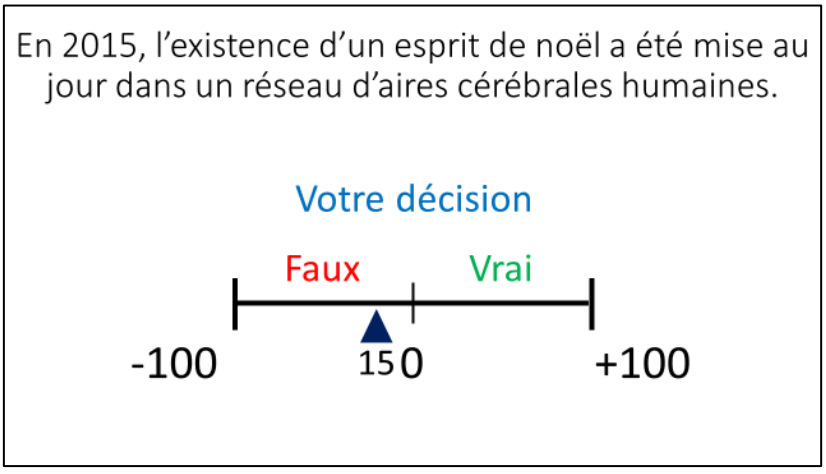
To help you answer this question, "robots" will be available during this step: There are 100 different robots. **Each robot has a** different **accuracy level** than the other robots. This accuracy level is the **number of chances out of 100 that the robot evaluates the short information correctly.**

This number is an integer Y between 1 and 100.

That is, there is one robot that has a 1 in 100 chance of answering correctly, there is one robot that has a 2 in 100 chance of answering correctly, and so on until the 100th robot that has a 100 in 100 chance of answering correctly. A robot that has 75 chances out of 100 to answer correctly will give a correct answer 75% of the time and will give a wrong answer 25% of the time.
At each period, the computer will draw one robot from the 100 robots. All robots have the same chance of being drawn. You will not know which one was drawn and the robot will change randomly each period.

You will have to choose which robots you would let answer the truthfulness question for you. To do this, you will decide how confident you are in your answer by choosing a threshold (a number X between 1 and 100) using the slider.

If the robot's Y number is **higher** than your X threshold, the robot will answer for you. If the robot's Y number is **lower than or equal** to your X threshold, **your answer will be submitted.**



Here, the participant answers that the content of the short information is wrong.
If the drawn robot has more than a 15% chance of answering correctly, the participant would prefer the submitted answer to be the robot's answer rather than their own.

The confidence threshold will be such that, for any robot with an accuracy level **higher than your threshold**, you would prefer the program to consider the **robot's answer** rather than yours.
For any robot with an accuracy level **lower than or equal to your threshold**, you would prefer the program to take **your answer** into account.

Thus, for example:

- If you choose 75 as your confidence threshold and the randomly selected robot has an accuracy level of 90, the program will consider that robot's answer. **The robot will have a 90 out of 100 chance of giving the right answer.**
- If you choose 75 as your confidence level and the randomly selected robot has an accuracy level of 20, the program will consider your answer.

You are therefore encouraged to say how correct you really think your answer is. The less sure you are of your answer, the more you are likely to choose a low threshold (i.e., if you think there is a 25 in 100 chance that your answer is correct, it is better to give a threshold of 25).

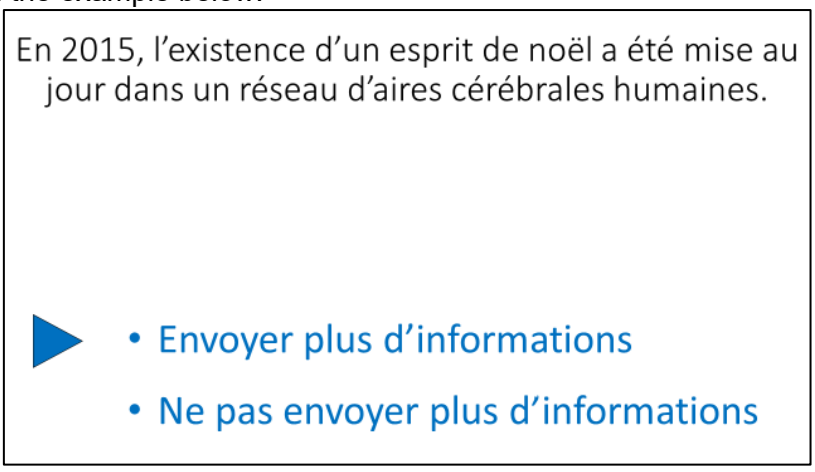
The program is indifferent to the sign of the confidence level. Choosing the numbers '75' or '- 75' will mean to the program that you think there is a 75 out of 100 chance that your assessment of the brief is correct. The direction of your response (positive or negative) is used to indicate your assessment of the brief ('False' or 'True').

Sixteen of your ratings out of 96 will be **randomly** selected at the end of the task. They will reward your performance. Your reward will be determined only on the randomly selected ratings. Each correct randomly selected assessment will earn you **50 ECUs**.

Participant B will have evaluated the same brief as you did, at the same time of the task as you did, and in the same way as you did.

B. <u>Sending Information:</u>

Each period, once you have evaluated the brief, you will be asked to **choose** between **sending** Participant B **additional information** about the content of the brief and **not sending anything**, as in the example below:



B Participants will have chosen between **receiving additional information** about the content of the brief and **not receiving anything**.

Sixteen of your sending choices out of 96 will be **randomly** selected at the end of the task. They will reward your performance in the following way:

For each of these sending choices,

- If your choice **matches** Participant B's choice, this period will **reward** you with 50 Experimental Currency Units (ECU).
- If your choice **does not match** the Participant B choice, this period will **not reward you** with ECUs.

B-Participants were asked to declare their desirability for each piece of information in the following manner:

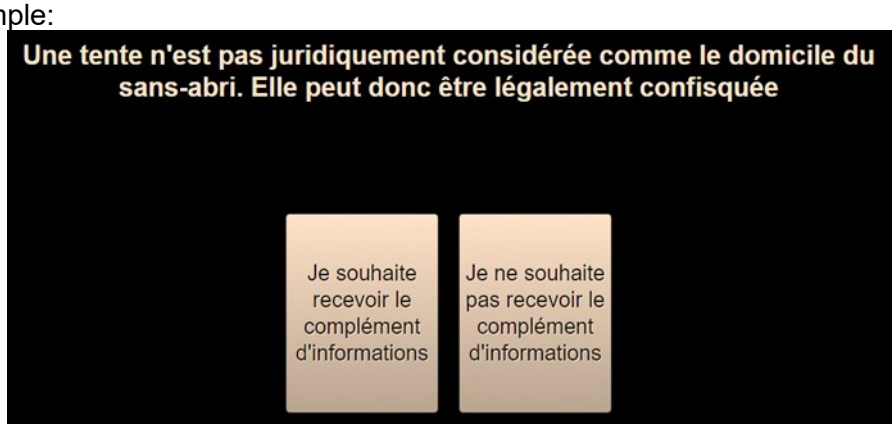
B-Participants were asked in **each period** to declare

1) their choice of reception (to receive or not to receive more information)

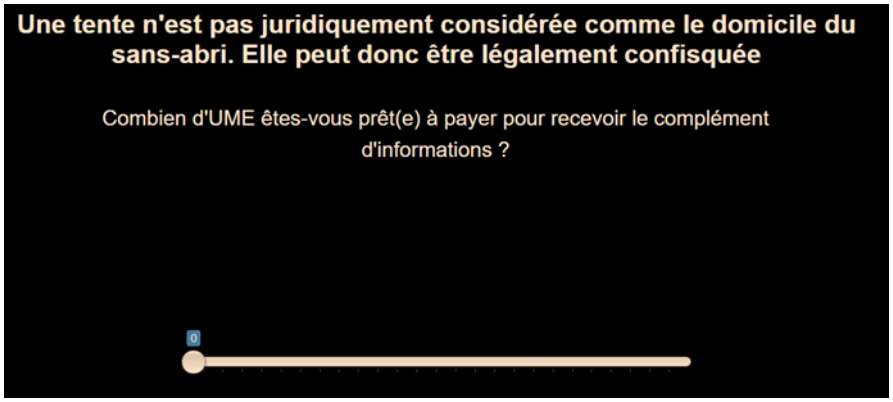
2) their willingness to see their choice fulfilled by indicating what fraction of their monetary allocation they were willing to spend to see their choice fulfilled.

B Participants thus had **monetary incentives** to reveal their **true preference**.

For example:



1) B Participants had to declare at each period their choice of reception (to receive or not to receive more information)



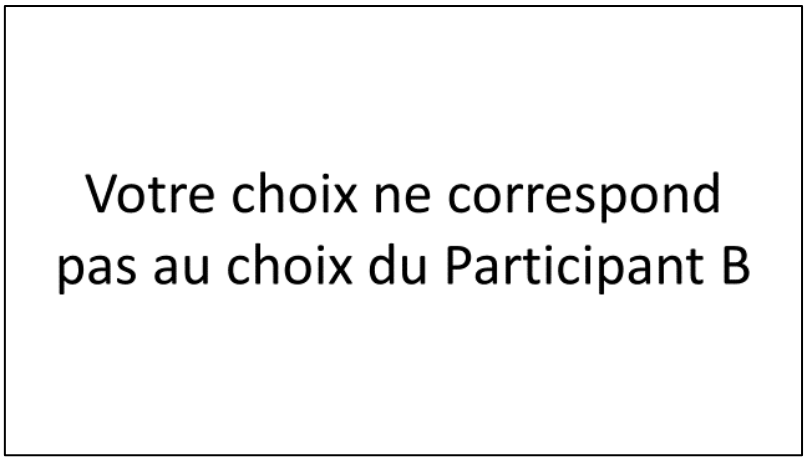
2) B Participants were then asked to declare their willingness to have their choice fulfilled by indicating what fraction of their monetary allowance they were willing to spend.

At the end of their task, a limited number of periods were randomly selected to make their choice. Their willingness to have their choice made then determined the chances that their choice would **actually happen**.

Their task was thus designed to lead them to choose, for each brief, whether they actually wanted to receive additional information.

C. <u>End of Period:</u>

You will know at the end of the period if your choice matches Participant B's choice:



D. <u>Information on Participant B :</u>

There are 2 types of periods:

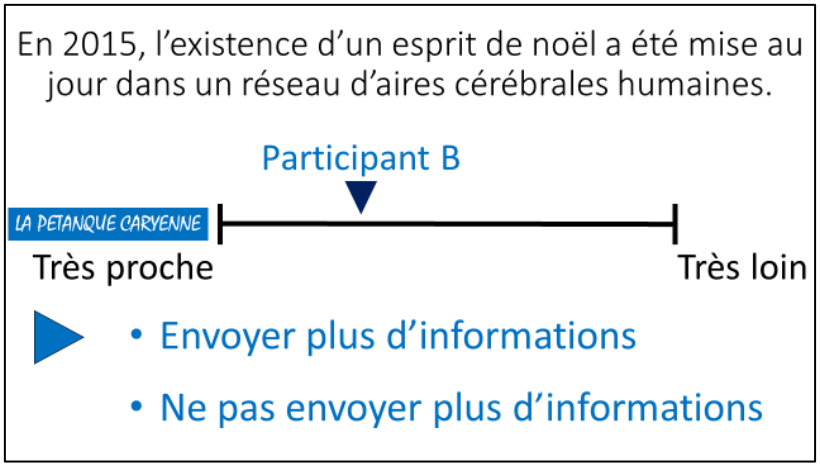
- The first half of the periods (48 periods) is characterized by the absence of information on Participant B.
- The second half of the periods (48 periods) is characterized by the display, in each period and at the time of the sending choice, of information about Participant B.

The information about Participant B will be displayed **during** your choice to send more information or not. It will take the form of the **distance** Participant B rated between him/her and a randomly selected organization.

