A Brief History of Risk

Ying Li,[a] Thomas Hills,[b] Ralph Hertwig[a]

[a]Center for Adaptive Rationality, Max Planck Institute for Human Development, 14195 Berlin, Germany; [b]Department of Psychology, University of Warwick, University Road, Coventry, CV4 7AL, United Kingdom

We thank Susannah Goss for editing the manuscript.

Corresponding Author: Ying Li
Contact: li@mpib-berlin.mpg.de

Abstract

Despite increasing life expectancy and high levels of welfare, health care, and public safety in most post-industrial countries, the public discourse often revolves around perceived threats. Terrorism, global pandemics, and environmental catastrophes are just a few of the risks that dominate media coverage. Is this public discourse on risk disconnected from reality? To examine this issue, we analyzed the dynamics of the risk discourse in two natural language text corpora. Specifically, we tracked latent semantic patterns over a period of 150 years to address four questions: First, we examined how the frequency of the word *risk* has changed over historical time. Is the construct of risk playing an ever-increasing role in the public discourse, as the sociological notion of a 'risk society' suggests? Second, we investigated how the sentiments for the words co-occurring with *risk* have changed. Are the connotations of *risk* becoming increasingly ominous? Third, how has the meaning of *risk* changed relative to close associates such as *danger* and *hazard*? Is *risk* more subject to semantic change? Finally, we decompose the construct of *risk* into the specific topics with which it has been associated and track those topics over historical time. This brief history of the semantics of risk reveals new and surprising insights—a fourfold increase in frequency, increasingly negative sentiment, a semantic drift towards forecasting and prevention, and a shift away from war toward chronic disease—reflecting the conceptual evolution of risk in the archeological records of public discourse.

Keywords: risk, danger, public discourse, content analysis, topic model, Ngram Corpus

A Brief History of the Semantics of Risk

1. Introduction

Humans have always been exposed to risks. Yet the nature of these risks has changed profoundly over the course of human biological and cultural evolution. Whereas the dominant risks were once starvation, infections, and violent conflict (Harari, 2016), many of today's risks are associated with lifestyle choices (e.g., obesity, cardiovascular disease, cancer). Although modern institutions such as hospitals, police and fire services, and international treaties now buffer people in industrialized nations from the worst consequences of risks, the "consequences of modernity" (Giddens, 1990) include new risks, such as nuclear weapons, global pandemics, deadly hospital bugs, terrorism, cyberattacks, and climate change. As we write this text, the world has been rocked by the coronavirus pandemic and Australia is recovering from its worst bush fires in recorded history. Over the past two centuries, however, rates of violent conflict, poverty, and starvation have decreased (Pinker, 2011) and life expectancy has doubled (Oeppen & Vaupel, 2002). Yet many people appear to feel that the world is more rife with dangers than ever (see Pinker, 2011): The historian Bourke (2005) has argued that "fear is the most pervasive emotion of modern society," and life in today's "risk society" (Beck, 1992) has been characterized as increasingly vigilant to a growing variety of risks and insecurities (e.g., the precautionary principle; Sunstein, 2005). It has been suggested that the "current climate of fear" (Stearns, 2012, p. x) is fueled at least partly by a range of players (e.g., politicians, media, federal agencies, businesses) who are desperate to capture public attention and are willing to use fear as a method to do so. The idea that people are more afraid then they used to be is also a regular topos in the cultural discourse (Rothman, 2016). Whether this is actually the case remains unclear, however: data on a population's fear level and risk perception only go back so far, and survey responses are influenced by the cultural context. It has been argued that fear is currently "in some ways slightly fashionable, so maybe people are even exaggerating a little bit" (Stearns in Rothman, 2016).

How does society identify risks? Cultural anthropologists and sociologists have emphasized that risks are not a natural kind but are socially constructed, based on norms, moral considerations, and structures of social organization (Douglas, 1992). What qualifies as a risk is therefore subject to dynamic social change. Religiously motivated terrorism is a striking example of how an "old" risk can transform into a new phenomenon and forcefully reappear on the collective radar. Bourke (2005) has documented a history of fears, from the Victorians' dread of being buried alive to the more recent fear of nuclear annihilation. These fears are preserved in cultural artifacts such as books and newspaper articles—records that provide insights into how risks are collectively identified and perceived. Taking a historical perspective on these artifacts reveals how and why society's attitudes to risk have changed and may indicate how they will change again in the future. Our goal in this study is to take a large-scale quantitative approach to the recent historical trajectory of the word *risk* with the aim of understanding the changing nature of its social construction.

Before we turn to our research questions, let us clarify that the term *risk* is often used to mean different things. In the risk management and actuarial literature, for instance, it describes a loss of a certain magnitude (e.g., injury, mortality) weighted by the probability of its occurrence (Rayner & Cantor, 1987; Short, 1984). By this actuarial measure, driving is riskier than flying because it is associated with a greater risk of injury per mile travelled. In the economic discourse, risk commonly refers to the variance in possible (positive or negative) returns. For instance, an investment option with higher return variance is deemed as riskier than an option with lower variance but the same expected mean return (Markowitz, 1952; Pratt 1964). Research in psychology, sociology, and anthropology has consistently demonstrated that these actuarial and economic definitions are too narrow to capture people's understanding of risk. Lay perceptions are multidimensional, encompassing higher order factors such as *dread* and *equitable exposure* (Bhatia, 2019; Slovic, 1987). *Dread risks*, as opposed to *chronic risks*, are defined by a perceived lack of control and potential large-scale loss of life, making flying a greater perceived risk than driving (e.g., Gaissmaier & Gigerenzer, 2012). Greater dread, in turn, is associated with greater perceived risk and a greater desire for regulation to reduce the risk (Slovic, 1987; Slovic et al., 1985; Sunstein, 2005). All these meanings and others are part of the public discourse and are included in the text corpora that we analyze. In other words, our focus is not on one definition at the expense of another, but rather endorses the rich and inclusive semantic history of *risk* in the natural language.

