Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding

Kang Zhang
zksoda@hotmail.com
Kyushu University
Fukuoka, Fukuoka, Japan

Hiroaki Shinden
shinden.hiroaki.652@gmail.com
Kyushu University
Fukuoka, Fukuoka, Japan

Tatsuki Mutsuro
mutsuro.tatsuki@kyudai.jp
Kyushu University
Fukuoka, Fukuoka, Japan

Einoshin Suzuki
suzuki@inf.kyushu-u.ac.jp
Kyushu University
Fukuoka, Fukuoka, Japan

ABSTRACT

This paper proposes 18 types of statistical data explanations and three kinds of procedures to investigate credibility in unethical and biased explanations due to exploitation of the 10 instincts proposed by Rosling et al. The explanation "women have lower math scores than men" accompanied with the averages and the distributions of their scores is an example of such an explanation, as it exploits the gap instinct, i.e., our tendency to divide all kinds of things into two distinct and often conflicting groups. It becomes much less credible if we replace the word "math" with "English", even if we keep the data as they are, as the exploitation seems to fail. Our judging procedures are based on phrase embedding and carefully designed comparisons to judge the credibility. The results of our experiments comparing the 18 types with their variants show promising results and clues for further developments.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Applied computing** → *Psychology*.

KEYWORDS

text classification, word embedding, exploitation of thinking traits, AI and ethics

ACM Reference Format:
Kang Zhang, Hiroaki Shinden, Tatsuki Mutsuro, and Einoshin Suzuki. 2022. Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3514094.3534171

1 INTRODUCTION

As the impact and the presence of AI systems on our societies increase, their unethical misconducts are prone to severe reproach. The hijacking event of the chatbot Tay clearly shows that pure benevolence could turn into an opposite outcome [39]. Inflammatory tweets by a chatbot are often unethical, harm the reputation of its producer, and challenge the moral of our society. For the last issue, several tweets are more influential than others, as their content are likely to be believed due to several reasons.

In this article, among such reasons, we tackle exploitation of human instincts in statistical data explanation. Rosling et al.'s book "Factfulness" has known a global success and emphasizes the importance of thinking based on facts and correct understandings [30]. The book includes examples of unethical and biased explanations each of which is denied by the accompanied statistical data. We, however, argue that such a thinking attitude is not always adopted and even accepted. Take as an example an explanation "women have lower math scores than men"[1] with chronological scores at SAT tests in the US and the score distributions of men and women in the 2016 test [30] in V in Figure 1. The latter statistical data clearly show the absurdness of discussing men and women in mass, as individually women or men who are good at math exist as those who are not. However, due to the gap instinct [30], i.e., our tendency to divide all kinds of things into two distinct and often conflicting groups, some portion of the public would believe the explanation, even though the statistical data clearly refute it. Such a situation deserves special attention as it highlights challenges to our rationality. In this paper, we are going to define 18 types of such credible and unethical explanations each with its statistical data.

We also provide countermeasures to such explanations. Word embedding [14, 17, 24–27], which projects a word in a high dimensional space, keeping its semantics relative to other words, has known notable successes [2, 13, 35]. We employ its extension, sentence embedding [9, 28, 36] to embed phrases, and devise three methods that judge whether an explanation is credible and unethical.

The organization of this paper is as follows. Section 2 reviews relevant works. We define the target problem in Section 3 and propose our methods in Section 4. Section 5 evaluates the methods by experiments and Section 6 concludes.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534171

[1] All unethical examples in this paper are either adopted from other sources or slightly modified from them and do not reflect the beliefs of the authors nor our organizations. In all cases, such examples are not believed by the authors of the sources, either.

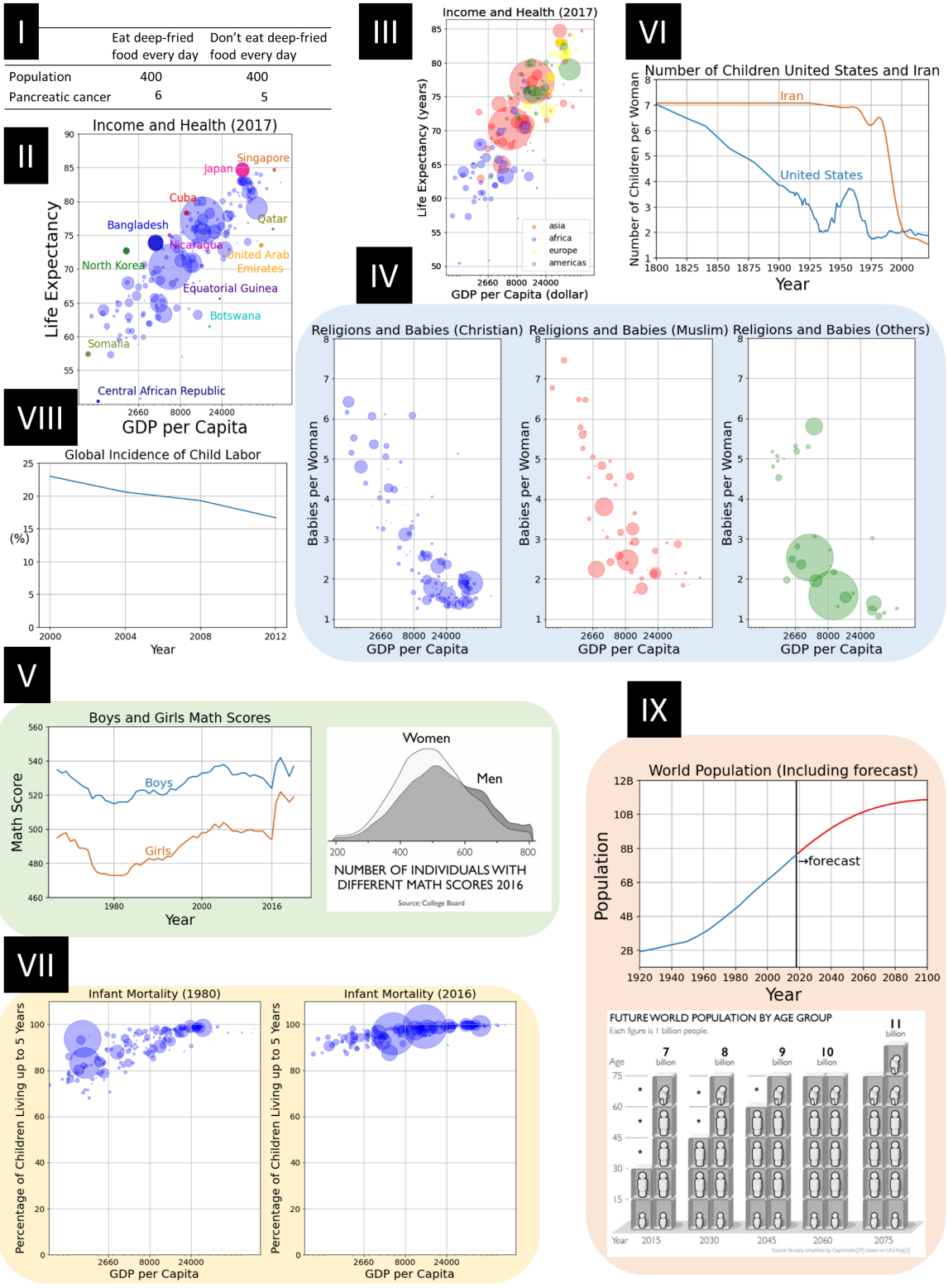

Figure 1: (best in color) Statistical data in explanations (I) - (IX). Data are adopted or modified from [30], [4], or Gapminder [31].