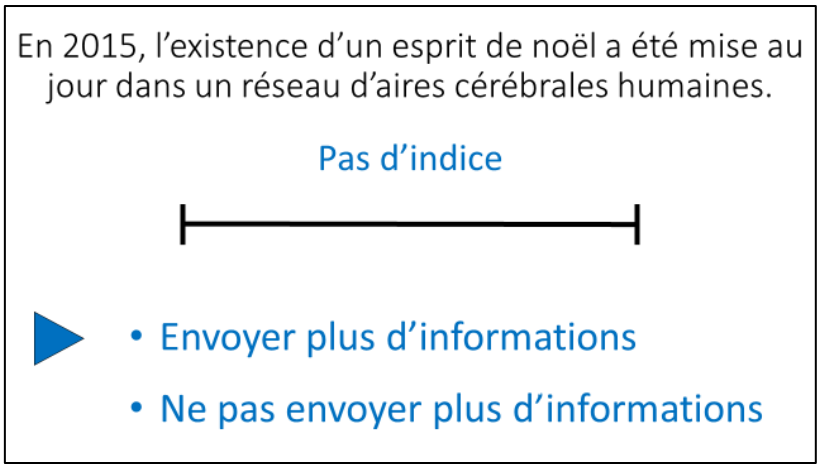
This organization will be **one of the ones** that you yourself have evaluated in the first part of this experiment. It will be related to the brief you have to choose to send or not to send.

The distance assessed corresponds to a projection on a scale of 0 to 100 of Participant B's answers to the **6 organization assessment questions**.

For example, here Participant B tends to feel close to the organization 'La Pétanque Carryenne':



Periods without information on Participant B will indicate this lack of information. They will take the following form:



E. End of Part 2:

Once this Part 2 is complete, we will ask you to complete several questionnaires. One of these will earn you several ECUs. You will be given instructions on these quizzes in due course.

Finally, you will be debriefed on the information and receive the verified version of the information you were exposed to.

3. Calculation of Final Compensation

A. Information Evaluation:

At the end of the experience, sixteen periods will be drawn from the 96 periods in Part 2. They will reward your judgment of truthfulness.

You will be rewarded, for each period drawn, as follows:

- If your answer is selected for review,

If your answer is correct, that period will be rewarded.

In case your answer is incorrect, that period will not be rewarded.

- If the drawn robot has a probability of answering correctly that is higher than your confidence level, the **probability** that this period will **reward** you will be the **probability Y** that the robot **answers correctly**.

A correct answer, either yours or the robot's, will get you 50 ECUs. An incorrect answer will get you 0 ECUs.

B. Sending information:

At the end of the experiment, sixteen more periods will be drawn from the 96 periods in Part 2. These will reward your performance in making a choice that matches Participant B's choice.

A choice that matches Participant B's choice will earn you 50 ECUs.

A choice that does not match Participant B's choice will earn you 0 ECUs.

C. Questionnaires:

At the end of the experiment, one of your decisions on one of the questionnaires will be drawn.

You will win the amount of ECUS that you have allocated for this decision.

D. Final Remuneration:

The ECUs will be converted into Euros at a ratio of 100 ECUs = 1€. The compensation for your performance and the compensation for answering the questionnaires will be added to your initial sum of 60 euros.

4. Experimentation

You will start the experiment with a training phase. Please ask your questions at this point in the task.

The information, themes and clues you see during the training phase will not be present during the experimentation phase. The answers you play with during the practice phase **will not be taken from** the answers of the 20 B participants. They will all have been chosen **at random** to allow you to familiarize yourself with the task.

Attention: **The task is constrained by a limited time to answer.** This constraint exists during the training phase and during the experimentation phase. You will become familiar with this time during the training phase. Throughout the task, you will have less than 10 seconds to respond to each decision. Each trial that you do not answer in time will come back at the end of the task.

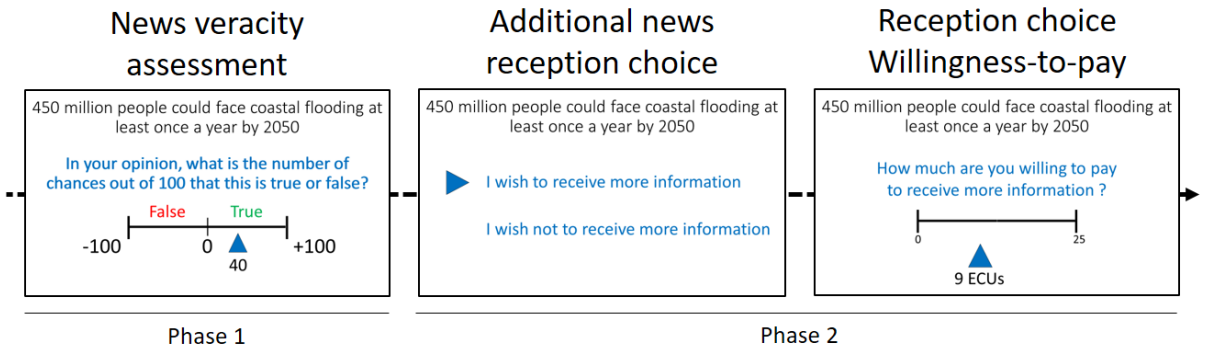


Figure S2. Veracity-judgment and information-choice stages from the *Receiving News* experiment. Participants first read brief news items and reported the probability that each news was true or false, allowing simultaneous assessment of veracity judgment and confidence. Correct evaluations (true news judged as true or false news judged as false) were rewarded with 50 ECU, while incorrect evaluations earned 0 ECU. Eight trials out of forty-eight were randomly selected for payment. In the second stage, participants chose whether to receive or avoid receiving additional information about the same news and indicated how much they were willing to pay (0–25 ECU, endowment = 200 ECU) to implement that choice. A Becker–DeGroot–Marschak (BDM) procedure determined whether and at what cost their choice was realized, ensuring that both information-seeking and information-avoidance carried potential cost. The figure is reproduced with the authors' authorization to illustrate the *Receiving news game*.

III. Supplementary Methods III : Presentation of the organizations (translated from French)

1. Democracy related organizations

France FREXIT:

France FREXIT is a private and independent initiative, created in March 2018 and aiming at informing, gathering, exchanging, proposing on the theme of Frexit. In all legality, without any violence, and in the respect of the Institutions. Taking advantage of the historical opportunity of the Frexit, FRANCE FREXIT proposes the complete reform of the Republic and the French State by a great FRENCH NATIONAL COORDINATION, based on a new formula of Power, including a truly democratic organization taking the best of the Republic and leaving the least good, and using modern means of communication: votes, electronic votes, draws, as well as some monarchical aspects, mainly in terms of spirituality and transmission of universal and traditional values of France. FRANCE FREXIT is politically opposed to the euro-extremist and euro-identitarian parties such as LaREM, LR, MODEM, EELV, PS, UDI, Parti Radical. FRANCE FREXIT does not support the euro-alternative parties and other decoy parties such as the RN, the LFI or DLF, the NPA, LO and some other euro-compatible parties.

Parti Libertarien:

Our observation is the same as the majority of French people: the weight of the state and its scope of action are constantly increasing, hindering our freedoms more and more, with the catastrophic results that everyone can see. We are also facing an extremely worrying legislative inflation, the accumulation of standards, regulations and laws make the system incomprehensible for the majority of French people and impracticable for entrepreneurs. Only large groups and multinationals benefit from this complexity and can expand without real competition. These regulations paralyze any personal initiative, block the social elevator and increase inequalities. Beyond the dramatic consequences of such an intrusion of the state in the life of individuals, the latter, by wanting to regulate every aspect of our daily life, goes beyond its prerogatives and violates our fundamental right to manage our life as we see fit. We are libertarians and we consider that it is up to free individuals to write their own history.

Le Mouvement Européen – France:

The European Movement - France has been mobilizing since 1950, across all generations, to bring to life a pluralist public debate on Europe. It deploys its activities around pedagogy, the organization of debate between citizens and the formulation of proposals to build Europe. It gathers thousands of volunteers gathered in more than 50 local sections, about twenty member organizations as well as a college of qualified personalities. Heir to the spirit of the founding fathers of Europe, the Movement is the first actor of civil society in France on European issues. It also mobilizes through its youth branch, the Young Europeans - France, which has 26 local groups throughout the territory. At the European level, the European Movement is also a member of the European Movement - International, a network of 39 organizations that make our commitment resonate throughout the continent. The European Movement-France is recognized as an association of general interest since July 22, 2016 and is also approved as an "educational association complementary to public education".

Fondation Robert Schuman:

Created in 1991 and recognised as a public utility, the Robert Schuman Foundation works in favour of European construction. As a reference research center the Foundation develops studies on the European Union and its policies and promotes their content in France, Europe and abroad. The Foundation is an open and multinational network. Its main mission is to keep alive the spirit and inspiration of one of the "Fathers of Europe", Robert Schuman, and to promote European values and ideals both within and beyond the borders of the Union. The Foundation produces numerous studies on European policies which constitute a valuable source of information for all those who want to understand European issues and challenges. Its independence allows it to deal with all current issues in an in-depth and objective manner. Its studies and analyses provide European decision-makers with information, arguments and food for thought that are appreciated for their usefulness and scientific quality. It multiplies initiatives in the field to advance the European democratic model.

2. Ecology-related organizations

Greenpeace:

Since its creation some 50 years ago, Greenpeace has acted on land and sea according to the principles of non-violence to protect the environment and promote peace. Today, we remain faithful to this mission, as well as to our total financial and ideological independence. Climate change, growing inequality, social injustice, migration and armed conflict... All the major challenges of our time, to which we must urgently respond, are intimately linked - as are the power structures that create them and the mentalities that accommodate them. This is why it is necessary to transform them together. Greenpeace is present in 55 countries, on all continents and oceans, through its 28 national and regional offices and its three boats. It has more than three million members and over 36,000 volunteers worldwide. We place citizen power at the heart of our campaigns by giving resonance to the work of all those who share

our vision, our hopes and our conviction that we need profound transformations in our societies.

WWF:

WWF is one of the world's leading independent environmental organizations. WWF works to stop the degradation of the planet's natural environment and to build a future in which humans live in harmony with nature, conserving the world's biological diversity, ensuring the sustainable use of renewable natural resources, and promoting the reduction of pollution and waste. Since 1973, WWF France has been carrying out concrete actions to safeguard natural environments and their species, promote sustainable lifestyles, train decision-makers, support companies in reducing their ecological footprint, and educate young people. WWF France, a public utility foundation, works for a living planet from Paris, Marseille, the Alps, Guyana and New Caledonia. WWF is committed to action based on dialogue and respect for others, and adopts a global approach that takes into account the interdependence between the state of the planet and human development.

Le Groupe d'experts non-gouvernemental sur l'évolution du climat (NIPCC):

The Non-Governmental Panel on Climate Change (NIPCC) is an international group of non-governmental scientists and academics who have come together to present a comprehensive, reliable and realistic assessment of the science and economics of global warming. Because it is not a government agency, and because its members are not predisposed to believe that climate change is caused by human greenhouse gas emissions, the NIPCC is able to offer an independent "second opinion" to the evidence reviewed - or not reviewed - by the Intergovernmental Panel on Climate Change (IPCC) on the issue of global warming. Since its founding in 2008, the NIPCC has been producing publications and reports for public policy. These reports aim, for example, to show that the impact of human-induced global warming is benign and could be beneficial to humanity and the natural world; that the evidence for rising sea levels is unreliable; or that there is no scientific consensus in the climate change debate.

L’association des climato-réalistes:

Appeared in France in 2015, climatorealism sees the climate as an object of science and not ideology. Climate change is multiple and poorly understood, so there is no evidence that our way of life would cause "climate disruption." To say so is not selfishness, denial or anti-environmentalism, but realism. We need to think about how best to use our resources, and put our efforts where they really matter. The purpose of the association of climato-realists is to promote an open and free debate on the evolution of the climate and the societal and environmental issues related to it, by encouraging the expression of rigorous and well-founded opinions in all its forms. The association aims to make citizens aware of the stakes of climate and energy policies conducted in the name of the fight against global warming. The association is apolitical and totally free in the expression of its ideas. It strives to disseminate reliable information gathered from serious sources.

3. Social justice – related organizations

SOS Méditerranée:

SOS Méditerranée is a European civil sea rescue association, independent of any political party and any religion, created in 2015 and made up of citizens mobilized to face the humanitarian emergency in the Mediterranean. SOS Méditerranée is based on the respect of man and his dignity, whatever his nationality, origin, social, religious, political or ethnic affiliation. The association's vocation is to provide assistance to any person in distress at sea who is within the scope of its action, without any discrimination. The persons concerned are men, women or children, migrants or refugees, who are in danger of death when crossing the Mediterranean Sea. The association also aims to ensure the protection of the survivors until their arrival in a safe port and to bear witness to the situation in the central Mediterranean. The

association is financed by private donations and public grants. The funds collected are allocated to the rental of the boat, daily maintenance and rescue costs.