2. Guiding Research Questions

Our goal in this study was to track change in the public discourse on risk over historical time by addressing four guiding questions. First, we examined how the frequency of the word *risk* has changed over historical time. Word frequency analysis has been used to capture patterns of usage associated with changes in cultural importance (Greenfield, 2013; Twenge et al., 2012; Uz, 2014). Here, it allows us to evaluate the idea that the construct of risk is playing an ever-increasing role in the public discourse, as suggested by the sociological concept of a "risk society" (Beck, 1992) and the anthropological observation that the scope of the word *risk* has broadened over time (Douglas, 1992, p. 14). Second, we investigated how the sentiments for the words co-occurring with *risk* have changed. This sentiment analysis allows us to evaluate the hypothesis that risk is becoming a more negative construct, associated with expectations that societies and policy makers should invest ever more in risk reduction and prevention (the precautionary principle; Sunstein, 2005). Third, we asked how the meaning of *risk* has changed by examining change in the semantic relationship between it and other words. The meaning of a word can be reliably inferred from the contexts in which it has been used (Firth, 1957). For example, analysis of the linguistic context of the verb *broadcast* shows that 150 years ago it referred to the spreading of seed, whereas it is now used to mean the spreading of information (Li et al., 2019). We examined the text corpora for indications that *risk* is more subject to semantic change than are close semantic associates such as *danger* and *hazard*. According to social

anthropologist Douglas (1992), the concept of *risk* has a strong cultural foundation, but this foundation is not static: Perspectives and social environments change; some dangers are politicized as risks while other worries are backgrounded. If *risk* has become a crucial construct for singling out certain objective dangers and designating them as social concerns, then tracking the use of the term in public discourse can reveal an underlying dynamic mechanism that is constantly responding to the changing sociocultural environment (Douglas, 1992). Fourth, we decomposed the construct of *risk* into the specific topics with which it has been associated and tracked those topics over historical time. Our purpose here was to identify the most prominent risk topics over time and to consider how they have changed in relation to world events.

We investigated these questions by analyzing latent semantic patterns in natural language. Tracing the historical meanings of words requires a corpus of texts published over a sufficiently long time period. The Google Books Ngram Corpus (Lin et al., 2012) is one of the few corpora that meet this requirement. Drawing on over 100 sources (e.g., libraries and publishers), it contains over 8 million books published from 1600 to 2008, or 6% of all books ever published. The corpus thus offers a *telescopic view* over a large time period. The corpus has been used to detect large-scale changes in language, which in turn correlate with social and demographic changes (Hills & Adelman, 2015; Hills et al., 2012; Hills et al., 2015; Michel et al., 2011). Any corpus, however, has its limitations. The Google Books Ngram Corpus offers limited contextual information due to a narrow window size (5-grams, or a contiguous sequence of five words); moreover, there has been a surge in the proportion of academic articles in the corpus (Pechenick et al., 2015). We therefore also examined *The New York Times Annotated Corpus* (NYT corpus; Sandhaus, 2008) to lend convergent validity to our results. This corpus contains all (1.8 million) articles published in the *New York Times* from 1987 to 2007, and offers a more *microscopic view* on the risks of modern life as reported in the most widely read U.S. newspaper. Let us emphasize that because our analysis draws on English texts only, the present results are limited to English-speaking cultures. Nevertheless, the Google Books Ngram Corpus, in particular, has the advantage of covering a relatively long time period, going beyond short-term analyses of, for instance, media coverage of risk and mortality (see the references in Young et al., 2008).

3. Materials and Methods

We used word co-occurrence to construct semantic representations of risk in each year of the analysis, such that the meaning of *risk* was approximated by the context in which it was used. The co-occurrence information allowed us to quantify how the sentiment and semantics of *risk* have changed over history. As *risk* may be used in multiple contexts, we used Latent Dirichlet Allocation (LDA, Blei et al., 2003) to identify the historical risk topics. This topic model algorithm detects underlying topics that best explain the structure of the language around *risk*, and allowed us to identify risk topics as they changed over time. In what follows, we describe this procedure in more detail. We begin by

briefly describing the Corpora we used. We provide our data on https://osf.io/jctn8/?view_only=988fccae28ca4995b6a3796002a888cc.

3.1 Google Books Ngram Corpus

The Google Books Ngram Corpus consists of *n*-grams: contiguous sequences of *n* items from a given text (*n* ranges from 1–5). We used the 5-grams of all English words in our analysis; each data entry therefore displays the number of times a 5-gram appears in the corpus during a specific year. We retrieved all 5-grams starting or ending with the word *risk*. As is standard procedure in many natural language processing tasks, we removed stop words, punctuation, digits, and words containing fewer than three characters before using the WordNet-based NLTK lemmatizer (Bird et al., 2009) to lemmatize each noun to its singular form and each verb to its present tense. Next, we aggregated the corpus by year so that each document contained all 5-grams in a specific year. Aggregating topics by years encourages the topic model to identify the underlying patterns that best explain differences among risk structures over years.

3.2 The *New York Times* Annotated Corpus

The NYT Corpus contains all articles published in the *New York Times* from 1987 to 2007. We constructed a risk corpus by selecting articles that mentioned the word *risk* or *risks* more than twice. Next, we pre-processed the corpus in the same way as we did the Google Books Ngram data, apart from aggregating articles by year: Each news article was treated as one document.

3.3 Corpus of Historical American English

The Corpus of Historical American English (COHA) is a large structured corpus of historical English. It contains 400 million words of text produced from the 1810s to 2000. COHA is balanced by genre decade by decade, which brings both benefits and concerns. On one hand, it alleviates concerns that insights gained from the corpus are driven by the changing compositions of genres. On the other hand, it may fail to map the reality that public preferences for genres change over history. Although it is difficult to argue whether COHA is a better corpus for analyzing culture change than the Google Books Ngram corpus or vice versa, consistency in the findings from both corpora would lend convergent validity to the results. Therefore, we used COHA to validate some of the historical analysis conducted with the Google Books Ngram Corpus, namely, the analysis of frequency and semantic shift.

3.4 Analysis of Frequency and Contextual Sentiment

Analyses of frequency, contextual sentiment, and semantic drift (Figures 1 and 2) were conducted using the Macroscope (Li et al., 2019), an interactive linguistic tool that analyzes historical sentiment and semantic change. The Macroscope was built on the basis of the historical word co-occurrence data made publicly available through the Google Books Ngram Corpus. Frequency was calculated by dividing the count of the selected words by the corpus size to control for the different corpus sizes for each year. Contextual sentiment for the selected words was computed in terms of the averaged valence ratings of

co-occurring words during a given year. The valence ratings were retrieved from data collected by Warriner et al. (2013), which contain valence scores for 13,915 English words, each rated on its "pleasantness" by around 30 participants. Using contemporary norms to estimate the valence of words used decades ago is potentially problematic, as all words may have changed their meaning or sentiment over history. In practice, however, it has been shown that historical sentiment as inferred from averaging contemporary valence norms of semantic neighbors is similar to the sentiment judged by historical language experts (Buechel, Hellrich, & Hahn, 2016).

3.5 Semantic Shift Analysis

The purpose of our semantic drift analysis was to examine how and to what extent the meaning of *risk* has changed over the past two centuries in relation to related concepts such as *danger*, *fear*, and *hazard*. It consisted of the following three steps: First, we retrieved the historical word embeddings for 50,000 common English Words trained by Li et al. (2019). Word embeddings provide a vector representation for each word based on its co-occurring relationship with other words; that is, they represent the context in which a word has been used. To derive word embeddings, Li et al. (2019) first, from Google Ngram Corpus, constructed a co-occurrence matrix for 50,000 common English words that records the number of times any two words were used within the same 5-gram. Next, they computed the positive pointwise mutual information (PPMI) for each pair of words and then constructed a PPMI matrix with entries given by:

$$PPMI(v_i, v_j) = \max(0, \log(\frac{p(v_i, v_j)}{p(v_i) \times (v_j)})), \qquad (1)$$

where v_i, v_j represents a pair of words from the corpus and $p(v)$ corresponds to the empirical probabilities of those words co-occurring within a sliding window of 5 over the original text. Finally, Li et al. reduced the dimension of word embeddings to 300 using singular value decomposition (SVD). This dimensionality reduction acts as a form of regularization and allowed us to compare word similarities by computing the cosine similarity of word embeddings.