2 RELATED WORK

Unethical and biased explanations are widely generated in diverse fields around the world, such as fake news and hoaxes[1]. Detecting fake news is a challenging Natural Language Processing (NLP) task involving two problems: characterization and detection [32]. Considering feature selection and extraction, Reis et al. [29] designed informative features, which consider semantic and syntactic properties, political biases, credibility, and environments of news, for automatic detection of fake news. Vlachos et al. [34] introduced fact checking tasks and discussed baseline approaches to assess truthfulness of explanations by measuring their semantic similarities. Detecting fake news is usually formulated as a classification task in a supervised manner [23, 38]. Through integrating meta data with texts, a hybrid Convolutional Neural Network (CNN) is devised to classify fake news based on surface-level linguistic patterns [37]. In this paper, we limit our attention to explanations of statistical data and focus on their unethical nature and credibility due to instinct exploitation.

As we explained above, our judging procedure is based on phrase embedding and carefully designed comparisons to judge the credibility. Measuring semantic similarity between various text components such as words, sentences, or documents has been explored in a wide range of downstream NLP tasks, such as machine translation [40], information retrieval [19] and question answering [22]. Li et al. [21] measured semantic similarity between words using multiple information sources, including attributes path lengths, depths, and local densities in a hierarchical semantic knowledge base. To reduce the ambiguity in words, a robust semantic similarity measure [5] was devised by utilizing information including page counts and lexico-syntactic patterns from text snippets of a Web search engine. Similar to [5], Normalized Google Distance (NGD) [10] was proposed to measure the similarity between two terms based on query results of Google search engine.

Semantic similarity methods have exploited the recent developments in neural networks and word embedding to enhance their performance [8]. In contrast to adopting traditional static word embedding [20, 25] for semantic similarity measurement between words [6], contextualized word embedding generated from modern neural language models, such as ELMo [26], GPT-2 [27], and BERT [14], has been widely employed for semantic similarity tasks [16]. The latter approach possesses over the former an advantage of capturing rich syntactic and semantic properties of words under diverse linguistic contexts. Moreover, for semantic similarity tasks between two sequences of multiple words, such as phrases and sentences, InferSent [11] employs a bi-directional Long-Short Term Memory (LSTM) with a max-pooling operator [18] as a sentence encoder to generate sentence embedding. Trained on a number of natural language prediction tasks, Universal Sentence Encoder [7] models the meaning of word sequences to encode sentences into high dimensional vectors. Sentence-BERT [28] adopted Siamese and triplet architectures based on the pre-trained BERT network to generate semantically meaningful embedding for sentences. Furthermore, the semantic similarities of sentences can be directly compared with cosine-similarity between their embeddings.

To the best of our knowledge, no previous work tackles the problem of judging credible and unethical explanations on statistical data with AI methods. This paper is the first work to define and investigate such explanations through AI techniques.

3 TARGET PROBLEM

3.1 Rosling et al.'s Ten Instincts

We focus our attention on Rosling et al.'s ten instincts [30], which are listed below. The ten instincts could be considered as innate, typically fixed patterns of human thinking.

(1) The gap instinct: our tendency to divide all kinds of things into two distinct and often conflicting groups, with an imagined, huge gap in between.
(2) The negativity instinct: our tendency to notice the bad more than the good.
(3) The straight line instinct: our tendency to believe that the increase is a straight line.
(4) The fear instinct: our tendency to focus our attention to what we are afraid of.
(5) The size instinct: our tendency to misjudge the size of things or the importance of a single number/instance.
(6) The generalization instinct: our tendency to categorize and generalize things all the time.
(7) The destiny instinct: our tendency to consider that several things never change due to their innate characteristics.
(8) The single perspective instinct: our tendency to prefer a single cause or solution.
(9) The blame instinct: our tendency to find a clear, simple reason for why something bad has happened.
(10) The urgency instinct: our tendency to want to take an immediate action in the face of a perceived imminent danger.

3.2 Credible and Unethical Explanation of Statistical Data that Exploits the Instincts

We assume the following five conditions for our credible and unethical explanation of statistical data.

(1) Data seem to be valid, ideally taken from an authoritative source, e.g., WHO.
(2) The explanation is significant.
(3) The explanation seems to be believed by a certain number of people.
(4) The data can prove why the explanation is not valid.
(5) The explanation exploits at least one of the ten instincts in Section 3.1.

Note that the first three conditions are mandatory for the validness, the significance, and the credibility of the explanation in this order. If we omit one of them, the explanation will be simply neglected. As we are going to consider variants for each explanation by replacing its phrases while keeping the data, we have inserted the phrase "seem to" in the first condition. Likewise, the third condition needs this phrase as it depends of subjectivity and knowledge of various people. The fourth condition assures that we can refute the explanation based on the accompanied data only. Without the condition, the unethical nature of the explanation depends on the beliefs and the knowledge of the judge, who are diverse. The 10

instincts in Section 3.1 are prone to unethical notions such as segregation, prejudice, fear, and inequality. Thus the last condition contributes to the unethical nature of the explanation.

Our target problem is to judge whether a given explanation is credible and unethical (class 1) or not (class 0). The types of the explanation will be defined in the next section. The math score example and its variant on the English score in Section 1 clearly show that these five conditions are not enough to solve this judgment problem. Our three kinds of solutions will be explained in Section 4.2.

4 PROPOSED METHODS

4.1 Eighteen Types

The 18 types of explanations (I)-(XVIII) describe 7 kinds of statistical data. The data are (A) values of a probabilistic variable under 2 conditions, (B) a scatter plot of 2 probabilistic variables, (C) scatter plots in different categories, (D) a probability density function of a probabilistic variable and a plot of its average value, (E) a time-series chart or scatter plots in chronological order, possibly with an additional one, (F) scatter plots of 2 probabilistic variables focusing on the total values and the average values, and (G) a funnel plot. Examples of the statistical data are shown in Figures 1 and 2.

For each type of the explanations, we code the exploited instincts and its statistical data. For example, in explanation (I), A-2 represents that the explanation exploits instinct (2) to explain the statistical data (A). In addition, we clarify why these explanations are not valid according to statistical data. *PhraseX* and *PhraseY* respectively represent an object and its characteristics. Lastly, we provide candidates for replacing *PhraseX* and *PhraseY* to generate variants.

(I) A-2, A-4: Deep-fried food boosts pancreatic cancer risk.
PhraseX: deep-fried food. *PhraseY*: pancreatic cancer.

(Clarification) The relative risk of pancreatic cancer is only increased by 0.25% [4]. Statistically testing the difference between the two groups will fail.

(Candidates for variants) *PhraseX*: alcohol abuse, heavy drinking, long-distance running. *PhraseY*: Alzheimer's disease, periodontal disease, flu, alopecia areata, bone fracture, nosebleeds.

We have two variations for explanations (II), which are used for upper-left and lower-right countries in Fig. 1 II.
(II-1) B-8: Cuba is the poorest of the healthiest countries.
PhraseX: Cuba. *PhraseY*: poorest.
(II-2) B-8: United Arab Emirates (UAE) is the richest of the unhealthiest countries.
PhraseX: United Arab Emirates. *PhraseY*: richest.

(Clarification) (II-1) Cuba is also the healthiest of the poorest countries. It is inappropriate to consider only one side.
(II-2) The same reason applies to UAE, which is also the unhealthiest of the richest countries.

(Candidates for variants) (II-1) *PhraseX*: Bangladesh, North Korea, Nicaragua. *PhraseY*: richest.
(II-2) *PhraseX*: Qatar, Equatorial Guinea, Botswana. *PhraseY*: poorest.
Note that Cuba and UAE are respectively compared among the healthiest and unhealthiest countries in these explanations.

(III) B-3: Life expectancy continues to grow in proportion to GDP per capita.
PhraseX: life expectancy. *PhraseY*: proportional to GDP.

(Clarification) Note that the horizontal axis in the figure is in a logarithmic scale, which is non-linear. Average life has an upper bound.