FEMEN:

FEMEN is an international movement of feminist political activists with bare torsos, painted with slogans, and heads crowned with flowers. Our slogans are short and punchy; our chests are our banners. From the militant necessity is born the accomplishment of powerful and provocative but always non-violent actions. The movement was born in 2008 in Kiev, Ukraine. Since 2010, the activists are politicizing and using their breasts as a support for their demands. With our provocative and resounding actions, we target the multiple manifestations of the patriarchal order: dictatorships, sex industry and religions. We are a female revenge against the sclerotic patriarchal culture. We are an expression of freedom and pluralism. Our ambition is to change mentalities and the public image of women by exposing our strength, our courage and our convictions. It is by developing our political action that we will succeed in changing even our most intimate reality.

Génération Identitaire:

Génération Identitaire is a political youth movement that brings together boys and girls across Europe. It was founded in September 2012. We call on young people to raise their heads: in the face of scum, in the face of those who want to control our lives and our thoughts, in the face of the standardization of peoples and cultures, in the face of the tidal wave of massive immigration, in the face of a school that hides the history of our people from us to prevent us from loving it, in the face of a so-called living together that turns into a nightmare... Génération Identitaire is the front line of resistance. Aware of the challenges we face, we do not refuse any battle. Proud of our heritage and confident in our destiny, we have only one watchword: we will not back down! We are the sacrificed generation, but not the lost generation. For we are going to war against all those who want to tear us away from our roots and make us forget who we are. Our ideal is reconquest, and we will carry it out to the end. Génération Identitaire is the barricade on which the youth in struggle for its identity stands.

La Manif Pour Tous:

La Manif Pour Tous defends marriage and filiation in coherence with the sexual reality of humanity, whose consequence is both the difference and the complementarity of the sexes, which is essential to conceive a child and to assume the difference between father and mother, fatherhood and motherhood. Our goal is the respect of the superior interest and the elementary needs of the child, today threatened by the societal reforms inspired by the gender ideology. La Manif Pour Tous takes a pragmatic approach whose objective is to promote the well-being and the future of the child, the adult and society as a whole, what is commonly called the general interest. It acts for present and future generations. Finally, La Manif Pour Tous reminds us that the family is the crossroads of the difference between the sexes and the difference between generations. Only this context meets the essential needs of the child to come into the world, to know its personal identity, to enter little by little in relations with the others, to insert itself in the group then in the society, to contribute to the social peace.

IV. Supplementary Methods IV : Pre-task questionnaire

1. Comprehension questionnaire

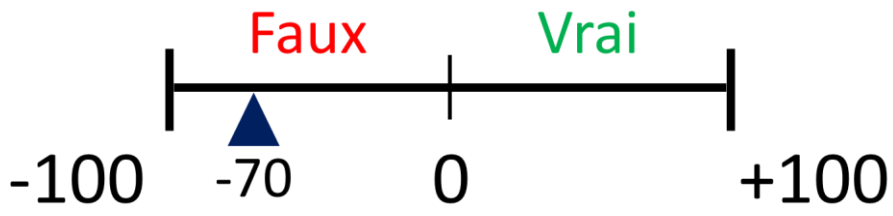
Please answer the following comprehension questions. Please call the experimenter if you have any questions.

1. General Questions

What is your goal during the information evaluation step?
What is your goal during the sending of additional information step?
How many B Participants are there? How were they chosen?

2. Information evaluation case n°1

Consider the case where you answer that a piece of information is false and you drag the slider to the value -70. If the randomly drawn robot has a 50% chance of answering correctly:



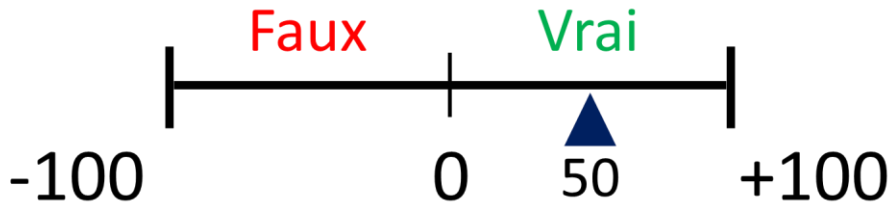
What evaluation will be submitted?

A. The submitted evaluation will be yours
B. The submitted evaluation will be the robot's

Which answer will be submitted between True and False?

3. Information evaluation case n°2

Consider the case where you answer that a piece of information is true and you drag the slider to the value 50. If the randomly drawn robot has a 75% chance of answering correctly:



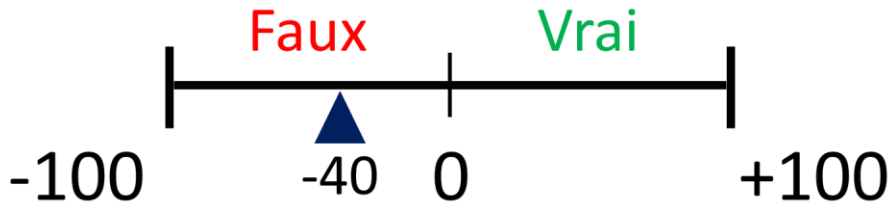
What evaluation will be submitted?

A. The evaluation submitted will be yours
B. The submitted evaluation will be the robot's

What will be the submitted answer if the submitted evaluation is the robot's?

4. Information evaluation case n°3

Consider the case where you answer that a piece of information is false and you drag the slider to the value -40. If the randomly drawn robot has a 60% chance of answering correctly:



What evaluation will be submitted?

A. The evaluation submitted will be yours
B. The submitted evaluation will be the robot's

What will be the submitted answer if the submitted evaluation is the robot's?

5. Information transmission case n°1

Consider the case where you choose not to send more information and Participant B has chosen not to receive more information.

Is the period a failure?
How many ECUs is this period worth?

6. Information transmission case n°2

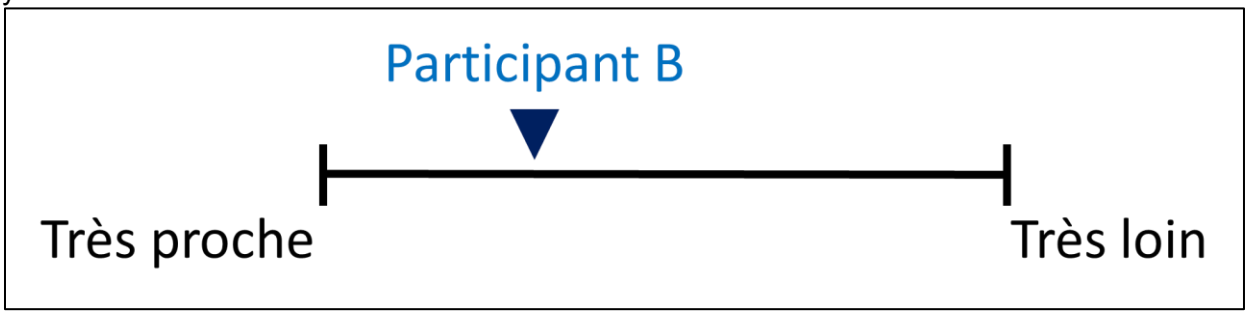
Consider the case where you choose not to send more information and Participant B has chosen to receive more information.
Is the period a failure?
How many ECUs is this period worth?

7. Information transmission case n°3

Consider the case where you choose to send more information and Participant B has chosen not to receive more information:
Is the period a failure?
How many ECUs is this period worth?

8. Information about Participant B

Let's consider the case where the following information about Participant B is provided to you before the choice to send information:



Participant B sees himself or herself as rather:

A. Very close to the organization
B. Moderately close to the organization
C. Neither close nor far from the organization
D. Somewhat distant from the organization
E. Very far from the organization

V. Supplementary Methods V : Post-task questionnaires

1. 11-20 game

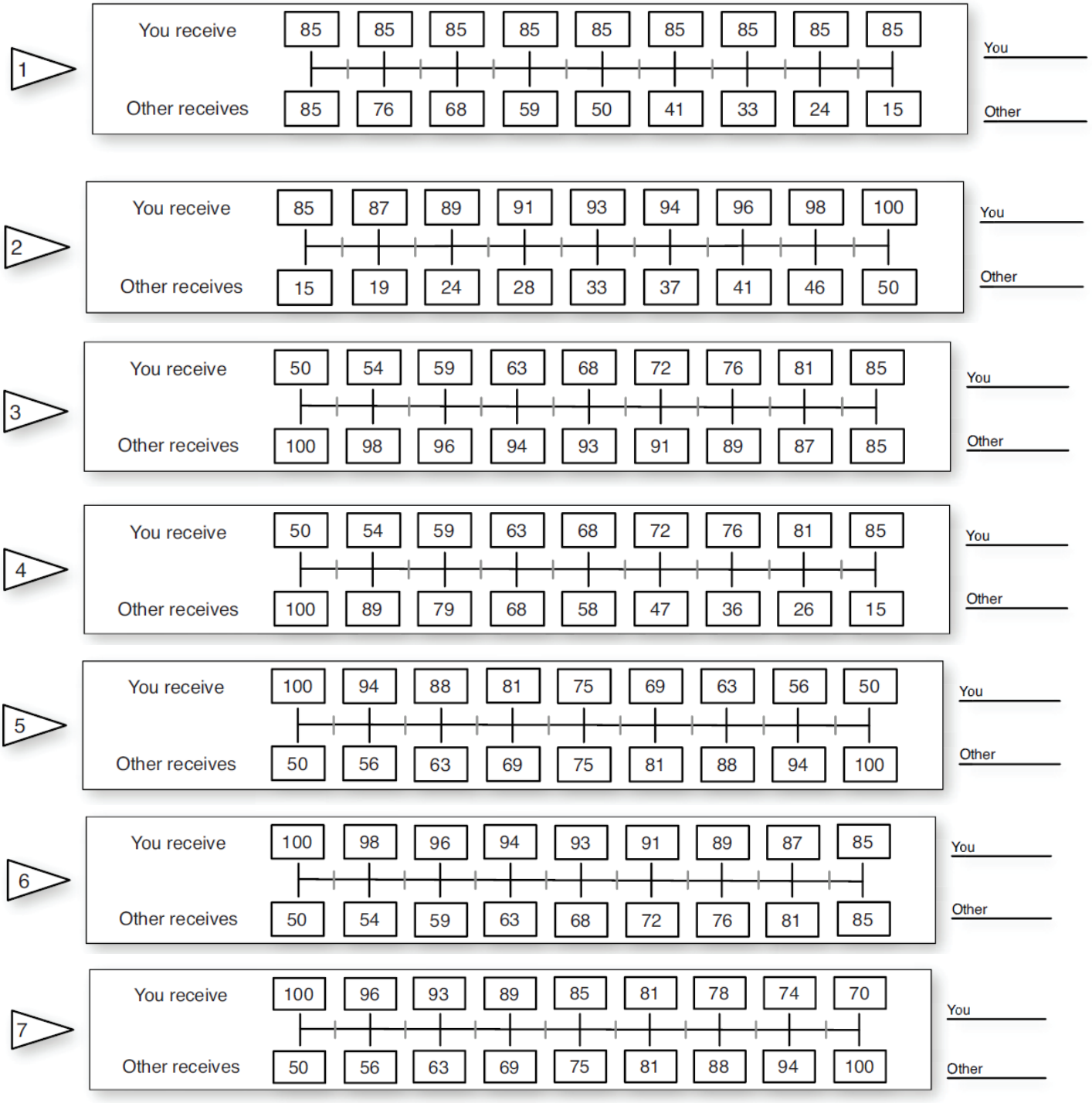
You and another person will participate in a game in which each player requests an amount of ECU. The amount must be a whole number between 11 and 20. The other person will be selected from the participants in this study. Each player will receive the amount he/she requests. A player will receive an additional 20 ECUs if he/she requests exactly 1 ECU less than the other player. The total amount will be added to the amount of ECUs earned so far. You will only participate in this game once and with only one other person.