Second, drawing on the historical word embeddings trained by Li et al. (2019), we identified the 8-nearest semantic neighbors for the words *risk*, *danger*, *fear*, and *hazard.* Specifically, we retrieved word embeddings for each of the four target words and their semantic neighbors in the years 1800 and 2000. For *risk*, we also retrieved the historical embeddings every 20 years between 1800 and 2000. In order to compare word embeddings from different time periods, we must ensure that the vectors are aligned to the same coordinate axes. We therefore used Orthogonal Procrustes to align the historical embeddings (Schönemann, 1966).

Third, we visualized semantic shift of words in two-dimensional space. To this end, we used principal component analysis to reduce the dimensions of word embeddings from 300 to 2. Figure 2 plots the word embeddings retrieved in the second step according to the two orthogonal principal components (PC1 and PC2). These two principal components represent compressed dimensions that best explain the variance of the raw data and are therefore not directly interpretable except in relation

to relative distance between word embeddings. The background words (semantic neighbors) are always shown in their "modern" (year 2000) positions. This approximation is necessary since, in reality, all words are moving. *Risk* and its synonyms are shown in their modern and historical positions. The path travelled through the semantic space is a proxy for change in historical meaning.

Finally, to validate our observations, we quantified semantic change in *risk* and its related concepts using historical word embeddings trained on COHA (Hamilton et al., 2016) and on the Google Books Ngram Corpus (Li et al., 2019). For each word, we computed cosine similarity between embeddings trained on the 1820[1] corpus and on the 2000 corpus.

3.6 Topic Modelling

We studied historical change in the meaning of the word *risk* by extracting risk topics from the Google Books Ngram Corpus (Lin et al., 2012) and the NYT corpus (Sandhaus, 2008). The topic model we used was Latent Dirichlet Allocation (LDA; Blei et al., 2003), a bag-of-words algorithm that identifies a set of topics that best describe/re-generate the corpus. We took two main steps in analyzing the data. First, we identified the structure of risk meanings by applying the topic model to the risk corpus. This step allowed us to understand the key events associated with risk. Next, we applied trend analysis to understand how the risk topics identified in the first step changed over time.

3.7 Interpreting Topics

To make sense of the meanings of the risk topics, we used Equation (2) to identify the words most relevant to each topic. The relevance of term *w* to topic *k* given a weight parameter λ was defined as:

$$\gamma(w, k|\lambda) = \lambda \log\left(P(w|k) + (1-\lambda)\log\left(\frac{P(w|k)}{P(w)}\right)\right), \qquad (2)$$

where $P(w|k)$ is the probability of term *w* being assigned to topic *k* and $P(w)$ is the marginal probability of term *w* being in the corpus. The first component of the equation, $P(w|k)$, prioritizes terms with high frequency in a topic. However, it does not consider how unique term *w* is to topic *k*, which can be captured by $\frac{P(w|k)}{P(w)}$, a quantity that Taddy (2012) called *lift*. We set λ to 0.5 to take both components into consideration; λ determines the weight given to the probability of term *w* under topic *k* relative to its lift.

One issue with topic models is that it is not clear which topics capture structures specific to the risk corpus and which topics capture general features of the source corpus. To find out, we used Equation (3) to compute the specificity of topic *k* to the risk corpus:

$$\text{Specificity(k)} = \sum_{i=1}^{n}\left(\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)} * \frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}\right), \qquad (3)$$

[1] We chose 1820 instead of 1800 because the frequency of *risk* in COHA in 1810 proved too small to train a stable model.

where $\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)}$ is the normalized relevance of word *w* to topic *k*, and $\frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}$ is the ratio of the frequency of word *w* in the risk corpus to its frequency in the source corpus. Specificity can range from 0 to almost infinity. A specificity of 1 means that, on average, the words characterizing the topic have the same frequency in both the risk corpus and the source corpus, suggesting that the topic reflects the underlying pattern of the source corpus, not risk. An example of a nonspecific topic is one that generates the words necessary to construct every document, such as articles and pronouns. The absolute value of topic specificity is heavily influenced by the data format: NYT articles are more likely than 5-grams to contain non-risk-specific words (noise) and therefore have smaller values of $\frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}$. Topic specificity is not comparable across corpora; instead, it should be used to compare topics from a same corpus.

3.8 Tracking Trends in Topics

To analyze trends in topics over time, we used the output from the LDA model on the Google Books Ngram Corpus to calculate the contribution of each topic *k* in each year by applying Equation (4). For each document (i.e., all 5-grams in a specific year), the equation controls for document length by dividing the number of words generated by each topic by the total number of words in the document. Thus, the yearly topic contribution estimate, $p_d(k)$, is defined as:

$$p_d(k) = \frac{|\{w \in d: topic(w) = \mathrm{k}\}|}{|d\,|}, \qquad (4)$$

where *k* is a topic and *w* is a word in a document d. The numerator is the number of words in document *d* that are generated by topic *k*; the denominator is the total number of words in document d.

4. Results

4.1 How Has the Frequency of *Risk* Changed Over Time?

We first investigated change in the frequency of the word *risk* over time, starting with the Google Books Ngram Corpus. As Figure 1A shows, use of the word *risk* has increased dramatically since about 1970, with an approximately fourfold increase in usage since the 1950s. We checked this trend in English against other languages and found similar increases in French, German, Italian, and Spanish (Figure 1B). In addition, we observed a similar proliferation of *risk* in the Corpus of Historical American English (COHA; Davies 2012). As COHA is balanced by genre and subgenre across decades,[2] these findings suggest that *risk* proliferation is not an artifact of increasing numbers of scientific journals being included in the Google Books Ngram Corpus (Figure 1A). There is, however, no sign that the public discourse has turned darker in general, as close semantic relatives signifying undesirable states such as *danger*, *fear*, and *hazard* are not being used more frequently. On the contrary,

[2] For example, fiction accounts for 48–55% of the total in each decade (1810s–2000s); subgenres such as prose, poetry, and drama are likewise balanced. This balance across genres and subgenres means that researchers can be reasonably certain that patterns in the data do not merely reflect artefacts of a changing genre balance.

the use of *danger* and *fear* has declined steadily over the past two centuries, while the use of *hazard* has remained relatively stable at a low frequency. These results are consistent with the idea that *risk*, more than other terms, has become a central concept in recent public and political discourse (Beck, 1992; Bourke, 2005; Douglas, 1992).