(Candidates for variants) *PhraseX*: healthy life expectancy. *PhraseY*: inversely proportional to GDP, not correlated to GDP.

(IV) C-1, C-6, C-7, C-8: Muslims have many babies compared to Christians.
PhraseX: Muslims. *PhraseY*: many babies.

(Clarification) All the 3 plots show that the number of babies decreases as the income increases, and there is no significant difference in the distribution. In fact, the average number of children per woman is 2.7 among Christians and 3.1 among Muslims.

(Candidates for variants) *PhraseX*: Judaisms, Christians. *PhraseY*: few babies.
Note that the target of the comparison should be changed in the variants, which will be shown in the Appendix.

(V) D-1, D-2, D-6, D-8: Women have lower math scores than men.
PhraseX: women. *PhraseY*: low math score.

(Clarification) The left plot shows that girls (women) have lower average scores than boys (men). However, the right plot shows that there exists an almost complete overlap between the two groups.

(Candidates for variants) *PhraseX*: men.
PhraseY: high math score, low English score, high English score.

(VI) E-7, E-8: Iranians have many children compared to Americans in the 21st century.
PhraseX: Iranians. *PhraseY*: many children.

(Clarification) In the past centuries, Iranians had more children than Americans. In this century the two groups are similar in the number of children.

(Candidates for variants) *PhraseX*: Afghans, Americans, French.
PhraseY: few children.

(VII) E-1, E-6, E-7, E-8: Infant mortality rates in developing countries are still significantly higher than in advanced countries.
PhraseX: developing countries. *PhraseY*: high infant mortality rates.

(Clarification) The percentage of children living up to 5 years is now over 85% in most countries, and there is no significant difference between advanced and developing countries.

(Candidates for variants) *PhraseX*: advanced countries.
PhraseY: low infant mortality rates, low enrollment rates, high enrollment rates.

(VIII) E-2, E-5: Child labor is about 15% and is not decreasing.
PhraseX: child labor. *PhraseY*: not decreasing.

(Clarification) The percentage of child labor is decreasing.

(Candidates for variants) *PhraseX*: child hunger, child mortality.
PhraseY: increasing, decreasing, not increasing, constant.

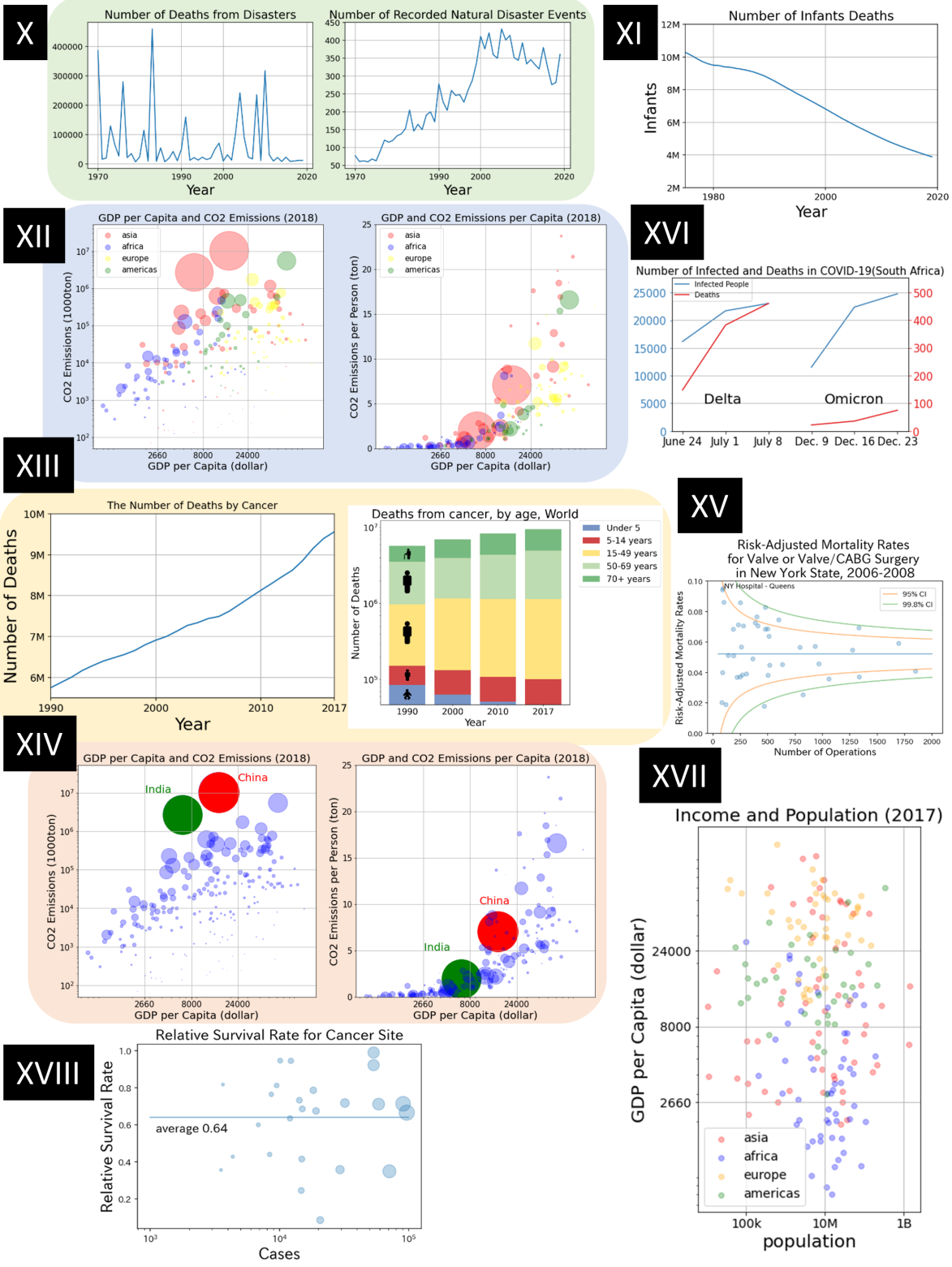

Figure 2: (best in color) Statistical data in explanations (X) - (XVIII). Data are adopted or modified from [30], [4], or Gapminder [31].

(IX) E-3: The world population will just increase.
PhraseX: world population. *PhraseY*: will just increase.

(Clarification) The bottom plot shows that the populations of younger generations are stable and those of older ones slowly increase. As the results, the population growth will be controlled.

(Candidates for variants) *PhraseY*: will rapidly increase, will just decrease, will rapidly decrease, will keep constant.

(X) E-2, E-4: Since year 2000, compared to 1980, there is an increasing in natural disasters and an increasing in deaths from natural disasters.
PhraseX: increasing in natural disasters. *PhraseY*: increasing in deaths from natural disasters.

(Clarification) The number of natural disasters is increasing, whereas the number of deaths from disasters is fluctuating and tends to decrease.

(Candidates for variants) From this type, we use {} as there are many candidates. A variant should discuss one of the three topics. *PhraseX*: {increasing in, decreasing in, constant} {natural disasters, epidemic damages, industrial accidents}. *PhraseY*: {increasing in, decreasing in, constant} deaths from {natural disasters, epidemic damages, industrial accidents}.

(XI) E-2, E-5, E-10: The death of many babies (4 million) is increasing.
PhraseX: death of many babies. *PhraseY*: increasing.

(Clarification) Nearly 10 million babies died 40 years ago, but recently the number has fallen to 4 million and the situation is improving.

(Candidates for variants) *PhraseX*: death of many {children, adults, old people}. *PhraseY*: not decreasing, decreasing, not increasing, constant.

(XII) F-1, F-6, F-7, F-8, F-9: Asia is the cause of the large amount of CO2 emissions.
PhraseX: Asia. *PhraseY*: large amount of CO2 emissions.