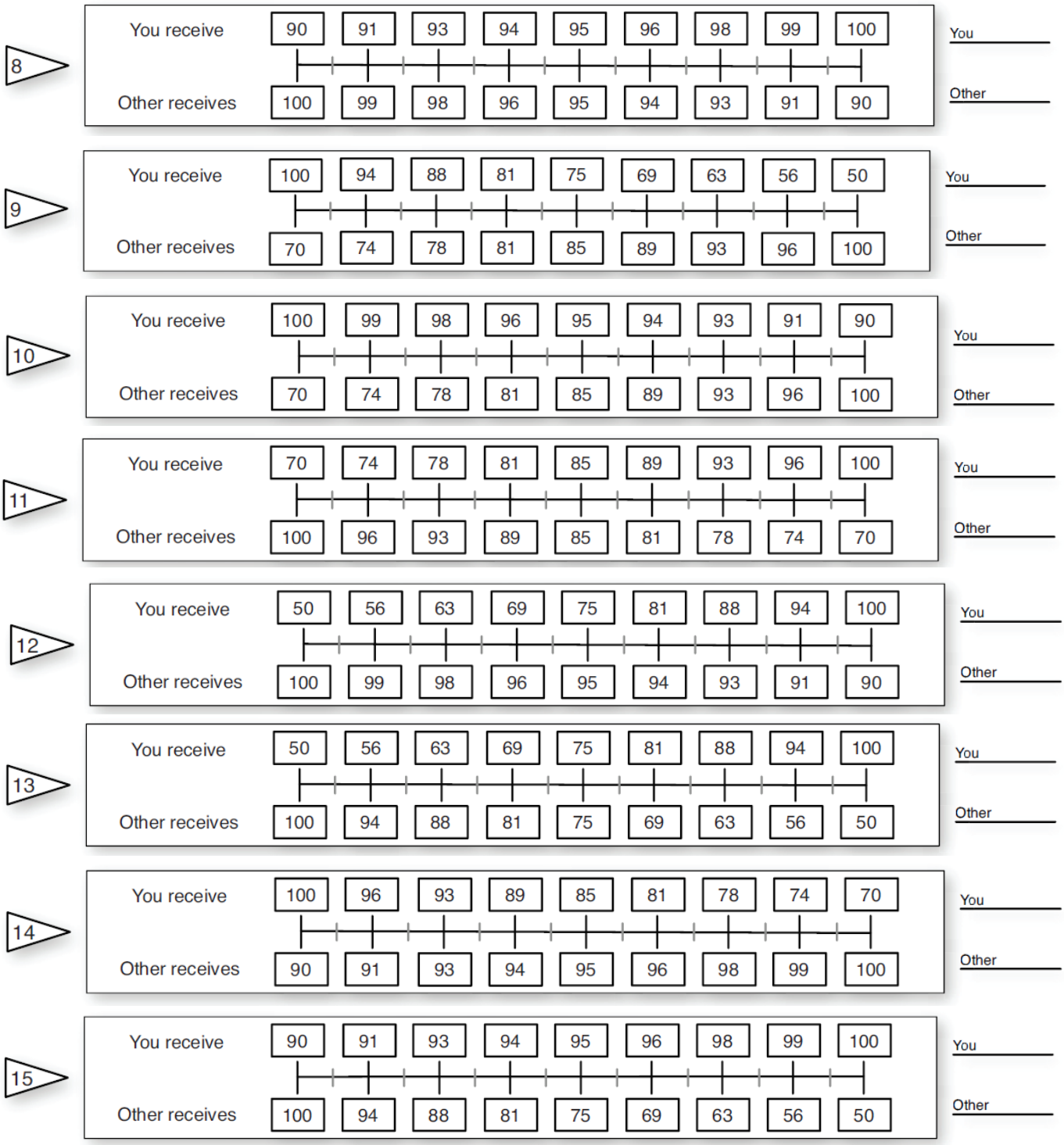
What amount are you requesting? Please answer as soon as possible.

2. Murphy's slider (SVO)

In this task you have been randomly paired with another person, whom we will refer to as the other. This other person is someone you do not know and will remain mutually anonymous. All of your choices are completely confidential. You will be making a series of

decisions about allocating resources between you and this other person. For each of the following questions, please indicate the distribution you prefer most by marking the respective position along the midline. You can only make one mark for each question. Your decisions will yield money for both yourself and the other person. In the example below, a person has chosen to distribute money so that he/she receives 50 dollars, while the anonymous other person receives 40 dollars. There are no right or wrong answers, this is all about personal preferences. After you have made your decision, write the resulting distribution of money on the spaces on the right. As you can see, your choices will influence both the amount of money you receive as well as the amount of money the other receives.





3. Manip check questionnaire

1. What was your strategy for choosing the number of chances out of 100 that the information was true or false?
2. What was your strategy for choosing what additional information to send?
3. How did you guess whether the receiver wanted or did not want to receive more information?
4. Did you understand that you were playing each trial with a different receiver's response, taken from all 20 receivers?

VI. Supplementary Methods VI : Behavioral analyses

1. Truthfulness estimation

Average response time to estimate truthfulness was 8.82 ± 1.52 seconds overall, with slower responses in the Control condition than in the Cue condition (Control: 9.03 ± 1.45; Cue: 8.61 ± 1.57). Participants estimated truthfulness with a bias: 61.61 ± 19.37% of information was declared as true overall (Control condition: 62.19 ± 18.41%; Cue condition: 61.04 ± 20.32%). They were not better than chance at this estimation, with performances centered around 50.98% success (sd = 4.18) and average success rate higher in the Control condition (52.48 ± 21.95%) than in the Cue condition (48.85 ± 23.11%).

Average confidence value was 49.9 ± 14.35. We looked at the impact of truthfulness judgment and news item features on confidence degree. Effect of truthfulness judgment was close to being statistically significant (estimate=1.48, se=.77, p=.054), and neither truthfulness nor theme (all *p*>.1) explained confidence degree (Table S1). We further tested whether confidence differed between conditions within each truthfulness judgment using Bayesian normal models with $\mu \sim N(50,10)$. For trials judged as true, the posterior Control-Cue difference in confidence was close to zero, with a delta δ = -0.016 (95% CrI [-7.234, 7.238]). That posterior difference was also close to zero for trials judged as false (δ = -0.489, 95% CrI [-7.727, 6.715]).

Therefore, despite a bias towards the true, participants' confidence in information truthfulness estimation did not differ reliably between judgments, conditions or news features. Table S1 summarizes the fixed-effect estimates from models of judgment and confidence.

Table S1: Summary of judgment and confidence MLMs.

	Judgment			**Confidence**		
Predictors	*Odds Ratios*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	1.67	1.21 – 2.31	**0.002**	-0.03	-0.52 – 0.45	0.887
Confidence	1.12	0.96 – 1.30	0.162			
Judgment: False	*Reference*			*Reference*		
Judgment: True				0.09	-0.03 – 0.20	0.138
Truthfulness: False	*Reference*			*Reference*		
Truthfulness: True	0.98	0.76 – 1.26	0.854	0.08	-0.04 – 0.21	0.178
Theme: Democracy	*Reference*			*Reference*		
Theme: Ecology	1.23	0.95 – 1.59	0.114	0.07	-0.03 – 0.17	0.169
Theme: SocJust	0.89	0.69 – 1.14	0.350	0.23	0.13 – 0.33	**<0.001**
TruthfulnessTrue:Confidence	1.04	0.90 – 1.22	0.578			
TruthfulnessTrue:JudgmentTrue				-0.04	-0.16 – 0.08	0.544
TruthfulnessTrue:ThemeEcology	1.13	0.78 – 1.62	0.528	0.01	-0.14 – 0.15	0.931
TruthfulnessTrue:ThemeSocJust	1.28	0.89 – 1.84	0.178	-0.24	-0.39 – -0.10	**0.001**
Age	0.99	0.89 – 1.09	0.783	-0.01	-0.22 – 0.20	0.935
Education	0.99	0.92 – 1.06	0.755	-0.02	-0.15 – 0.10	0.742
Sex: Female	*Reference*			*Reference*		
Sex: Male	0.87	0.68 – 1.11	0.260	-0.09	-0.55 – 0.36	0.691
Random Effects						
σ^2	3.29			0.66		
τ_{00}	0.14 $_{Subject}$			0.37 $_{Subject}$		
τ_{11}	0.08 $_{Subject.Confidence}$			0.05 $_{Subject.JudgmentTrue}$		
ρ_{01}	1.00 $_{Subject}$			-0.16 $_{Subject}$		
ICC				0.36		
N	32 $_{Subject}$			32 $_{Subject}$		
Observations	3072			3072		
Marginal R^2 / Conditional R^2	0.014 / NA			0.012 / 0.368		

After establishing that Senders' confidence in their truthfulness judgments did not differ reliably across judgments or conditions, we compared Senders and Receivers on truthfulness judgments confidence.

Receivers judged slightly more items as true than Senders (61.98 ± 12.20% versus 61.20 ± 9.83%). We compared truthfulness judgments using beta-binomial models. The posterior difference in the probability of judging an item as true favored Receivers, δ = .03 (95% CrI [.006, .064]).

Receivers also reported higher average confidence than Senders (61.11 ± 12.28 versus 49.90 ± 14.35), corresponding to an 11.2-point raw difference. We compared Senders' and Receivers' confidence values using Bayesian normal models. The sender-receiver confidence comparison was sensitive to prior specification. Under $\mu_1, \mu_2 \sim N(50,10)$ and common variance, the estimated difference was close to zero, δ = -0.12 (95% CrI [-0.07, 0.05], DIC = 432.2, Cohen's d = -0.008). Under $\mu_1, \mu_2 \sim Uniform(0,100)$ with $S \sim half-t(scale = 10)$, the estimated difference was larger, δ = -11.18 (95% CrI [-18.91, -3.25], DIC = 423.2, Cohen's d = -0.824). Because the conclusion depends on prior specification and both groups are small, we do not treat this comparison as robust evidence for a sender-receiver difference in confidence (Figure S3). The difference in values may partly reflect differences in the response scales used across the two tasks: Senders reported confidence in 10-point bins, whereas Receivers could report any integer value from 1 to 100.

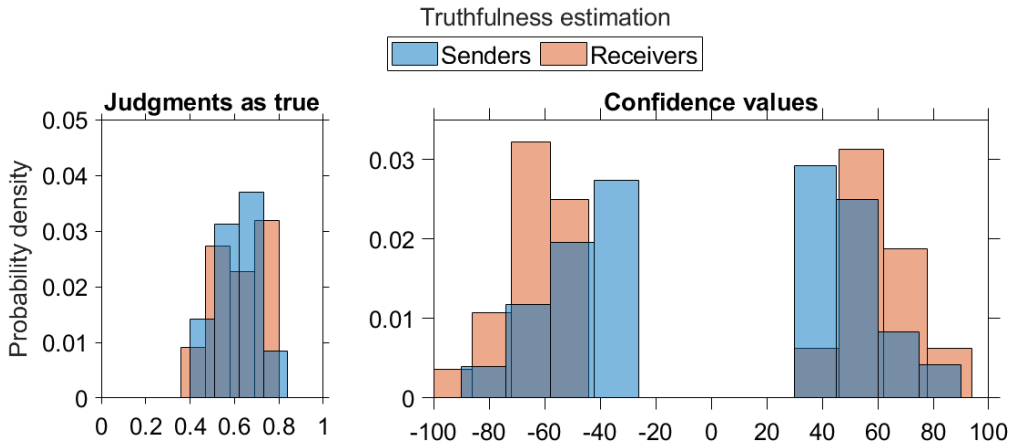


Figure S3: Plots of senders' proportion of judgments as true and confidence values against receivers. Senders and receivers reporting binomial judgments of news truthfulness. Senders reported confidence in 10-point bins, whereas Receivers could report any integer value from 1 to 100.

2. Choices to send

In the Cue condition, Receiver-specific information was provided as the Receiver's social distance from an organization associated with the news domain. Higher values indicated greater distance from the organization. The mean cue value across Cue trials was 43.35 ± 21.51. Cue values differed descriptively across domains, with mean values of 64.73 ± 10.94 for the democracy-related organization, France FREXIT; 30.94 ± 23.02 for the ecology-related organization, NIPCC; and 36.18 ± 10.96 for the social-justice-related organization, FEMEN. The mean absolute Sender–Receiver distance difference was 28.80 ± 13.14 overall, with values of 31.75 ± 12.73 for FREXIT, 23.92 ± 8.07 for NIPCC, and 30.74 ± 16.04 for FEMEN. Thus, democracy-related Cue trials involved greater average social distance than ecology- or social-justice-related Cue trials. These descriptive statistics characterize the cue environment; the main behavioral analyses test whether trial-by-trial Receiver social distance predicted Senders' choices.

Choices to send supplementary information were analyzed using mixed-effects logistic models predicting trial-by-trial send decisions. These models tested how condition, truthfulness judgment, confidence, cue value, trial order, and their interactions related to the probability of sending the news. The main effect of confidence on sending probability interacted with truthfulness judgment and with condition (all $p < .05$). Figure S4 illustrates the fixed effects of confidence value, interacting with judgment and condition, on sending probability. Table S2 summarizes the fixed-effect estimates from these models.