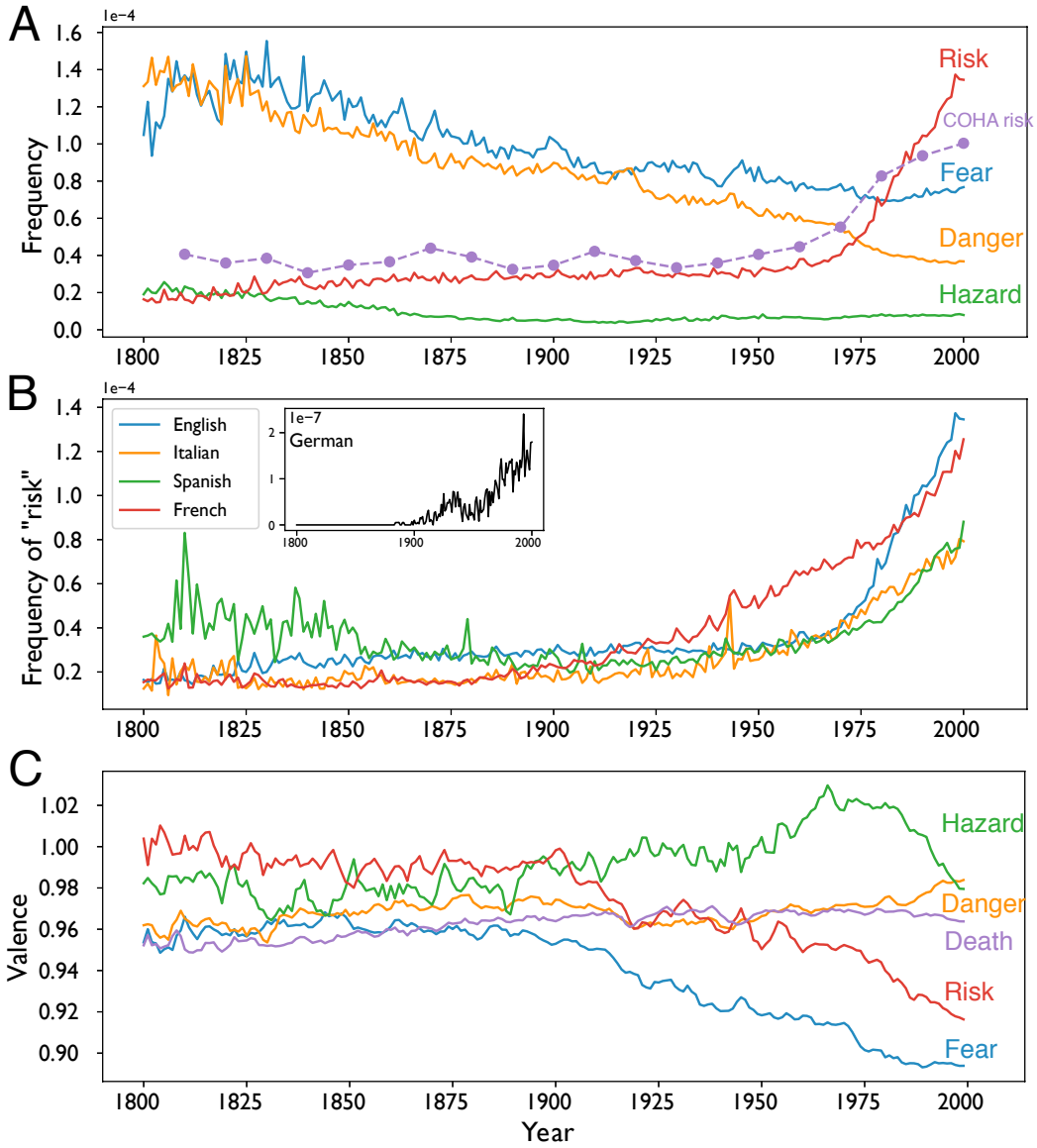

Figure 1. Historical change in the frequency and sentiment for the word *risk* and its close semantic neighbors in the Google Books Ngram Corpus. (A) Frequency of *risk*, *fear*, *danger*, and *hazard* in the Google Books Ngram Corpus and frequency of *risk* in the Corpus of Historical American English (COHA). (B) Frequency of *risk* in five languages—English, Italian, Spanish, French, and German—in the Google Books Ngram Corpus. German is presented in a separate box because the frequency of *risk* is much lower in German than in the other languages. (C). Change in the sentiment for words co-occurring with *risk, fear, danger, hazard*, and *death*. Sentiment was adjusted to mean score of all words, such that valences > 1 indicate a more positive context than average. The word *death* is included to provide a sentiment benchmark, as its meaning and sentiment have remained stable over history.

4.2 How Have the Sentiments Associated with *Risk* Changed?

Next, we examined whether the sentiments[3] associated with *risk* have changed over time. For example, is it possible—in line with a more economic interpretation of risk—that the use of the word *risk* is increasingly associated with an appreciation of the large potential rewards that make some risks worth taking (Pleskac & Hertwig, 2014)? This is not the case, as the results presented in Figure 1C show. Computing the frequency-weighted average valence of the words that co-occurred with *risk* over the past 200 years revealed that the sentiment associated with risk has become increasingly negative, showing a roughly monotonic decline from 1800 to 2000. To provide points of comparison, we also analyzed the related concepts of *danger*, *fear*, *hazard* as well as *death* as a benchmark. The sentiment analysis shows that *risk* has undergone a much larger change over time than these inherently undesirable concepts (with the exception of *fear*). In the early 1800s, the sentiment for words co-occurring with *risk* was more positive than that of any of the four comparison words; by the end of 20th century, it was more negative than that of *danger*, *hazard*, or *death* (Figure 1C). In other words, the word *risk* has become not only more prevalent but also more negative in meaning.

4.3 How Have the Semantic Relationships of Risk Changed?

The increasing negativity of *risk*'s sentiment, relative to the stability of the sentiment for *danger*, *fear*, and *hazard*, might be driven by the changing contexts in which these words have been used. In this section, we therefore turn to an analysis of semantic drift, which likewise suggests that the *risk* has experienced more semantic change over historical time than have its close semantic associates. Specifically, Figure 2 visualizes the semantic associates of *risk*, *danger*, *fear*, and *hazard* in two-dimensional space relative to their k most similar words in 1800 and 2000 (k = 9 for each word). A larger distance between two words suggests less similarity in the contexts in which they appeared. The pattern is clear: *risk*, *danger*, and *hazard* started as close semantic neighbors in 1800 and moved apart over time. By the year 2000, the underlying semantics of *risk* had grown more similar to those of *prevalence* and *prevention*, terms associated with the quantification, reduction, and avoidance of risk. *Danger* and *hazard*, in contrast, remained in the semantic area defined by words such as *harm, threat*, *adverse*, and *peril*. This finding suggests that the word *risk* has moved from merely representing the presence of threats to also being associated with the scientific examination, quantification, and prevention of threats.