(Clarification) Asian countries seem to be the cause in the view of total emissions, which is denied by the per person emission view with respect to the GDP per capita.

(Candidates for variants) *PhraseX*: Africa, Europe[2].
PhraseY: small amount of CO2 emissions[3].

(XIII) E-2, E-8: The risk of death from cancer is increasing worldwide.
PhraseX: risk of death from cancer. *PhraseY*: increasing.

(Clarification) The number of deaths from cancer is increasing, which is the result of the increase of the elderly in number.

(Candidates for variants) *PhraseX*: risk of death from {Alzheimer's, heart} disease. *PhraseY*: decreasing, constant.

(XIV) F-8, F-9, F-10: China is the cause of the large amount of CO2 emissions. *PhraseX*: China. *PhraseY*: large amount of CO2 emissions.

(Clarification) A large population inevitably leads to an increase in CO2 emissions. In terms of CO2 emissions per person, the explanation is denied.

(Candidates for variants) *PhraseX*: United Kingdom, India, United States. *PhraseY*: small amount of CO2 emissions.
Note that this type gives a more precise view than type (XII), which explains continents.

(XV) G-1, G-8, G-9: Small hospitals are dangerous hospitals[4]. *PhraseX*: small hospitals. *PhraseY*: dangerous hospitals.

(Clarification) The funnel plot shows that most of the data points are within the confidence interval [4]. Thus there is no such tendency.

(Candidates for variants) *PhraseX*: large hospitals. *PhraseY*: safe hospitals.

(XVI) E-1, E-6, E-7, E-8: Omicron strain of COVID-19 is less dangerous than Delta strain.
PhraseX: Omicron strain. *PhraseY*: less dangerous.

(Clarification) Judging the dangerous degree of Omicron strain only by the number of deaths is inadequate. Omicron strain is more dangerous than Delta strain in the view of infections.

(Candidates for variants) *PhraseX*: Alpha strain, Beta strain, Delta strain, Gamma strain. *PhraseY*: more dangerous.
Note that COVID-19 hasn't been clarified scientifically and is a subject of fierce debate in 2022.

(XVII) B-1, B-6, B-7: Africa has lower GDP per capita than other regions.
PhraseX: Africa. *PhraseY*: low GDP.

(Clarification) Not all African countries have lower GDP per capita than other regions.

(Candidates for variants) *PhraseX*: Asia, Americas, Europe. *PhraseY*: high GDP.
Note that we compare *PhraseX* with all the rest, which will be discussed in the Appendix.

(XVIII) B-2, B-6, B-7, B-8: The average 5-year survival rate for cancer is 64% so short life expectancy is predicted than other diseases.
PhraseX: cancer. *PhraseY*: short life expectancy.

(Clarification) The explanation is an overgeneralization because the survival rates for several less dangerous cancers are higher.

(Candidates for variants) *PhraseX*: Alzheimer's disease, periodontal disease, heart disease, pneumonia. *PhraseY*: long life expectancy.
We hesitated between short and long in *PhraseY* but finally chose the former as is is more credible than the latter. The term "than other diseases" has been added as without it one would compare cancer patients with people with no disease. Note that this type is kept to show the difficulty of the target problem despite its flaw, i.e., the statistical data do not contain survival rates of other diseases, which reflects the difficulty of discussing the remaining diseases all at once.

[2]We omitted Americas, which is diverse.
[3]We initially considered methane and freon gas emissions as variants but gave up as they are known to largely depend on other factors than population.
[4]We repeated the word "hospitals" to correctly measure the relevance between *PhraseX* and *PhraseY*.

4.2 Three Judgement Methods

We devise three judgement methods (α), (β) and (γ) for our target problem. They mainly assess the credibility of the 18 types of explanations and their variants. Detection method (α) is devised to judge explanation (I) on a habit *PhraseX* and a disease *PhraseY*. The credibility of explanation (I), unlike others can be decomposed into those of *PhraseX*, *PhraseY*, and the relevance between them.

Detection method (β) is devised to judge explanations (II), (IV)-(VII), (XII), and (XIV)-(XVIII), which describe the subject *PhraseX* has a property *PhraseY* compared to many other subjects. The credibility of these 11 types of explanations can be decomposed into those of the relevance between *PhraseX* and *PhraseY*, and *PhraseX* compared to its typical variants. The variant is either explicitly specified in the explanation or can be imagined implicitly.

Except explanation (II), the property is represented by essentially an adjective followed by a noun phrase. Note that explanation (II) has a property *PhraseY* in a form of the superlative of an adjective. Thus we will modify the detection method (β) for explanation (II).

Detection method (γ) is devised to judge explanations (III), (VIII)-(XI), and (XIII), which mostly describe the subject *PhraseX* has a trend *PhraseY* unlike other kinds of trends[5]. The credibility of these 6 types of explanations corresponds to that of *PhraseY* compared to its typical variants. The variants are usually clear from *PhraseY*. Thus, comparing relevant trends will be the key for a successful judgement.

Recall that the class of credible and unethical (class 1) depends on the phrases used in the explanation. Thus the three methods all employ relevance degrees as the basis of their judgments. Each relevance degree is either a semantic similarity between a pair of phrases or a ratio of such semantic similarities. The semantic similarity is a cosine-similarity of the embedded vectors of the phrases by SBERT [28]. Specifically the semantic similarity $\mathrm{Sim}(\cdot)$ between *PhraseX* and *PhraseY* is given as follows.

$$\mathrm{Sim}(PhraseX,\ PhraseY) = \frac{s(PhraseX)\cdot s(PhraseY)}{\|s(PhraseX)\|\,\|s(PhraseY)\|}, \quad (1)$$

where $s(\cdot)$ represents the embedding vector by SBERT.

4.2.1 Detection method (α). This method judges an explanation as credible and unethical if and only if the habit *PhraseX* is bad, the disease *PhraseY* is dangerous, and the two are highly relevant, which correspond to the three kinds of credibility.

$$\begin{aligned}&\text{IF } (\theta_{\text{relevance}} > \theta_1) \wedge (\theta_{\text{fear}} > \theta_2) \wedge (\theta_{\text{bad habit}} > \theta_3)\\ &\text{THEN 1}\\ &\text{ELSE 0,}\end{aligned} \quad (2)$$

where $\theta_1, \theta_2, \theta_3$ are user-supplied thresholds and $\theta_{\text{relevance}}$, θ_{fear}, $\theta_{\text{bad habit}}$ are the relevance ratio, the fear ratio, and the bad habit ratio, respectively, given below. For instance, explanation (I) is judged credible and unethical because eating deep-fried food is a bad habit, pancreatic cancer is a dangerous disease, and the two seem to be highly relevant.

$\theta_{\text{relevance}}$ represents the relevance between *PhraseX* and *PhraseY*. The values of the semantic similarity $\mathrm{Sim}(\cdot)$ largely depend on its arguments, i.e., *PhraseX* and *PhraseY*, which forbids to use the same value for θ_1 in different explanations. To mitigate this variety,

[5]We use the term "mostly" as type (X) has no subject as its *PhraseX* is also a trend.