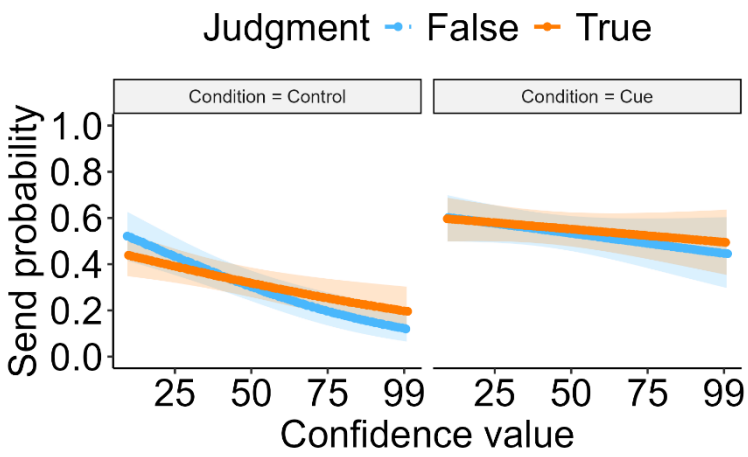


Figure S4: Linear fixed effects of confidence value on sending probabilities for news judged as true and false in Control and Cue conditions. Predicted sending probabilities were estimated from mixed-effects logistic regression models of trial-by-trial sending choices. Continuous predictors were mean-centered and scaled before model fitting. For visualization, predicted probabilities were back-transformed and displayed on the original task scales. In the Control and Cue conditions, lower confidence was associated with higher sending probability. This contrast was stronger for items judged false than for items judged true. Confidence-related differences were also more pronounced in the Control condition than in the Cue condition. Estimates show fixed-effect predictions from models with participant-level random effects. Error bars/shaded intervals indicate 95% confidence intervals.

Table S2: Summary of send MLMs.

	Send			Send			Send			Send			Send		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.68	0.56 – 0.84	**<0.001**	1.51	1.11 – 2.04	**0.008**	0.60	0.48 – 0.74	**<0.001**	0.70	0.60 – 0.83	**<0.001**	0.69	0.52 – 0.92	**0.012**
Condition: Control	*Reference*			*Reference*			*Reference*			*Reference*			*Reference*		
Condition: Cue													2.64	1.81 – 3.85	**<0.001**
Judgment: False	*Reference*			*Reference*			*Reference*			*Reference*			*Reference*		
Judgment: True	1.06	0.90 – 1.24	0.500										1.08	0.92 – 1.27	0.371
Confidence	0.55	0.43 – 0.71	**<0.001**										0.57	0.44 – 0.72	**<0.001**
Confidence:JudgmentTrue	1.31	1.03 – 1.67	**0.030**										1.28	1.01 – 1.63	**0.041**
Confidence:JudgmentFalse:ConditionCue	1.53	1.18 – 1.98	**0.001**										1.49	1.14 – 1.93	**0.003**
Confidence:JudgmentTrue:ConditionCue	1.31	1.08 – 1.60	**0.007**										1.23	1.01 – 1.50	**0.044**
Cue				0.35	0.27 – 0.45	**<0.001**							0.35	0.27 – 0.45	**<0.001**
Socdistcue							1.22	1.08 – 1.38	**0.001**						
Order										0.84	0.78 – 0.90	**<0.001**			
ConditionControl:Order													0.88	0.71 – 1.09	0.239
ConditionCue:Order													0.77	0.62 – 0.96	**0.018**
Random Effects															
σ^2	3.29			3.29			3.29			3.29			3.29		
τ_{00}	0.19_{Subject}			0.28_{Subject}			0.30_{Subject}			0.19_{Subject}			0.20_{Subject}		
τ_{11}	$0.18_{\text{Subject.Confidence}}$												$0.17_{\text{Subject.Confidence}}$		
ρ_{01}	0.71_{Subject}												0.70_{Subject}		
ICC	0.10			0.08			0.08			0.05			0.10		
N	32_{Subject}			32_{Subject}			32_{Subject}			32_{Subject}			32_{Subject}		
Observations	3072			1536			1536			3072			3072		
Marginal R^2 / Conditional R^2	0.028 / 0.127			0.057 / 0.130			0.011 / 0.093			0.009 / 0.063			0.061 / 0.157		

We also estimated an additional mixed-effects logistic model including sociodemographic and questionnaire-derived covariates. Sending probability decreased as Social Value Orientation increased (OR = 0.850, 95% CI [0.749, 0.963], p = .0110), corresponding to a 15.0% reduction in the odds of sending for a one-standard-deviation increase in SVO. Male participants showed higher sending probability than female participants (OR = 1.545, 95% CI [1.185, 2.014], p = .0013). These effects indicate that individual differences contributed to overall sending propensity.

3. Success in estimating others' preferences

Relying on truthfulness estimation had a reliable effect on Senders' success in estimating Receivers' preferences in the Control condition (p = .004) but not in the Cue condition (p = .139). Senders were more successful in inferring Receivers' preferences in the Cue condition than in the Control condition (59.51 ± 20.46% versus 50.28 ± 17.75%). Although we did not detect a significant effect of cue value itself, we did find a significant effect of condition on success (p = .011).

We may have failed to detect a direct effect of cue value because the raw cue does not have a stable directional meaning across stimuli. The cue always indicates proximity to an organization, but the behavioral meaning of that proximity depends on the stimulus class and the stance of the relevant organization in that class. As a result, a higher cue can support either receiving or not receiving depending on the item, which can attenuate a single pooled cue coefficient unless the cue is first recoded into a class-aligned congruence measure. Implementing such a recoding prospectively would have required a substantially different task design, which was not feasible under the constraints of the present experiment. By contrast, a condition effect can still emerge because the Cue condition changes the information structure of the task overall, even if raw cue values are not coded in a way that yields a consistent signed effect across all stimuli.

Together, these results complement the main findings by indicating that Senders' own confidence-based beliefs predicted success in inferring Receivers' preferences only in the Control condition. Receiver-specific social-distance information was incorporated effectively when available, improving preference inference.

Table S3: Summary of send success MLMs.

	Success			**Success_control**			**Success_cue**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.71	0.57 – 0.88	**0.002**	0.64	0.46 – 0.90	**0.011**	0.83	0.55 – 1.25	0.374
Truthfulness: False	*Reference*			*Reference*			*Reference*		
Truthfulness: True	2.50	2.16 – 2.89	**<0.001**	2.76	2.24 – 3.41	**<0.001**	2.42	1.96 – 2.99	**<0.001**
Condition: Control	*Reference*			*Reference*			*Reference*		
Condition: Cue	0.83	0.72 – 0.96	**0.011**						
Confidence	1.06	0.94 – 1.19	0.363	1.29	1.08 – 1.53	**0.004**	0.88	0.74 – 1.04	0.139
Judgment: False	*Reference*			*Reference*			*Reference*		
Judgment: True	0.96	0.83 – 1.11	0.591	0.86	0.69 – 1.07	0.177	1.07	0.87 – 1.32	0.529
Confidence:JudgmentTrue	0.94	0.81 – 1.09	0.434	0.81	0.65 – 1.01	0.062	1.06	0.86 – 1.31	0.592
Cue							0.80	0.63 – 1.02	0.069
Order				0.91	0.74 – 1.12	0.376	0.84	0.67 – 1.06	0.138
Sex: Female	*Reference*			*Reference*			*Reference*		
Sex: Male	1.05	0.89 – 1.22	0.579	1.12	0.90 – 1.41	0.309	0.98	0.77 – 1.24	0.846
Age	1.01	0.94 – 1.09	0.737	1.03	0.92 – 1.14	0.615	1.00	0.90 – 1.12	0.986
Education	1.01	0.96 – 1.05	0.781	1.01	0.95 – 1.07	0.798	1.00	0.94 – 1.07	0.907
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	0.00_{Subject}			0.00_{Subject}			0.09_{Subject}		
τ_{11}				$0.00_{\text{Subject.Order}}$			$0.08_{\text{Subject.Order}}$		
ρ_{01}							-0.96_{Subject}		
ICC							0.01		
N	32_{Subject}			32_{Subject}			32_{Subject}		
Observations	3072			1536			1536		
Marginal R^2 / Conditional R^2	0.063 / NA			0.079 / NA			0.059 / 0.067		

VII. Supplementary Methods VII : Modelling

1. Parameter recovery

For parameter-recovery analyses, we generated synthetic datasets from each candidate model in the model space and fit each dataset with its generating model. Simulations reproduced the structure of the Sending News task while increasing the number of observations available for recovery. We simulated 50 synthetic Senders, each facing 20 synthetic Receivers per condition and 50 stimuli per Receiver, yielding 1,000 trials per condition for each synthetic Sender. Both Control and Cue conditions were simulated. For each synthetic Sender, a ground-truth parameter vector was sampled once and held fixed across 10 repeated datasets, so repeated simulations indexed recovery variability for the same underlying parameters. Receiver behavior was generated first, separately by condition, by assigning cue values on a 1–100 scale and sampling reception choices from a logistic function of cue value (slope = -7, intercept = 2). Candidate receiver tables were accepted only if they yielded a reception rate close to 0.33 (tolerance = 0.005; maximum = 10,000 attempts) and, in the Cue condition, if cue values significantly predicted reception choices ($p < 0.001$). Sender-side inputs were then generated from sender-specific latent tendencies, so that each synthetic Sender showed a stable within-subject pattern of evaluations and binary judgments across trials, while still contributing trials from both judgment categories. The cue environment was reduced to a single stimulus-class setting. Synthetic sending choices were generated trial-by-trial from the corresponding model using the task input matrix. On each trial, models used the current trial inputs and the available history from previous trials, including previous Receiver choice, previous Sender choice, and previous success; after each simulated choice, feedback was computed and the history variables were updated for the next trial. Recovery fits were performed with the same variational Bayesian estimation framework and family-specific prior structure used in the empirical modelling pipeline. Parameter identifiability was quantified by comparing known generating parameters with recovered posterior estimates, focusing on transformed participant-level summaries across repeated simulations. Recoverability was summarized using Pearson correlations, root mean squared error, signed bias, and empirical 95% coverage, defined as the proportion of simulations in which the 95% posterior interval contained the known generating value. The RLtruth and the static Bayes model were not included in the parameter-recovery benchmark to limit computational burden. They are respectively a variant of the RLtruthCue model in which the cue component is omitted, and a restricted variant of the BayesRL model in which the learning component is omitted.

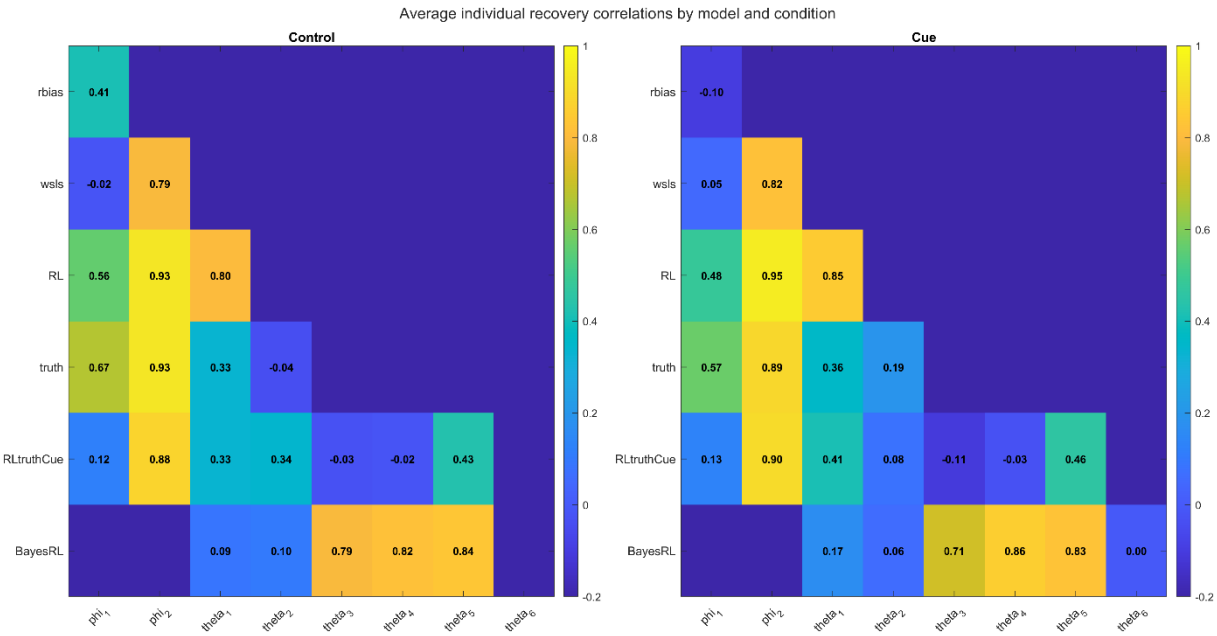


Figure S5: Parameter Recovery Across Models and Conditions. Heatmaps summarize parameter recoverability for each candidate model in the Control condition (left) and Cue condition (right). Rows correspond to models and columns to model parameters. Cell color and overlaid values indicate Pearson correlations between known generating parameters and recovered posterior means, computed on transformed participant-level averages across repeated simulations. Higher values indicate better parameter recovery. Parameters not defined for a given model, not evaluated in a given condition, or fixed by design are left blank.