It is possible that this pattern is a result of an increase in the number of academic (especially medical) articles in the Google Books Ngram Corpus (Pechenick et al., 2015). Therefore, we again used COHA, a smaller yet genre-balanced corpus, to validate our findings. We analyzed the semantic shift

[3] Because we inferred historical sentiment by averaging the valence of contextual neighbors, what we measured is sentiment of the context associated with *risk*, not directly sentiment of the word *risk*. However, the two are conceptually related: Because the meaning of a word can be learnt from the linguistic companions it keeps (Firth, 1957), words used in negative contexts are likely to carry negative connotations.

of *risk* using historical word embeddings trained on COHA (Hamilton et al., 2016) and compared the results with results derived from embeddings trained on the Google Books Ngram Corpus (Li et al., 2019). For each word, we quantified semantic similarity over history by computing the cosine similarity of embeddings trained on the 1820 corpus and the 2000 corpus. Cosine similarity scores range from 0 to 1, with larger scores indicating greater semantic similarity. Comparison of results from the two corpora confirmed that the semantics of *risk* was much less stable than the semantics of *danger*, *fear*, and *hazard* (Table 1). In addition, we searched for the nearest semantic neighbors for risk in COHA in 1820 and 2000. Again, we found that *risk* acquired associations with medical concepts over time: its top-5 nearest semantic neighbors changed from *loss*, *expense*, *danger*, *trouble*, and *run* in 1820 to *disease*, *diabetes*, *cancer*, *rate*, and *factors* in 2000.

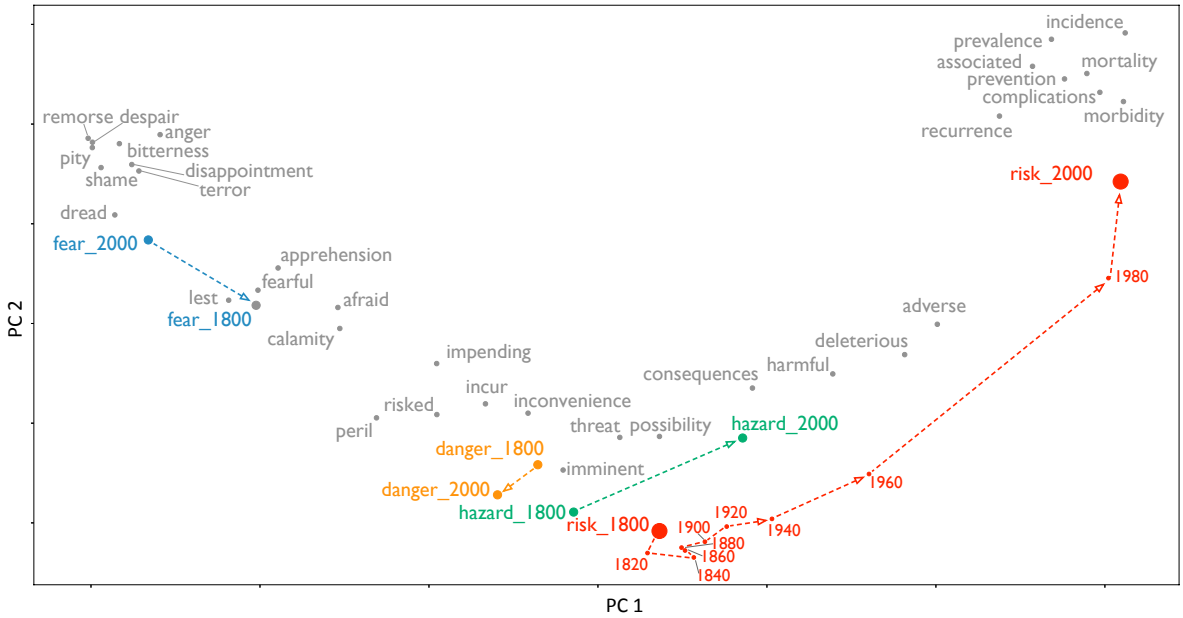

Figure 2. Semantic drift of *risk*, *danger*, *fear*, and *hazard* from 1800 to 2000 in the Google Books Ngram Corpus. The target words (in color) are shown in relation to their near associates[4] (in gray) in the years 1800 and 2000. The meaning of *Risk* is shown at 11 historical points from 1800 to 2000 with a 20-year interval. PCA was performed to reduce the dimension of word embeddings from 300 to 2 so that words can be visualized in two-dimensional space. The axes represent the two principal components. A larger distance between two words indicates lower semantic similarity. The words *risk*, *danger*, and *hazard* started as near neighbors in 1800 but moved apart over time.

[4] The semantic neighbors of risk identified from Google Ngram Corpus (as shown in the top right corner of this figure) are different from the words people report in the Small World of Word project (SWOW; De Deyne, Navarro, Perfors, Brysbaert & Storms, 2019) where they were required to "enter the first word that comes to mind" when reading the word *risk* (the most popular responses in SWOW are *danger, game, reward, gamble, chance, money*, to name a few). We argue that language corpora may have greater ecological validity because it reflects real situations where people find *risk* most appropriate to label an object. In contrast, behaviors in a free association task are likely to be influenced by non-semantic cues such as word frequency, phonological overlap, etc. However, we acknowledge that the need to be informative in language production may cause language corpora to underrepresent certain information, such as obvious fact like banana is yellow.

Table 1

Semantic Similarity Between 1820 and 2000

	Risk	Danger	Fear	Hazard
Google Books Ngram	0.36	0.61	0.58	0.56
COHA	0.42	0.81	0.80	0.54

Note. For each word, semantic similarity was quantified by calculating the cosine similarity of word embeddings between 1820 and 2000. The embeddings were normalized such that the similarity scores range from 0 to 1, with 1 and 0 representing maximum and minimum similarity, respectively.

4.4 How Have Risk Topics Changed Over Time?

The semantic drift analysis shows how *risk* has diverged from its semantic neighbors over the last two centuries, but it cannot provide detailed insights into the topical dimensionality of risk in this period. As noted by Blais and Weber (2006), risk is a multidimensional concept encompassing numerous topics. We therefore applied LDA to investigate the topics that have driven the proliferation of *risk* in the public discourse and its increasingly negative sentiment. We inferred topic meanings by inspecting their most relevant words (see Equation 2 in the Methods section), as summarized for each topic in Table 2. Applying the topic model to the Google Books Ngram Corpus identified six risk categories: **war** (topic 1, 2, 3), **nuclear** (topic 4), **health** (topic 5, 6, 7, 8, 9), **HIV/AIDS** (topic 10, 11), **risk society** (topic 12), **economy** (topic 13, 14), and a non-specific topic on risk analysis (topic 15).