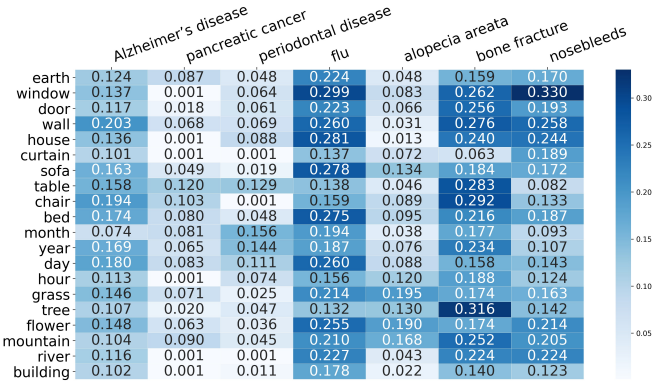

Figure 3: Relevance degrees between diseases (*PhraseY*) and base words for *PhraseX* for method (α).

we define $\theta_{\text{relevance}}$ as follows, by selecting a base word of which semantic similarity is neutral to major diseases.

$$\theta_{\text{relevance}} = \frac{\mathrm{Sim}(PhraseX,\ PhraseY)}{\mathrm{Sim}(\text{base word}, PhraseY)} \quad (3)$$

We conducted preliminary experiments on the relevance degrees between a disease and such words. The results are shown in Figure 3, where each row and column represent a candidate of the base word and a major disease[6], respectively. We see that "day" and "earth" have neither small nor large degrees to all the tested diseases. We select the latter as the base word. Similar degrees of relevance using base words can be found in a sentiment analysis paper [3]. However, Araque et al. adopted the maximum value for a set of base words for each lexicon word for normalization [3]. On the other hand, we select a base word of which value is not extreme for all diseases to use it in Eq. (3). This difference is due to the nature of the two target problems.

$\theta_{\text{bad habit}}$ represents the bad habit ratio of *PhraseX*.

$$\theta_{\text{bad habit}} = \frac{\mathrm{Sim}(PhraseX,\ \text{"bad habit"})}{\mathrm{Sim}(PhraseX,\ \text{"good habit"})} \quad (4)$$

Note that $\theta_{\text{bad habit}}$ compares the closeness of the habit in *PhraseX* to the phrase "bad habit" than "good habit" through the ratio of the two semantic similarities.

Finally, θ_{fear} represents the fear ratio of *PhraseY*. We first devised a similar definition as follows.

$$\theta_{\text{fear}} = \frac{\mathrm{Sim}(PhraseY,\ \text{"major illness"})}{\mathrm{Sim}(PhraseY,\ \text{"minor illness"})} \quad (5)$$

However a series of preliminary experiments proved that Eq. (5) shows quite counter-intuitive results, probably due to our highly-variable subjectivity in assessing the risk of diseases[7]. Thus we use a summary (GBD Cause and Risk Summaries in https://www.thelancet.com/gbd/summaries) of DALYs (Disability Adjusted Life Years) in Burden of Disease (https://ourworldindata.org/burden-of-disease) as θ_{fear}. DALYs measures the loss in health quantitatively,

[6]The diseases are sorted from the heaviest one to the lightest one from the leftmost to the rightmost with a method explained later.

[7]For instance, we have witnessed a young man and a middle-aged man having very different opinions on alopecia areata.

allowing us to compare different diseases and other kinds of damages. DALYs is equivalent to losing one year in good health because of either premature death or disease or disability. One DALYs represents one lost year of healthy life. DALYs is calculated by the sum of Years of Life Lost (YLL) and Years Lost due to Disability (YLD).

4.2.2 Detection method (β). An explanation of types (IV)-(VII), (XII), and (XIV)-(XVIII) states that subject $PhraseX$ possesses property $PhraseY$, where $PhraseY$ is essentially an adjective followed by a noun phrase. This method judges an explanation as credible and unethical if and only if the following two conditions hold. The first one is that $PhraseX$ is more relevant to $PhraseY$ than its inverse $Phrase\overline{Y}$. The second one is that the property $PhraseY$ is more relevant to the subject $PhraseX$ than any other subject $PhraseX^{'}$ belonging to the opposite class.

$$\text{IF } (\theta^{\beta}_{XY} > \theta^{\beta}_{X\overline{Y}}) \wedge \forall X^{'} (\theta^{\beta}_{XY} > \theta^{\beta}_{X^{'}Y}) \text{ THEN 1} \\ \text{ELSE 0} \tag{6}$$

We implement the first condition in Eq. (6) as the relevance degree θ^{β}_{XY} between $PhraseX$ and $PhraseY$ being larger than the relevance degree $\theta^{\beta}_{X\overline{Y}}$ between $PhraseX$ and $Phrase\overline{Y}$. Specifying $Phrase\overline{Y}$ for each explanation is straightforward, as we only have to choose essentially an opposite adjective.

Note that the variety problem of the semantic similarity $\text{Sim}(\cdot)$ also exists in this method (β) and the meaning of an adjective is decided by the following noun phrase. We thus select the noun in $PhraseY$ as $PhraseY_{\text{base}}$. The relevance degrees are defined as follows.

$$\theta^{\beta}_{XY} = \frac{\text{Sim}(PhraseX, PhraseY)}{\text{Sim}(PhraseX, PhraseY_{\text{base}})} \tag{7}$$

$$\theta^{\beta}_{X\overline{Y}} = \frac{\text{Sim}(PhraseX, Phrase\overline{Y})}{\text{Sim}(PhraseX, PhraseY_{\text{base}})} \tag{8}$$

$$\theta^{\beta}_{X^{'}Y} = \frac{\text{Sim}(PhraseX^{'}, PhraseY)}{\text{Sim}(PhraseX^{'}, PhraseY_{\text{base}})} \tag{9}$$

Note that $PhraseY_{\text{base}}$ serves as a base in comparison. $PhraseY_{\text{base}}$ are "babies" for (IV), "math score" for (V), "children" for (VI), "infant mortality rates" for (VII), "CO2 emissions" for (XII) and (XIV), "hospitals" for (XV), "strain" for (XVI), "GDP" for (XVII), and "life expectancy" for (XVIII).

Recall explanations (II-1) and (II-2) each has $PhraseY$ in a form of the superlative of an adjective. Hence we have no $PhraseY_{\text{base}}$ in these types. We simply adopt the similarity measure instead of the ratio of the similarity measures in Eqs. (7) - (9).

Likewise, the second condition in Eq. (6) is implemented as θ^{β}_{XY} being larger than $\theta^{\beta}_{X^{'}Y}$. $PhraseX^{'}$ cannot be any value otherwise only the most relevant subject $PhraseX$ is judged as credible and unethical. Here we adopt the closed-world assumption and specify $\{PhraseX^{'}\}$ for each $PhraseX$. We will explain the assignment in the Appendix.

Similar ideas can be found in human bias papers [15, 16, 33], which use distinct groups of subjects to investigate the target group. Unlike these works, we use a distinct group to evaluate the credibility of an explanation through the second condition of Eq. (6).

For instance, explanation (XII) is judged credible and unethical because the above conditions are satisfied for $PhraseX$: Asia, $PhraseY$: large amount of CO2 emissions, $Phrase\overline{Y}$: small amount of CO2 emissions, $PhraseY_{\text{base}}$: CO2 emissions, and $PhraseX^{'} \in$ {Africa, Europe}.

4.2.3 Detection method (γ). An explanation of types (III), (VIII)-(XI), (XIII) describes the subject $PhraseX$ has a trend $PhraseY$, where $PhraseY$ is essentially a phrase with a verb in the present participle tense. This method judges an explanation as credible and unethical if and only if the subject $PhraseX$ is more relevant to trend $PhraseY$ than any different trend $PhraseY^{'}$.

We implement the condition as the relevance degree θ^{γ}_{XY} between subject $PhraseX$ and trend $PhraseY$ being larger than relevance degree $\theta^{\gamma}_{XY^{'}}$ between $PhraseX$ and any other relevant trend $PhraseY^{'}$.

$$\text{IF } \forall Y^{'} (\theta^{\gamma}_{XY} > \theta^{\gamma}_{XY^{'}}) \text{ THEN 1} \\ \text{ELSE 0} \tag{10}$$

We typically specify the relevant trends in $PhraseY^{'}$ by replacing the verb or the adverb used in $PhraseY$ with the opposite one, keeping other words as they are[8]. Consequently, the relevance degrees in Eq. (10) are defined as the corresponding relevance degrees, as the variety problem is not serious unlike methods (α) and (β) due to the forms of $PhraseY$ and $PhraseY'$.

$$\theta^{\gamma}_{XY} = \text{Sim}(PhraseX, PhraseY) \\ \theta^{\gamma}_{XY^{'}} = \text{Sim}(PhraseX, PhraseY^{'}) \tag{11}$$

For instance, explanation (XIII) is judged credible and unethical because the above conditions are satisfied for $PhraseX$: "risk of death from cancer", $PhraseY$: "increasing", and $PhraseY^{'} \in$ {"decreasing", "constant"}.