The uneven recovery pattern suggests that recoverability was parameter-specific and depended on both the generating parameter regime and the structure of the model. One source of difficulty was that some generating parameter values produced behavior that was extreme or only weakly distinguishable under the present task. Recovery was poorer in these regions of parameter space and improved when generating values were restricted to behaviorally plausible ranges, suggesting that part of the initial weak recovery reflected weakly informative simulations. A second source of difficulty was structural overlap between parameters. In the Bayesian model, theta_1 and theta_2 both contribute to the same truth-evidence computation from the same confidence input. Changes in one parameter can therefore be partly compensated by changes in the other, making them difficult to separate from observed choices alone. A related issue arises for the cue parameter theta_6, whose effect on cue-dependent evidence can be partly mimicked by cue-side parameters that shift baseline evidence levels. Thus, different parameter combinations can produce similar cue-dependent behavior, weakening recovery of the cue-weight parameter specifically. Overall, the results indicate that weak recovery was localized to parameters whose effects were either weakly expressed by the generative task structure used in the recovery protocol, or partially confounded with other parameters in the same latent computation. Parameters with more distinct behavioral consequences were recovered more reliably, whereas parameters contributing to overlapping truth- or cue-evidence branches remained only partially identifiable.

2. Pairwise model distinguishability

We assessed model separability using targeted pairwise simulation-based distinguishability analyses under known generating models. We used this approach because the candidate model roster was structurally overlapping, making a single global confusion matrix difficult to interpret as a summary of model separability. Utility models were cumulative extensions of simpler utility models, beta-binomial and utility models used overlapping task inputs, and richer models could absorb simpler models by reducing the contribution of additional components.

Analyses were conducted by comparing pairs of models using free energy. For each model pair, synthetic datasets were generated from one model, both candidate models were fitted to the same synthetic data, and model evidence was compared. We defined the true-model winner rate as the proportion of simulations in which the generating model obtained higher free energy than its paired alternative. Values close to 1 indicate strong recovery of the generating model, values close to .5 indicate weak distinguishability, and values close to 0 indicate systematic absorption by the alternative model.

Because the heuristic models served as control models, the pairwise distinguishability analysis focused on the two theoretically relevant model families: utility models and beta-binomial models. This analysis focused on a constrained sequence of utility and beta-binomial comparisons that tested whether the main computational additions in the model space (reinforcement learning, confidence-based valuation, cue-based valuation, and beta-binomial belief updating) produced separable behavioral signatures.

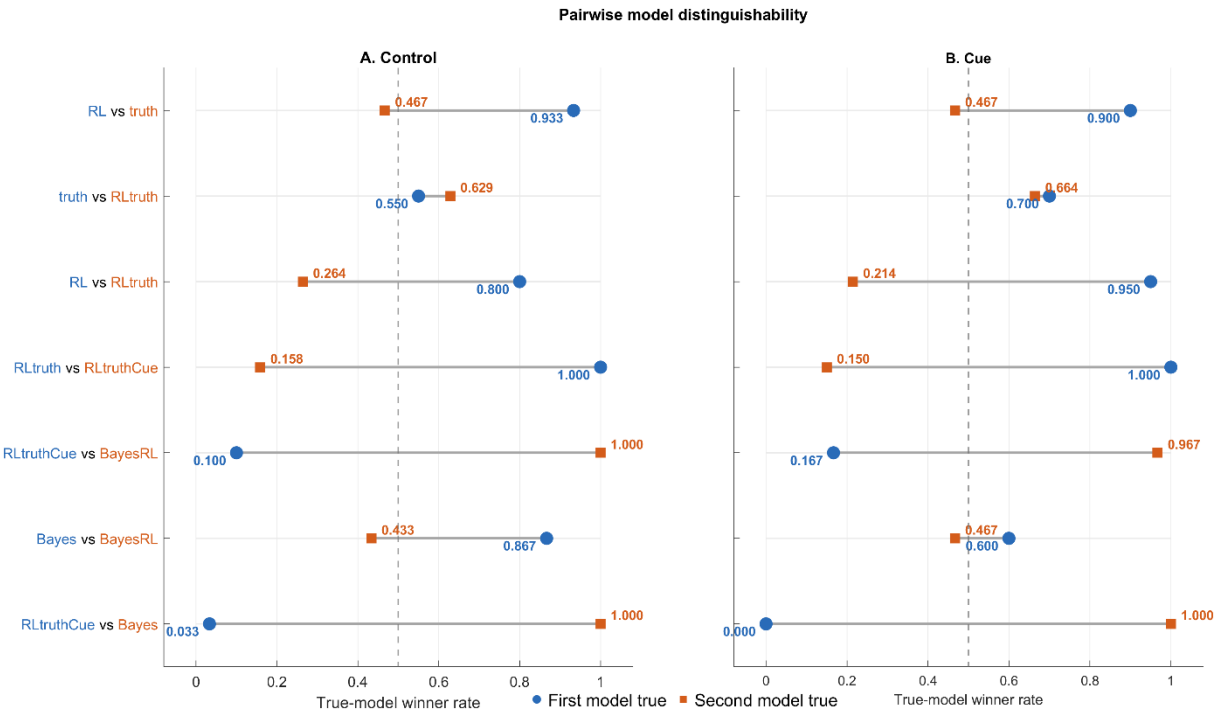


Figure S6: Pairwise Model Distinguishability Across Conditions. Pairwise simulation-based distinguishability results for the candidate models in the Control condition (left) and Cue condition (right). Each row corresponds to one tested model pair. For each pair, the blue circle shows the true-model winner rate when the first-named model generated the synthetic data, and the orange square shows the true-model winner rate when the second-named model generated the synthetic data. Winner rate is defined as the proportion of simulations in which the generating model achieved higher free energy than its paired alternative after both models were fit to the same synthetic dataset.

The pairwise analyses showed that model distinguishability varied across contrasts. Within the utility family, simpler contrasts showed clearer recovery than richer extensions, consistent with the partially nested structure of these models. The richer utility models did not always generate behavior that was separable from neighboring utility models under the present task structure.

Cross-family comparisons favored the beta-binomial models over the richest utility models. When beta-binomial models generated the synthetic data, they were reliably recovered against the richest utility model. Conversely, when the richest utility model generated the synthetic data, beta-binomial models were often selected instead. This indicates that the beta-binomial formulation captured behavioral structure also present in the richest utility model, while providing stronger recovery when beta-binomial data were generated.

Within the beta-binomial family, the social and non-social variants showed partial and asymmetric distinguishability. This pattern is consistent with the fact that both variants share the same beta-binomial observation structure and several task inputs.

Overall, these results indicate that the current model space contains both recoverable and structurally overlapping contrasts. Distinguishability was strongest for the simplest utility contrast, weaker for richer nested utility extensions, and strongly asymmetric in comparisons between the richest utility model and the beta-binomial family.

3. Model calibration

Support for a better fit of the BayesRL model comes from Mean Log Evidence and Bayesian model comparison exceedance probabilities (Manuscript Figure 4a and 4b).

We examined how closely the candidate models matched the observed sending behavior. Model comparison identifies the best-fitting model, but calibration shows whether predicted sending probabilities track the empirical probabilities. We compared observed and predicted

sending probabilities across subjects and models to assess how accurately each model captured participants' choices (Figure S7).

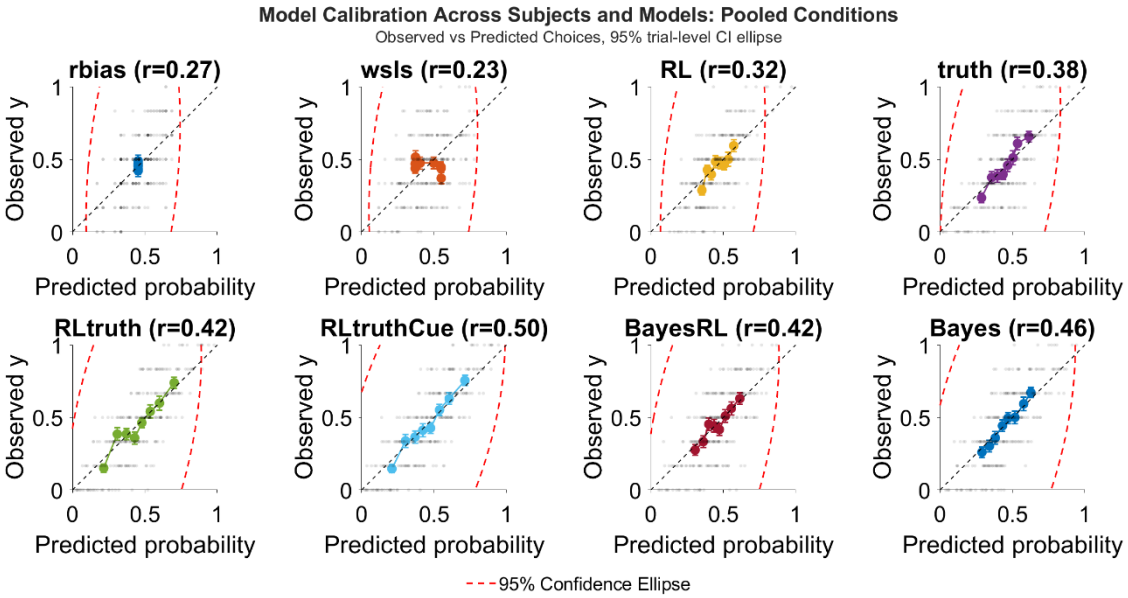


Figure S7: Model Calibration Across Subjects and Models. Observed versus predicted sending probabilities for each candidate model, pooled across conditions. Points show binned predicted and observed probabilities, the diagonal indicates perfect calibration, and the red dashed ellipse shows the 95% trial-level confidence ellipse. Titles report Pearson correlations between predicted and observed responses for each model.

4. BayesRL model

We analyzed the contribution of learning to fitting the participants' data. Support for a contribution of $\theta^{learning}$ to the estimation of participants' behavior comes from the better fit of the BayesRL model than the Bayes model in both conditions (Control: PEP = .863, estimated frequency = .966; Cue: PEP = .992, estimated frequency = .977; Manuscript Figure 4b). The corresponding learning-rate term, which governs how much previously accumulated learning is retained across trials, averaged 0.722 ± 0.111 in the Control condition and 0.670 ± 0.064 in the Cue condition. These results indicate that participants' sending behavior was better captured by a model allowing beliefs to be updated across trials than by an otherwise matched Bayesian account without learning, or by a non-Bayesian model. Note that we fixed all participants' priors at $Beta(0.5,0.5)$, equivalent to p = .5.

We examined what the winning Bayesian model (BayesRL) implies about the latent beliefs underlying participants' choices. In this model, the mean of the beta distribution indexes participants' belief about receivers' preferences, whereas the variance indexes uncertainty around that belief. Figure S8 shows how these latent quantities differ across conditions and across the learning and cue components of the model. The VBA method approximates participants' true priors, revealing as well that participants held prior beliefs about *Receivers*' distance from organizations (average: α = 4.46, β = 6.64). Relatively to the other parameters, the bias, with mean hyper-parameters (α = .48, β = .5) and (α = .45, β = .49) for Control and Cue conditions, respectively, can be interpreted as non-informative.