Each topic represents a probability distribution over all words. In order to validate our interpretation of risk topics from the Google Books Ngram Corpus, we selected a collection of words (see the left column of Figure 3A) that characterize each of the risk categories identified above and examined how those words were distributed over topics (see the left panel of Figure 3A). Instead of selecting words from Table 2, we chose a different set of associates that we felt exemplified our interpretation of the topics, based on events occurring at the time the topics peaked. For example, under the war category, we selected words that reflect the major participants in 20th century wars (e.g., *Soviet, American, Japan, Germany*) as well as war-related words (e.g., *battle, invasion, army*). For the cancer category, we included names of the most common cancers. If our interpretation was correct, topics grouped under the same category should be more likely to generate corresponding words but not others. This is indeed what we found. For example, Figure 3A shows that topics 1, 2 and 3 in the Google Books Ngram corpus (identified as *war* topics in Table 2) were associated with the set of words we selected under the *war* category. This pattern, visualized as probability loadings on the diagonal of the word-topic probability heat map in Figure 3A, lends further support to our interpretation of topic meanings presented in Table 2.

Table 2

Most Relevant Words for Each Risk Topic, Ordered by Relevance as Defined in Equation 2

Index	**Google Books Ngram Corpus**	Index	**NYT Corpus**
1	Life, imminent, battle, resolve	1	Military, war, Iraq, troop
2	Life, war, bureau, loss	2	China, Japan, country, foreign
3	War, uncertainty, loss, prepare		
4	Nuclear, carcinogenic, patient, infant	3	Environmental, plant, energy, gas
5	Heart, coronary, injury, bear	4	Cancer, woman, study, breast
6	Breast, cancer, osteoporosis, fetus	5	Drug, patient, doctor, hospital
7	Stroke, cancer, disease, capital		
8	Prostate, cancer, event, Alzheimer		
9	Management, diabetes, cardiovascular, overweight		
10	AIDS, nation, HIV, immunodeficiency	6	AIDS, virus, infect, vaccine
11	HIV, deficit, assess, volume		
12	Management, value, assessment, society	7	Child, school, parent, student
13	Confrontation, return, equilibrium, preference	8	Fund, stock, investor, market
14	Rate, free, interest, return		
15	Behavio[u]r, group, death, population		
		9	Food, fat, eat, diet
		10	Insurance, bank, loan, insurer
		11	Law, court, abortion, tobacco
		12	Airline, flight, shuttle, space
		13	Company, business, executive, industry
		14	Investigation, Enron, prison, police
		15	Think, people, way, thing
		16	Republican, Clinton, Bush, Democrat
		17	Game, player, sport, team
		18	Day, car, hour, walk
		19	City, build, York, new
		20	Film, art, movie, theater

Note: Topics 15–20 of the NYT corpus are shown in gray to indicate that these topics are not specific to articles that contain the word *risk* and its inflections. Topic specificity is defined in equation 3.

How replicable is this category structure? To find out, we also analyzed the NYT Corpus. Applying the same procedure to the NYT Corpus confirmed all risk categories inferred for the Google Books Ngram Corpus (visualized as probability loadings on the diagonal of the right panel of Figure 3A). We can therefore conclude that the meanings of risk derived in our analysis of the Google Books Ngram dataset are not corpus-specific results associated with a non-representative sample, but reflect general trends in the topicality of risk over both relatively long and short time scales.

In order to ensure that the topics were risk-specific and did not just reflect the background features of the corpus, we next computed *topic specificity* (see Equation 3 in the Methods section) to quantify the relative correspondence of each topic with the risk corpus as compared with the entire

corpus (see Figure 3B). A topic specificity score around or below 1 means that the topic has a distribution of words similar to that seen in the entire corpus; the topic therefore represents the general features of the entire corpus. For the Google Books Ngram Corpus, we found the topic specificity of all risk topics to be above 1 (ranging from 50 to 650), suggesting that all topics were risk-relevant. In contrast, the specificity of NYT topics ranged from 0.7 to 2.5, with six topics being irrelevant to risk (the specificity scores of topics 15–20 were close to or less than 1). This notable difference in the topic specificity of the two corpora may be attributable to differences in data format: Recall that the Google Books Ngram data contain words that co-occurred with *risk* within a narrow window size, whereas the NYT data contain entire articles that mention the word *risk*. As such, NYT articles are more likely than Google Books Ngrams to contain words not specific to *risk*.

Nevertheless, both corpora rendered a similar set of high-specificity topics: nuclear, heart disease, cancer, diabetes, and HIV/AIDS. War-related topics had low specificity in the NYT Corpus. This result is not surprising because, as we show in the following analysis, war topics have gradually disassociated from *risk* since World War II, and the NYT Corpus only dates back to 1987. Beyond the risk topics identified for the Google Books Ngrams, we found only one additional topic in the NYT Corpus with specificity clearly above 1 (topic 9, featuring words such as *food, fat, eat,* and *diet*), and four additional NYT topics slightly above 1 (topics 11–14, which we interpreted as legal, flight, commercial, and fraud, respectively). Correspondingly, the key words associated with topics 11–14 showed low co-occurrence with *risk* in the Google Books Ngram Corpus throughout history. This comparison suggests that, overall, both corpora converged on a similar set of important risk categories.