5 EXPERIMENTS

We choose a SBERT model named "all-mpnet-base-v2"[9] trained on a large amount of data (more than 1 billion training pairs) which can map each phrase to a 768 dimensional dense vector. In detection method (α), the thresholds $\theta_1, \theta_2, \theta_3$ are all set to 1.

Table 1 shows the confusion matrices of our three methods. The accuracies of our detection methods (α), (β) and (γ) on the 18 types of explanations are 1, 0.510, and 0.914, respectively. We see that (α) and (γ) exhibit relatively high accuracies, probably due to the simpler nature of their target explanations. Though (β) needs a substantial refinement, we believe that the results are quite promising as the first step toward judging credible and unethical explanations of statistical data.

We show detailed results for the three methods in the Appendix. In summary, our method (α) exhibits the perfect results, which proves the effectiveness of our proposed ratios $\theta_{\text{relevance}}$, θ_{fear} and $\theta_{\text{bad habit}}$ for representing the semantic relatedness between diseases and habits. Likewise, our method (γ) exhibits accuracy close to 92%. Most of the mistakes (6 out of 7) are committed on explanation (VIII), possibly due to the difficulty in handling a negated

[8] We used the term "typically" as we also replace a specific trend with a neutral trend. See the last example in this section.

[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Table 1: Results of methods (α), (β), and (γ)

(α)	Predicted Positive	Predicted Negative	(β)	Predicted Positive	Predicted Negative	(γ)	Predicted Positive	Predicted Negative
Actual Positive	9	0	Actual Positive	12	32	Actual Positive	13	4
Actual Negative	0	19	Actual Negative	13	35	Actual Negative	3	62

phrase, i.e., "not decreasing". With a hindsight, we see that changing "not decreasing" in *PhraseY* from class label 1 to 0 reduces 3 mistakes and "increasing" from class label 0 to 1 reduces 1 mistake. We do not adopt such a modification but learn the difficulty in handling these expressions and capturing the credibility of people. In overall, however, our method (γ) achieves impressive results of close to 92 %, which proves the effectiveness of our approach.

On the contrary, the performance of our method (β) seems to be unacceptable because it is below the accuracy (0.521) of the baseline method which always predicting the majority class label 0. A closer look at the performance of each type reveals a binary classification of the 11 types. The first group consists of easier types, i.e., (V) (7-1), (XIV) (7-1), (II-1) (5-3), (VI) (5-3), (XVII) (5-3), and (XII) (4-2), where the numbers in parentheses represent correct and wrong predictions in this order. The accuracy of our method (β) is 0.717 (33/46), which we believe relatively high due to the difficulty of the target problems. On the other hand, the second group consists of more difficult types, i.e., (IV) (3-3), (XVIII) (4-6), (II-2) (3-5), (VII) (3-5), (XVI) (1-9), and (XV) (0-4). The accuracy of our method (β) is 0.304 (14/46).

As we explained in Section 4.1, type (XVIII) has serious a flaw, i.e., the difficulty of discussing the remaining diseases all at once, but was kept to show the difficulty of the problem. Type (XVI) also shows the difficulty in handling a serious issue related to the recent pandemic, which hasn't been clarified scientifically and is a subject of fierce debate. Type (XV) also poses a challenge because small hospitals are in general less well-equipped but receives fewer serious patients than large hospitals. Their degrees of safety are controversial, which might have influenced our phrase embedding. Omitting these three types yields accuracy 0.617 (42/68) for the remaining types, which we believe promising.

A closer look at the remaining difficult types revealed challenges on our semantic similarity Sim and relevance degree $\theta^{\beta}_{X\overline{Y}}$. In type (II-2), 4 out of the 5 mistakes were due to the fact that Sim("Somalia", "richest") = 0.29137 being larger than that for UAE (0.29136), Qatar (0.244), Equatorial Guinea (0.238), and Botswana (0.271). These results are against the fact shown in Fig. 1 II. The remaining mistake was caused by our choice of Japan and Singapore for *PraseX′* in "Botswana is the poorest of the healthiest countries" (class 0). The semantic similarities with "poorest" are 0.179, 0.162, and 0.278 for Japan, Singapore, and Botswana, respectively, which match the fact in Fig. 1 II. However, this explanation was labeled as 0 due to the numerous countries that have longer life expectancy (healthier) and smaller GDP per capita (poorer) than Botswana and not due to Japan and Singapore. We believe our choice of *PraseX′* is correct, as the two countries are representatives of the healthiest ones. We thought of comparing Botswana with countries of similar life expectancy, but gave it up as they wouldn't recognized as "the healthiest countries". Similarly, the majority of the mistakes in types (IV) and (VII) were due to semantic similarity that are against our class labeling, e.g., Muslims are more similar to few babies than many babies, Christians are more similar to many babies than few babies. The remaining mistakes were due to the division by the semantic similarity to the base word, e.g., Sim("Judaisms", "babies") = 0.278 is small and thus boosts the relevance degree while that for Christians (0.446) and Muslims (0.437) don't. Note that our semantic similarity Sim and relevance degree $\theta^{\beta}_{X\overline{Y}}$ are effective for many explanations, which proves the difficulty in handling these exceptions.

In summary, the examples of counter-intuitive semantic similarities may be attributed to our phrase embeddings which are directly generated from the pre-trained SBERT. SBERT utilizes multiple language datasets for training, without fine-tuning on task-specific corpus. Thus there may exist several phrase embeddings that harm our credibility judgment task. In general, the relevance degree defined by the semantic similarity exhibits encouraging performance on judging the credibility of the explanations on statistical data. The underlying semantic relatedness between phrases is worth exploring in the next step[10].

6 CONCLUSIONS

In this paper we have investigated exploitation of ten instincts [30] in statistical data explanations as a first yet important step toward ethical AI. Our goal is not in abusing our investigation but to prevent such unethical conducts through deep understanding. Our 18 prototypes together with their variants, our three kinds of judgement methods, and our experiments serve as a milestone toward the goal.

This paper opens promising avenues for further research. Beyond the judgement methods, transforming a credible and unethical explanation to a correct one represents a challenging and yet important problem. A fully automatic generation of credible and unethical explanations will be the next step, though our earlier investigations with generative deep neural networks knew limited success. Large-scale cognitive experiments on the credibility of the variants of each explanation is definitively highly rewarding. Targeting at the right population is the key to success, though the authors feel that our communities lack diversity in this respect. Web services such

[10]Two technical problems are related with synonyms and singular/plural forms. In type (V), we initially used "boys" and "girls" for "men" and "women", respectively. We decided to use more official nouns as we believe the examinees are not usually called boys and girls at their ages. This replacement boosted the performance to 7-1 from 3-5. In types (IV), (VII), and (XV), we use the form used in the explanation. We also tried to use singular forms to the nouns in our phrases and obtained different performance. Note that these facts cannot be generalized to other cases.

as Amazon Mechanical Turk (https://www.mturk.com/) provide a powerful solution, though a careful design is mandatory [12].

ACKNOWLEDGMENTS

A part of this work was supported by JSPS KAKENHI Grant Number JP21K19795. The first author is supported by China Scholarship Council (Grant No. 201906330075).