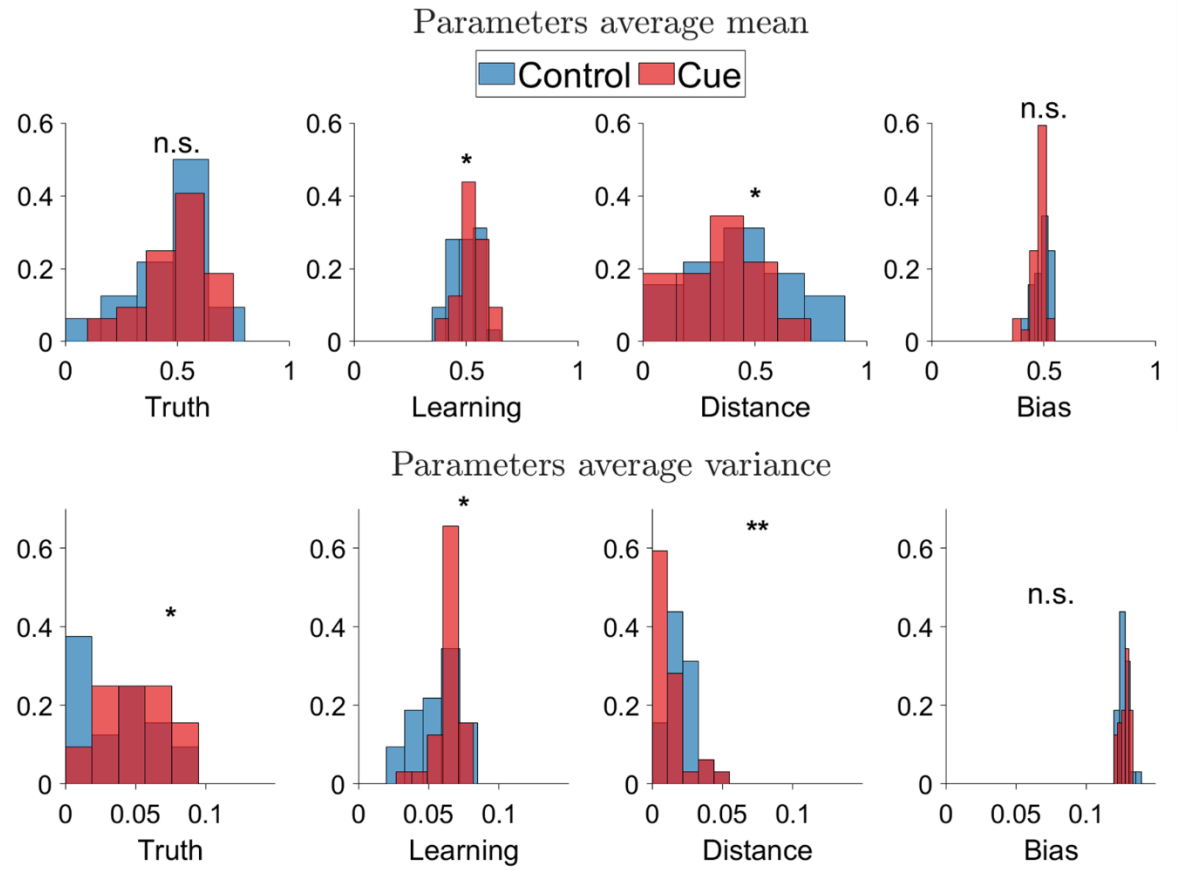


Figure S8: Participants update their beliefs and the confidence about their beliefs in the cue condition. The mean is a property of beta distributions central tendency. In the context of our study, it represents participants' beliefs about the receivers' preferences. The variance represents the dispersion around the mean, representing participants' confidence in their beliefs. In the cue condition, average beliefs about receivers' preferences given the learning increase towards action $a = to\ receive$. However, participants' confidence in their beliefs decrease. In the contrary; average beliefs given the distance increase away from action $a = to\ receive$, with higher confidence, compared to control condition.

To clarify how these latent components map onto choice behavior, we show in Figure S9 the group-level distributions of the model parameters themselves. This allows us to visualize the average relative contribution of truth, learning, cue, and bias terms to estimated send probability, and how these contributions differ between the Control and Cue conditions.

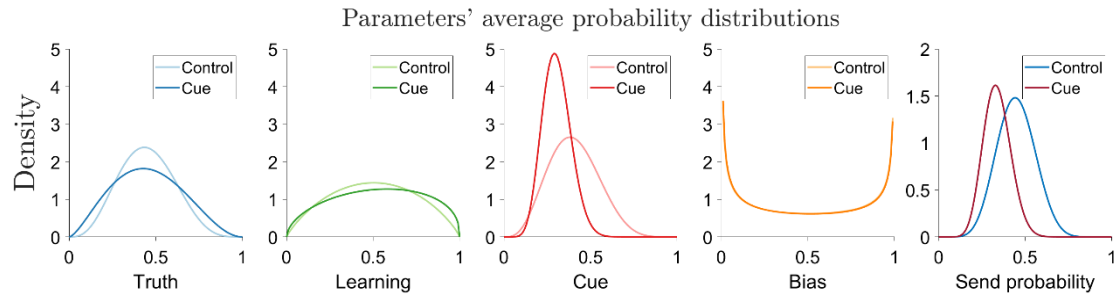


Figure S9: Parameters' Average Probability Distributions. Group-averaged beta distributions for the Bayesian model parameters in the Control and Cue conditions. Panels show the distributions associated with truth, learning, cue, and bias parameters, together with the implied distribution over send probability. Rightward shifts indicate higher estimated probability of sending, whereas leftward shifts indicate lower estimated probability of sending.

To quantify these effects formally, we tested whether between-subject variation in the modal values of the Bayesian parameters predicted send probability. Table S4 reports the mixed-effects model linking send probability to the modes of the fitted Bayesian parameters, considering the effects of response times during the sending decision, news theme and participants news truthfulness judgment.

Table S4: MLM of send probability explained by modes of the Bayesian model parameters.

	Send_control			**Send_cue**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.94	0.75 – 1.18	0.606	0.41	0.31 – 0.56	**<0.001**
Mode_learning	1.05	0.93 – 1.19	0.397	1.00	0.89 – 1.13	0.955
Mode_truth	1.51	1.29 – 1.76	**<0.001**	1.05	0.86 – 1.29	0.621
Mode_distance	2.33	1.48 – 3.65	**<0.001**	2.88	2.32 – 3.59	**<0.001**
Mode_bias	1.42	0.91 – 2.23	0.123	1.21	0.95 – 1.54	0.116
Send_RT	1.15	1.03 – 1.29	**0.016**	1.16	1.03 – 1.31	**0.014**
Theme: Democracy	*Reference*			*Reference*		
Theme: Ecology	0.90	0.70 – 1.17	0.438	1.00	0.72 – 1.40	1.000
Theme: SocJust	1.02	0.79 – 1.33	0.871	1.39	1.02 – 1.91	**0.039**
Judgment: False	*Reference*			*Reference*		
Judgment: True	0.86	0.69 – 1.08	0.191	1.41	1.10 – 1.79	**0.006**
Random Effects						
σ^2	3.29			3.29		
τ_{00}	0.00 $_{Subject}$			0.00 $_{Subject}$		
τ_{11}	0.00 $_{Subject.Mode\_bias}$			0.05 $_{Subject.Mode\_bias}$		
	0.09 $_{Subject.Mode\_truth}$			0.20 $_{Subject.Mode\_truth}$		
	0.00 $_{Subject.Mode\_distance}$			0.03 $_{Subject.Mode\_distance}$		
	0.02 $_{Subject.Mode\_learning}$			0.00 $_{Subject.Mode\_learning}$		
ρ_{01}	1.00			-1.00		
	-1.00			1.00		
	1.00			-1.00		
	1.00			0.93		
ICC	0.03			0.07		
N	32 $_{Subject}$			32 $_{Subject}$		
Observations	1536			1536		
Marginal R^2 / Conditional R^2	0.118 / 0.147			0.219 / 0.271		

We finally asked whether model fit improved when the mode of θ^{action}, representing participants' integrated beliefs about the Receiver's preferences, was added explicitly to the mixed-effects model. Table S5 reports the corresponding analysis. The Marginal R^2 increased from 0.118 to 0.191 in the Control condition, and from 0.219 to 0.316 in the Cue condition, indicating that θ^{action} accounts for substantially more variance in send probability than the parameterization reported in Table S4. This suggests that the individual parameter modes capture only part of the decision process, whereas θ^{action} distribution reflects their integrated contribution more directly.

Table S5: MLM of send probability explained by modes of the Bayesian model parameters after introducing the mode of θ beta distribution.

	Send_control			**Send_cue**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.91	0.73 – 1.14	0.428	0.40	0.30 – 0.53	**<0.001**
Mode_learning	0.88	0.79 – 0.99	**0.039**	0.81	0.72 – 0.92	**0.002**
Mode_truth	0.89	0.76 – 1.04	0.148	0.78	0.68 – 0.89	**<0.001**
Mode_distance	0.76	0.46 – 1.24	0.269	0.67	0.48 – 0.95	**0.024**
Mode_bias	0.87	0.56 – 1.35	0.535	0.97	0.81 – 1.17	0.752
Mode_action	2.76	2.19 – 3.46	**<0.001**	4.84	3.48 – 6.74	**<0.001**
Send_RT	1.13	1.01 – 1.26	**0.039**	1.08	0.96 – 1.21	0.226
Theme: Democracy	*Reference*			*Reference*		
Theme: Ecology	0.90	0.69 – 1.17	0.412	0.95	0.68 – 1.33	0.786
Theme: SocJust	1.04	0.80 – 1.36	0.763	1.32	0.97 – 1.81	0.080
Judgment: False	*Reference*			*Reference*		
Judgment: True	0.87	0.69 – 1.09	0.224	1.42	1.11 – 1.81	**0.005**
Random Effects						
σ^2	3.29			3.29		
τ_{00}	0.00 $_{Subject}$			0.00 $_{Subject}$		
τ_{11}	0.00 $_{Subject.Mode\_bias}$			0.00 $_{Subject.Mode\_bias}$		
	0.00 $_{Subject.Mode\_truth}$			0.00 $_{Subject.Mode\_truth}$		
	0.00 $_{Subject.Mode\_distance}$			0.00 $_{Subject.Mode\_distance}$		
	0.00 $_{Subject.Mode\_learning}$			0.00 $_{Subject.Mode\_learning}$		
	0.00 $_{Subject.Mode\_action}$			0.00 $_{Subject.Mode\_action}$		
ρ_{01}	-0.30			-0.49		
	-0.82			0.37		
	-0.80			0.23		
	0.64			-0.32		
	0.79			-0.30		
N	32 $_{Subject}$			32 $_{Subject}$		
Observations	1536			1536		
Marginal R^2 / Conditional R^2	0.191 / NA			0.316 / NA		

VIII. Supplementary Methods VIII : fMRI analyses

1. Behavior-based fMRI analyses

We analyzed brain activity associated with confidence during news evaluation. When participants exhibited higher confidence in their own assessment of news veracity, increased activity was observed in ventral striatum (all *p* voxel-level < .001 uncorrected, cluster-level < .05; Figure S10).

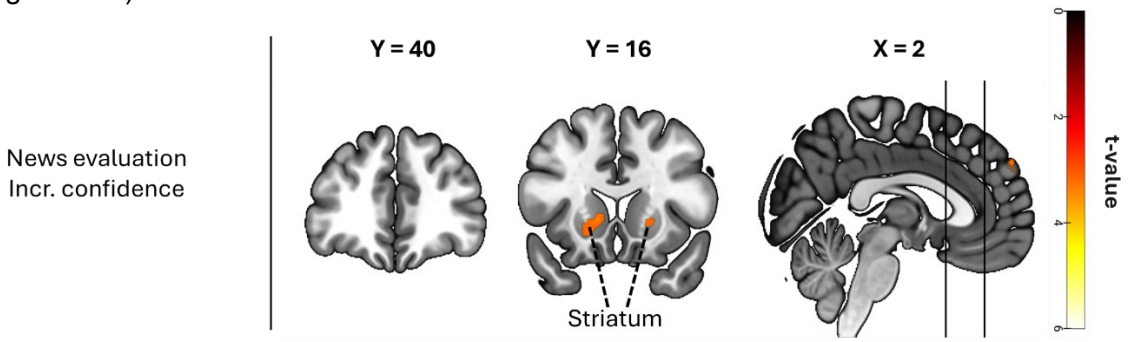


Figure S10: The confidence in veracity estimation elicited brain activity in ventral striatum. When participants evaluated news item veracity, an increase in the confidence in their evaluation modulated the BOLD signal in ventral striatum. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to p<.001 and (cluster-level thresholds at p<.05 uncorrected, filter k>30).

We also compared brain activity during sending decisions (regardless of the decision to send or withhold additional information) when participants succeeded in matching the receiver's preference, with sending decisions when participants failed. In a single whole-brain analysis, the contrast elicited higher activity in the ventromedial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC), and in the left amygdala (all *p* voxel-level < .005 uncorrected, cluster-level < .05; Figure S11).