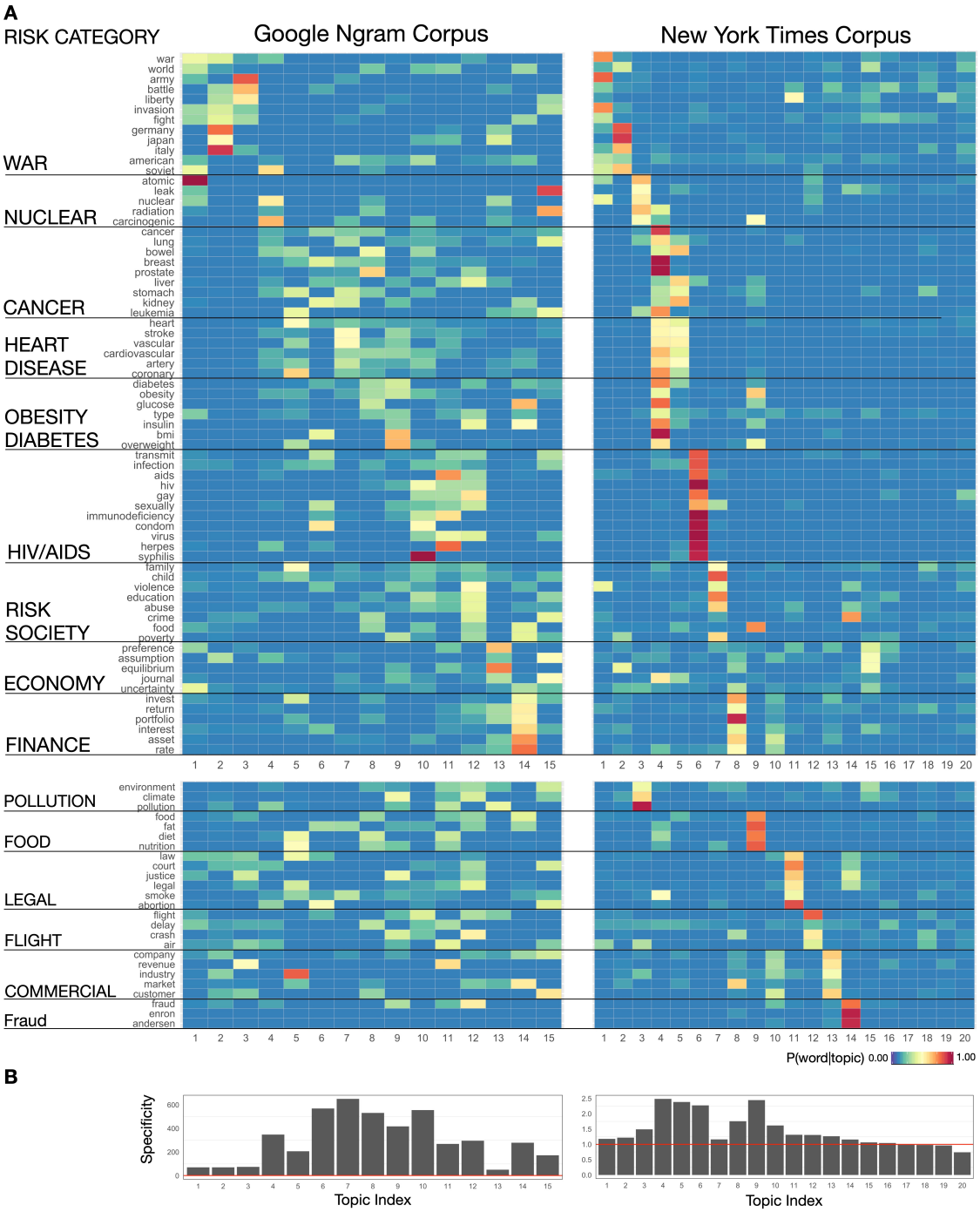

Figure 3. Visual quantification of risk topics. (A) Heatmap of the probability that word *w* was generated by topic *k* in models derived from the Google Books Ngram Corpus (left) and the NYT Corpus (right). Words on the y-axis were selected by referring to the list of most relevant words for each topic (relevance defined by Equation 2) and they were grouped by categories. (B) Topic specificity (as defined by Equation 3). The red horizontal line indicates topic specificity equal to 1. Topics with specificity above this reference line can be considered risk-specific and therefore capture one or more aspects of the meaning of risk. Topics with specificity below 1 can be considered generic words that are not informative with respect to risk meanings.

4.5 How Are Changes in Risk Categories Associated with Other Events and Developments?

One advantage of Google Books Ngram Corpus is that it allows us to investigate change in the sources of risk across a period of over 150 years and to speculate on how those changes relate to other historical events and developments. Specifically, we performed a trend analysis on the topic model derived from the Google Books Ngram Corpus over the years 1850 to 2008. As Figure 4 shows, the structure of the Google Books Ngram risk topics underwent major changes over this period. The three war-related topics emerge early in the distribution: Topic 1 (*life, imminent, battle, resolve*) dominated the risk structure in the second half of the 19th century, which witnessed several major wars (e.g., Crimean War, American Civil War). Topic 2 (*life, war, bureau, loss*) emerged and reached its peak during World Wars I and II. Topic 3 (*war, uncertainty, loss, prepare*) reached its peak during the Vietnam War. Topic 4 (*nuclear, carcinogenic, patient, infant*) peaked around 1985, capturing the risks associated with the proliferation of nuclear weapons during the Cold War (see the histogram in Figure 4) and the growing use of nuclear power in the 1970s and 1980s.

Chronic diseases such as heart disease and cancer are now the leading global risks for mortality (World Health Organization, 2009). Topics reflecting this development (topics 5–9) started to emerge from the 1970s and remain the most prominent risk topics. Due to the large proportion of shared words associated with the different health conditions, topics 5, 6, 7, and 8 show considerable overlap, that is, they share words that describe cancer, heart and coronary issues, and other severe diseases. Topic 9, associated with obesity and diabetes, emerged after 2000. The data for topics 10 and 11 show that concerns over AIDS and HIV emerged within 2 years of the first AIDS diagnosis in the US in 1981 and soon reached a peak around 1995, when the reported annual mortality from HIV/AIDS peaked in the United States (Centers for Disease Control and Prevention, 2010). Potentially reflecting the dramatic medical advances in treatments for HIV and the ensuing drop in mortality rates, this risk topic decreased in prominence after 2000 (see the histogram of AIDS-related deaths in the US in Figure 4).

Finally, topic 12 (*management, value, assessment, society*) is about management of various social risks. It seems to relate to Beck's conceptualization of the *risk society*, being associated with words such as *Ulrich, Beck,* and *modernity*. Topics 13 and 14 relate to the economy, and emerged from the 1970s: topic 13 features words like *preference*, *assumption, equilibrium*, and *journal*, whereas topic 14 features words such as *return*, *portfolio,* and *interest*. Lastly, topic 15 (*behavior, group, death, population*) seems to be concerned with general risk analysis, without reference to any specific risk event.