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Mandelbaum Amit and Shalev Adi. 2016. Word Embeddings and their Use in Sentence Classification Tasks. arXiv:1610.08229
[3] O. Araque, G. Zhu, and C. A. Iglesias. 2019. A Semantic Similarity-Based Perspective of Affect Lexicons for Sentiment Analysis. *Knowledge-Based Systems* 165 (2019), 346––359.
[4] Michael Blastland and David Spiegelhalter. 2014. *The Norm Chronicles: Stories and Numbers about Danger.* Basic Books, London.
[5] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring Semantic Similarity between Words Using Web Search Engines. In *Proc. WWW*. 757–766.
[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science* 356, 6334 (2017), 183–186.
[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal Sentence Encoder. arXiv:1803.11175
[8] Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of Semantic Similarity - A Survey. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–37.
[9] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-Training Tasks for Embedding-Based Large-Scale Retrieval. In *Proceedings of the 8th International Conference on Learning Representations*.
[10] Rudi L. Cilibrasi and Paul M.B. Vitanyi. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19, 3 (2007), 370–383.
[11] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 670–680.
[12] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3 (2013).
[13] Andrew M. Dai and Quoc V. Le. 2015. Semi-Supervised Sequence Learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 3079–3087.
[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805
[15] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635––E3644.
[16] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 122–133.
[17] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1986. Distributed Representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, Maas., 77–109.
[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation* 9, 8 (1997), 1735–1780.
[19] Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. 2017. Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping PubMed Queries to Documents. *Journal of Biomedical Informatics* 75 (2017), 122–127.
[20] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning*. 1188–1196.
[21] Yuhua Li, Zuhair A. Bandar, and David McLean. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering* 15, 4 (2003), 871–882.
[22] Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable Semantic Textual Similarity: Finding and Explaining Differences between Sentences. *Knowledge-Based Systems* 119 (2017), 186–199.
[23] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. 1751–1754.
[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781
[25] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. ACL, 1532–1543.
[26] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. arXiv:1802.05365
[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1, 8 (2019), 9.
[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 3980–3990.
[29] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81.
[30] Hans Rosling, Ola Rosling, and Anna Rosling Roennlund. 2018. *Factfulness: Ten Reasons We're Wrong about the World - and Why Things are Better than You Think.* Sceptre, London.
[31] Ola Rosling, Anna Rosling Rönnlund, and Hans Rosling. 2005. Gapminder Download the Data. https://www.gapminder.org/data/.
[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[33] Y. C. Tan and L. E. Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Proc. NeurIPS*. 13209–13220.
[34] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task Definition and Dataset Construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. 18–22.
[35] Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng-Lin Liu, and Hongwei Hao. 2016. Semantic Expansion Using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification. *Neurocomputing* 174 (2016), 806–814.
[36] Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10837–10851.
[37] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 422–426.
[38] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 637–645.
[39] Tomáš Zemčík. 2021. Failure of Chatbot Tay was Evil, Ugliness and Uselessness in its Nature or Do We Judge it through Cognitive Shortcuts and Biases? *AI & SOCIETY* 36, 1 (2021), 361–367.
[40] Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1393–1398.

A DETAILED EXPERIMENTAL RESULTS

We show detailed results for the three methods in Tables 2 - 4. In the Tables, a phrase in parentheses represents that the explanation belongs to class 0. The authors discussed carefully in assigning a class label, e.g., both explanations on English score for men in type (V) were judged class 0, as the authors agree that it is widely known that men who are good at English exist as those who are not[11].

FP and FN represent that the explanation is judged as a false positive and a false negative, respectively. A blanc in Result column either represents a true positive or a true negative[12].

In our method (β), as we explained in Section 4.2.2, we specify $PhraseX'$ for each $PhraseX$. We denote the assignment in the form of ($PhraseX$, $PhraseX'$), though a phrase could be a set of phrases

[11] This example highlights the prejudice behind the math score example for women.
[12] Their distinction is easy due to the parentheses.

Table 2: Results by method (α)

Type	PhraseX	PhraseY	$\theta_{\text{relevance}}$	θ_{fear}	$\theta_{\text{bad habit}}$	Result
I	deep-fried food	pancreatic cancer	2.309	11.5	2.070	
28-0		Alzheimer's disease	1.968	25.3	2.070	
		periodontal disease	1.715	7.10	2.070	
		(flu)	0.941	0.00	2.070	
		(alopecia areata)	2.736	0.60	2.070	
		(bone fracture)	1.374	0.00	2.070	
		(nosebleeds)	1.363	0.00	2.070	
	alcohol abuse	pancreatic cancer	1.921	11.5	1.755	
		Alzheimer's disease	4.211	25.3	1.755	
		periodontal disease	4.055	7.10	1.755	
		(flu)	1.474	0.00	1.755	
		(alopecia areata)	4.522	0.60	1.755	
		(bone fracture)	1.852	0.00	1.755	
		(nosebleeds)	1.933	0.00	1.755	
	heavy drinking	pancreatic cancer	1.624	11.5	1.528	
		Alzheimer's disease	3.557	25.3	1.528	
		periodontal disease	2.965	7.10	1.528	
		(flu)	1.656	0.00	1.528	
		(alopecia areata)	3.159	0.60	1.528	
		(bone fracture)	1.893	0.00	1.528	
		(nosebleeds)	1.962	0.00	1.528	
	long distance running	(pancreatic cancer)	0.048	11.5	0.922	
		(Alzheimer's disease)	0.509	25.3	0.922	
		(periodontal disease)	-0.133	7.10	0.922	
		(flu)	0.769	0.00	0.922	
		(alopecia areata)	1.415	0.60	0.922	
		(bone fracture)	1.509	0.00	0.922	
		(nosebleeds)	1.262	0.00	0.922	

to save space. Refer to Table 3 for the following assignments. In type (IV), ({Muslims, Judaisms}, Christians) and vice versa. (V), (men, women) and vice versa. (VI), ({Iranians, Afghans}, Americans) and ({Americans, French}, Iranians). Note that in this type $PhraseX^{'}$ is specified with the word "than" to one country. (VII), (developing countries, advanced countries) and vice versa. Note that $PhraseX^{'}$ is specified in the text. (XII), (Asia, {Africa, Europe}) and vice versa. (XIV), ({China, India, United States}, United Kingdom) and vice versa. Note that UK is considered to be the representative of the countries with a small amount of CO2 emissions in the current era. (XV), (large hospitals, small hospitals) and vice versa. (XVI), ({Omicron strain, Alpha strain, Beta strain, Gamma strain}, Delta strain) and vice versa. (XVII), (Africa, {Asia, Americas, Europe}), (Asia, {Africa, Americas, Europe}), (Americas, {Africa, Asia, Europe}), and (Europe, {Africa, Asia, Americas}). Note that this assignment is a result of using the expression "than other regions" in the explanation. (XVIII), ({cancer, Alzheimer's disease, heart disease, pneumonia}, periodontal disease) and vice versa. Note that this assignment follows a similar reason to (XIV).

Note that in types (II-1) and (II-2), $PhraseX^{'}$ is rather specified at the end of the explanation as "of the healthiest" or "of the unhealthiest", which has a fixed correspondence to $PhraseY$, i.e., "poorest" or "richest", respectively. Thus for these types we denote the assignment in the form of ($PhraseY$, $PhraseX^{'}$), which are (poorest, {Japan, Singapore}) and (richest, {Central African Republic, Somalia}). These 4 countries are selected as the representatives of the healthiest or unhealthiest as they are located in the upper-right or lower-left corners in Fig. 1 II, respectively.

In our method (γ), as we explained in Section 4.2.3, we specify $PhraseY^{'}$ for each $PhraseY$. The assignment is straightforward, as for each $PhraseY$, $PhraseY^{'}$ consists of its variants belonging to the opposite class. For instance, as shown in Table 4, in type (III), when $PhraseY$ is "proportional to GDP", $PhraseY^{'}$ is {"inversely proportional to GDP", "not correlated to GDP"} and vice versa.