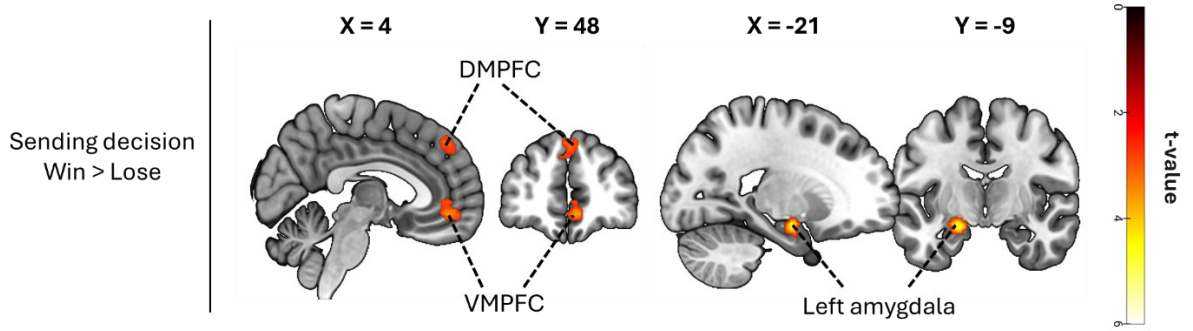


Figure S11: A successful decision to send or withhold additional information was associated with prefrontal and amygdala activity. Trials in which participants succeeded in matching the receiver's preference for receiving or not receiving additional information were associated with increased activity in dmPFC, vmPFC and left amygdala – compared to when they failed to match the receiver's preference. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to p<.005 (cluster-level thresholds at p<.05 uncorrected, filter k>30).

Then, we analyzed brain activations associated with decisions to withhold additional information versus to send extra-information, modulated by participants' behavioral measures—confidence in the *Control* condition and social distance in the *Cue* condition. In a first whole-brain analysis, during decisions to withhold (relative to send) when participants exhibited higher confidence in their own assessment of news veracity, increased activity was observed in the midbrain, ventral striatum, ventromedial prefrontal cortex (vmPFC), and

dorsolateral prefrontal cortex (dlPFC) (all *p* voxel-level < .001 uncorrected, cluster-level < .05; Figure S12a). In a complementary whole-brain analysis comparing decisions to withhold versus to send extra-information, greater activity was observed in the superior temporal sulcus (STS) and dorsomedial prefrontal cortex (dmPFC) when participants viewed cues signaling larger social distance between the receiver's opinions and the news content (all *p* voxel-level < .001 uncorrected, cluster-level < .05; Figure S12b).

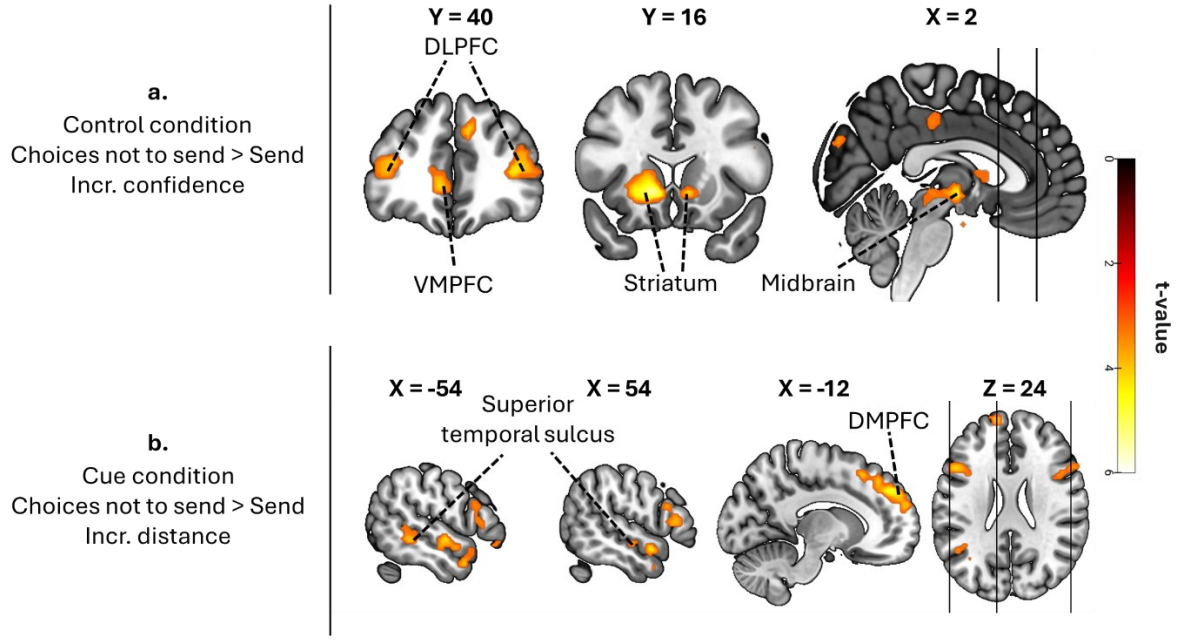


Figure S12: The confidence in veracity estimation and the Receivers' social distance to the relevant organisation elicited corresponding activity in an information valuation network and a social processing network. Throughout the task participants learned to decrease their probability to send extra information. **a)** When participants had no cue about *receivers*, an increase in confidence in the news item veracity judgement was associated with a decrease in the probability. This increase in confidence modulated the BOLD signal in midbrain, striatum, vmPFC and dlPFC, with greater effect during choices not to send. **b)** When they received the social distance cue, it anti-correlated with the probability to send additional information. The *Receiver*'s social distance to the relevant organization modulated the BOLD signal in STS and dmPFC, with greater effect during choices not to send. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to p<.001 (cluster-level thresholds at p<.05 uncorrected, filter k>30).

Table S6: Summary of significant activations for fMRI GLM with behavioural variables. Voxel-level threshold: p<.001, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. DLPFC: dorsolateral prefrontal cortex. VMPFC: ventromedial prefrontal cortex. DMPFC: dorsomedial prefrontal cortex. STS: superior temporal sulcus.

contrast: Send > Not send	label	voxels at p<.001	peak z-score	p (cluster FWE corrected)	p (cluster uncorrected)	peak voxel MNI coordinates			
Increasing confidence, control condition	midbrain	104	4.62	.035	.006	6	27	-6	
	striatum	765	5.25	.000	.000	-12	15	-3	LR
	DLPFC	68	4.02	.118	.022	-42	42	15	L
	DLPFC	149	4.5	.009	.002	39	36	12	R
	VMPFC	55	4.04	.188	.036	-9	39	0	
Increasing distance, cue condition	DMPFC	368	4.61	.000	.000	-15	48	36	
	STS	569	4.31	.000	.000	-51	-9	-6	L
			4.5	.000	.000	-42	33	-12	L
	STS	93	4.15	.032	.005	-51	-42	-6	L
	STS	63	4.19	.105	.016	-39	-57	18	L
	STS	55	4.23	.147	.023	54	3	-15	R
	STS	57	4.16	.135	.021	63	-15	-12	R

2. Bayesian-based fMRI analyses

We extended analyses shown in figure 7 by analyzing brain activations associated with decisions to withhold additional information versus to send extra-information, modulated by participants' Bayesian parameters—θ^{truth} in the *Control* condition and $\theta^{distance}$ in the *Cue* condition. In a first whole-brain analysis, during decisions to withhold (relative to send) when θ^{truth} was decreasing (associated with a decreasing probability to send), increased activity was found in the ventral striatum, ventromedial prefrontal cortex (vmPFC), and dorsolateral prefrontal cortex (dlPFC) (all *p* voxel-level < .001 uncorrected, cluster-level < .05; Figure S13a). In the second whole-brain analysis, when $\theta^{distance}$ was decreasing (associated with a decreasing probability to send), greater activity was observed in the right temporoparietal junction (rTPJ) and right superior temporal sulcus (rSTS) only (all *p* voxel-level < .001 uncorrected, cluster-level < .05; Figure S13b).

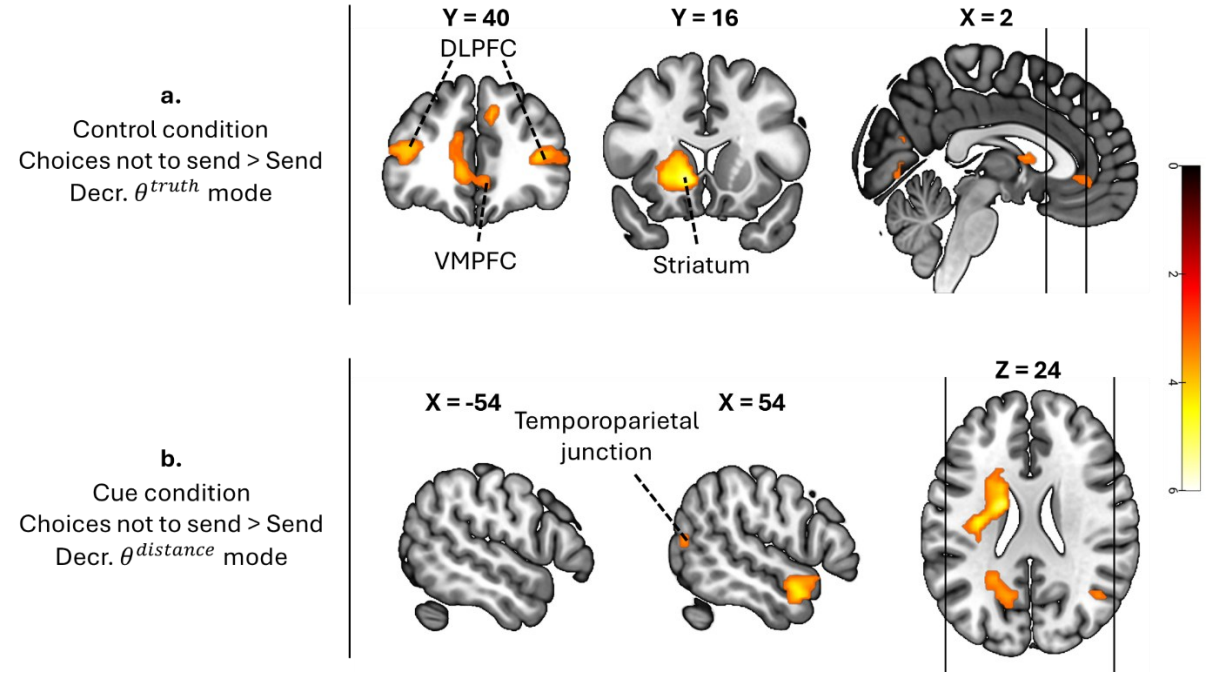


Figure S13: Confidence-based and Social-distance-based beliefs about *Receivers*' preferences during choices not to send compared to choices to send. Throughout the task participants learned to decrease their probability to send extra information. **a)** The increase of brain activity in dmPFC, vmPFC and ventral striatum associated with decreasing confidence-based beliefs about Receiver's preferences for receiving additional information (represented by θ^{truth}) was stronger when participants chose to withhold sending information compared to when they chose to send. **b)** When social-distance-based beliefs about

Receiver's preferences for receiving additional information (represented by $\theta^{distance}$) decreased, brain activity during choices to withhold, relative to send, was stronger in the rTPJ and rSTS. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to p<.001 (cluster-level thresholds at p<.05 uncorrected, filter k>30).

Table S7: Summary of significant activations for fMRI GLM with Bayesian parameters. Voxel-level threshold: p<.001, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. VLPFC: ventrolateral prefrontal cortex. VMPFC: ventromedial prefrontal cortex. DMPFC: dorsomedial prefrontal cortex. TPJ: temporo-parietal junction.

contrast: Decisions not to send	label	voxels at p<.001	peak z-score	p (cluster FWE corrected)	p (cluster uncorrected)	peak voxel MNI coordinates			
Decreasing θ(*truth), control condition*	striatum	73	3.97	.101	.016	-9	15	-6	L
	VMPFC	52	3.49	.218	.037	-3	30	-9	
	DLPFC	53	3.91	.210	.035	39	36	12	R
Decreasing θ(distance*), cue condition*	TPJ	324	3.8	.000	.000	-39	-69	18	L
	TPJ	228	4.76	.001	.000	45	-63	15	R
contrast: Time of sending decision									
Decreasing θ, both conditions	VLPFC	44	3.63	.161	.017	24	51	0	R