Table 3: Results by method (β), where MR, ER, and CO2E represent mortality rates, enrollment rates, and CO2 emissions, respectively.

Type	PhraseX	PhraseY	θ^{β}_{XY}	Result	Type	PhraseX	PhraseY	θ^{β}_{XY}	Result
II-1	Cuba	poorest	0.273		XII	Asia	large amount of CO2E	0.848	
5-3		(richest)	0.266		4-2		(small amount of CO2E)	0.227	
	Nicaragua	poorest	0.245			Africa	small amount of CO2E	0.193	FN
		(richest)	0.192				(large amount of CO2E)	0.534	
	Bangladesh	poorest	0.267	FN		Europe	small amount of CO2E	0.216	FN
		(richest)	0.284				(large amount of CO2E)	0.812	
	North Korea	poorest	0.258	FN	XIV	China	large amount of CO2E	0.951	
		(richest)	0.314	FP	7-1		(small amount of CO2E)	0.265	
II-2	United Arab Emirates	richest	0.291	FN		India	large amount of CO2E	0.741	
3-5		(poorest)	0.185				(small amount of CO2E)	0.435	
	Qatar	richest	0.244	FN		United States	large amount of CO2E	0.520	FN
		(poorest)	0.223				(small amount of CO2E)	0.143	
	Equatorial Guinea	richest	0.238	FN		United Kingdom	(small amount of CO2E)	0.481	
		(poorest)	0.180				(large amount of CO2E)	0.570	
	Botswana	richest	0.271	FN	XV	small hospitals	dangerous hospitals	0.826	FN
		(poorest)	0.278	FP	0-4		(safe hospitals)	0.861	FP
IV	Muslims	many babies	0.536	FN		large hospitals	safe hospitals	0.860	FN
3-3		(few babies)	0.561	FP			(dangerous hospitals)	0.867	FP
	Judaisms	many babies	0.621		XVI	Omicron strain	less dangerous	1.012	FN
		(few babies)	0.570		1-9		(more dangerous)	1.138	FP
	Christians	few babies	0.552	FN		Alpha strain	less dangerous	0.855	FN
		(many babies)	0.526				(more dangerous)	0.935	FP
V	women	low math score	0.020	FN		Beta strain	less dangerous	1.023	FN
7-1		(high math score)	0.299				(more dangerous)	1.155	FP
	men	high math score	0.327			Gamma strain	less dangerous	0.812	FN
		(low math score)	0.163				(more dangerous)	0.845	FP
	women	(low English score)	0.106			Delta strain	more dangerous	0.672	FN
		high English score	0.270				(less dangerous)	0.709	
	men	(high English score)	0.199		XVII	Africa	low GDP	0.809	FN
		(low English score)	0.081		5-3		(high GDP)	0.875	
VI	Iranians	many children	0.597	FN		Asia	high GDP	1.009	
5-3		(few children)	0.616	FP			(low GDP)	0.709	
	Afghans	many children	0.634			Americas	high GDP	0.884	FN
		(few children)	0.593				(low GDP)	0.699	
	French	few children	0.670			Europe	high GDP	0.955	FN
		(many children)	0.662				(low GDP)	0.776	
	Americans	few children	0.505	FN	XVIII	cancer	(long life expectancy)	0.910	
		(many children)	0.531		4-6		short life expectancy	1.027	FN
VII	developing countries	high infant MR	1.264	FN		Alzheimer's disease	(long life expectancy)	0.972	
3-5		(low infant MR)	1.365	FP			short life expectancy	1.108	FN
	advanced countries	low infant MR	1.355	FN		heart disease	(long life expectancy)	0.901	
		(high infant MR)	1.407	FP			short life expectancy	1.066	FN
	developing countries	low ER	1.206	FN		pneumonia	(long life expectancy)	0.926	
		(high ER)	1.272				short life expectancy	1.107	FN
	advanced countries	high ER	1.484			periodontal disease	(short life expectancy)	1.500	FP
		(low ER)	1.033				long life expectancy	1.224	FN

Table 4: Results by method (γ), where ND, ED, and IA, represent natural disasters, epidemic damages, and industrial accidents, respectively.

Type	PhraseX	PhraseY	θ^{γ}_{XY}	Result
III	life expectancy	proportional to GDP	0.144	
6-0		(inversely proportional to GDP)	0.033	
		(not correlated to GDP)	0.079	
	healthy life expectancy	proportional to GDP	0.168	
		(inversely proportional to GDP)	0.059	
		(not correlated to GDP)	0.150	
VIII	child labor	not decreasing	0.061	FN
9-6		(decreasing)	0.062	
		(not increasing)	0.090	
		(increasing)	0.111	FP
		(constant)	0.111	
	child hunger	not decreasing	0.146	FN
		(decreasing)	0.148	
		(not increasing)	0.187	
		(increasing)	0.177	
		(constant)	0.188	FP
	child mortality	not decreasing	0.195	FN
		(decreasing)	0.206	
		(not increasing)	0.207	
		(increasing)	0.217	FP
		(constant)	0.127	
IX	world population	will just increase	0.203	FN
4-1		will rapidly increase	0.221	
		(will just decrease)	0.130	
		(will rapidly decrease)	0.126	
		(will keep constant)	0.211	
X	increasing	increasing in deaths from ND	0.898	
27-0	in ND	(decreasing in deaths from ND)	0.820	
		(constant deaths from ND)	0.650	
	decreasing in ND	(increasing in deaths from ND)	0.793	
		(decreasing in deaths from ND)	0.896	
		(constant deaths from ND)	0.638	
	constant ND	(increasing in deaths from ND)	0.636	
		(decreasing in deaths from ND)	0.563	
		(constant deaths from ND)	0.844	
	increasing in ED	increasing in deaths from ED	0.940	
		(decreasing in deaths from ED)	0.851	
		(constant deaths from ED)	0.801	
	decreasing in ED	(increasing in deaths from ED)	0.881	
		(decreasing in deaths from ED)	0.934	
		(constant deaths from ED)	0.787	
X	constant ED	(increasing in deaths from ED)	0.838	
		(decreasing in deaths from ED)	0.778	
		(constant deaths from ED)	0.915	
	increasing in IA	increasing in deaths from IA	0.945	
		(decreasing in deaths from IA)	0.891	
		(constant deaths from IA)	0.791	
	decreasing in IA	(increasing in deaths from IA)	0.862	
		(decreasing in deaths from IA)	0.942	
		(constant deaths from IA)	0.752	
	constant IA	(increasing in deaths from IA)	0.764	
		(decreasing in deaths from IA)	0.714	
		(constant deaths from IA)	0.879	
XI	death of many babies	increasing	0.133	
20-0		(decreasing)	0.114	
		(not increasing)	0.129	
		(not decreasing)	0.112	
		(constant)	0.064	
	death of many children	increasing	0.129	
		(decreasing)	0.102	
		(not increasing)	0.116	
		(not decreasing)	0.092	
		(constant)	0.076	
	death of many adults	increasing	0.169	
		(decreasing)	0.114	
		(not increasing)	0.158	
		(not decreasing)	0.141	
		(constant)	0.052	
	death of many old people	increasing	0.156	
		(decreasing)	0.122	
		(not increasing)	0.138	
		(not decreasing)	0.130	
		(constant)	0.030	
XIII	risk of death from cancer	increasing	0.082	
9-0		(decreasing)	0.033	
		(constant)	-0.012	
	risk of death from Alzheimer's disease	increasing	0.064	
		(decreasing)	0.018	
		(constant)	-0.042	
	risk of death from heart disease	increasing	0.094	
		(decreasing)	0.063	
		(constant)	0.006	