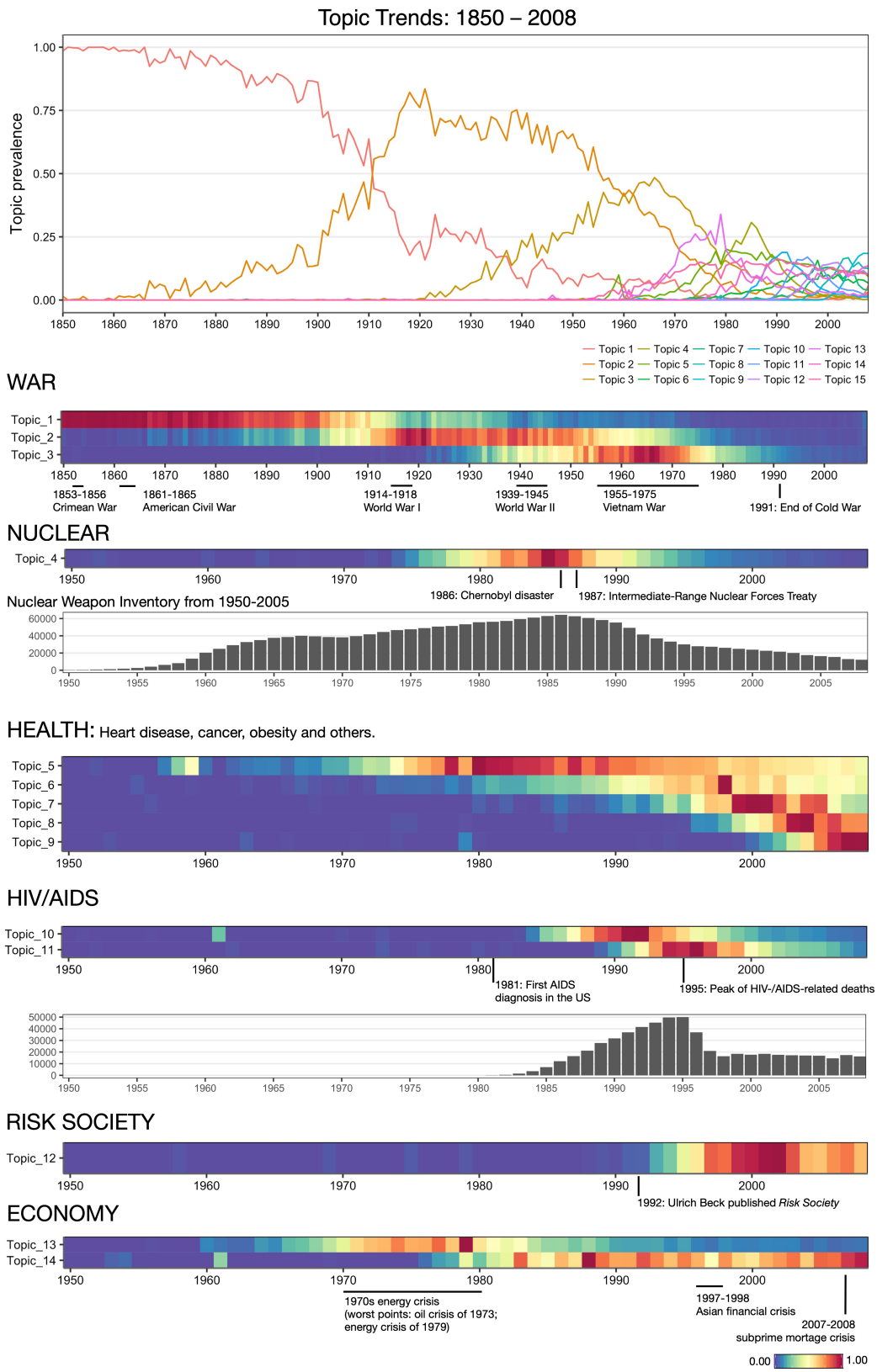

Figure 4. Trend analysis on risk topics derived from the Google Books Ngram Corpus. Topics are grouped into six categories: war, nuclear, health, HIV/AIDS, risk society, and economy. Relevant historical events are labeled to indicate how changes in the meanings of risk were associated with historical events and developments. Top panel: historical trends of 15 risk topics (computed using Equation 4). Bottom panel: normalized topic trend for each individual topic. Topic 15 is not included as it does not refer to a specific risk topic.

5. Discussion

Risk is a complex multidimensional construct. It takes a variety of forms in public discourse and has, accordingly, been investigated in various ways. Each approach focuses on some aspects of the discourse at the expense of others. One common approach has been to analyze media coverage of risk as a leading source of information for the general public and experts alike (see, e.g., Combs & Slovic, 1979, and various references in Young et al., 2008). Our approach consisted in a large-scale analysis of historical text corpora. Such corpora are attractive because they collate a vast array of perspectives on an extensive historical time window: in the case of the Google Book Ngrams Corpus, over 8 million books and 150 years. What did we learn about the risk-related discourse in English-speaking countries?

First, we found—consistent with Beck's (1992) diagnosis of post-industrialist Western societies as risk societies facing a wide variety of unique and human-made risks and with Gidden's (1990) idea that society is increasingly preoccupied with the future and its safety—that the word *risk* has become much more prevalent (Figure 1A). There is evidence of an approximately fourfold increase in its usage since the 1950s. Beck also stressed that risks in the post-modern world are increasingly unknowable and unpredictable due to scientific and technological innovations having unanticipated consequences. It is possible that this process has contributed to our second major observation, namely, that the sentiments associated with risk have become much more negative, starting around 1900 and confirming Pinker's (2011) observation that humans have become increasingly preoccupied with the negative aspects of risk. Interestingly, the same does not apply to its close semantic relatives (Figure 1C). What is also puzzling is that this change in sentiments is happening at a time when the semantics of risk have become increasingly associated with notions of quantification, reduction, and prevention—findings that also challenge the idea that the increase in negative sentiments has been caused by the unknowability of risks. In addition, we found that the risk categories to some extent reflect real-world changes in the prevalence and magnitude of the respective risks (see Figure 4 and our analyses of nuclear proliferation and AIDS-related deaths). Finally, we also found a shift from macro-risks, such as war and battle, to more individual-specific, chronic risks such as disease (Holzmann & Jørgenson, 2000) as well as shift toward more variability in risk topics. The strong focus on modern diseases suggests that the public discourse is generally oriented toward the most prevalent causes of death and harm. This is noteworthy, as several authors have argued that people tend to be afraid of the wrong things (see Glassner, 2018; Renn, 2014; Schröder, 2018).

Many of these patterns observed are remarkable in part because they are monotonic: the notable increase in the frequency and negativity of the risk construct, and the increase in number of topics it encompasses. These changes are perhaps related to one another. One potential underlying mechanism is the social amplification of risk (Jagiello & Hills, 2018; Kasperson et al., 1988; Moussaïd et al., 2015): as information is transferred from one individual to another, people tend to share the more negative aspects of a risk at the expense of potential gains. In Jagiello and Hills (2018), an individual exposed to

a balanced argument on nuclear power shared that information with another individual. As information was communicated from one individual to the next, the focus shifted increasingly to the downsides of nuclear power and away from its benefits. This pattern is consistent with the substantial evidence that negative information has more influence on decision making than positive information (Baumeister et al., 2001; Ito et al., 1998; Rozin & Royzman, 2001). A second, related factor is that this effect may be further amplified by increasing communication over the period of our analysis. As Herbert Simon (1971) noted, "a wealth of information creates a poverty of attention" (pp. 40–41). With the unprecedented amounts of information now available, all other things being equal, the absolute amount of negative information has increased. In this environment, information that is better at being received, remembered, and reproduced has a selective advantage (Hills, 2019). This mechanism may apply particularly to information on prominent risks, which may self-reinforce more rapidly via intensified social communication (Jagiello & Hills, 2018).

What can be concluded from our results about the state of the public discourse on risk? First and foremost, our analysis can offer only a glimpse of this complex and multi-dimensional construct. Yet, we found results that were both disconcerting and reassuring. Primarily, the increasing prevalence of the word *risk* is an indicator of its growing significance, which is in itself a double-edged sword. Classifying something as a potential risk is likely to burden it with negative sentiments. Yet, branding something a risk also appears to imply the chance of changing our fortune in relation to it. Importantly, the text corpus analyses suggest that risk categories track real threats over the 20th and 21st century, shifting from violent death to chronic disease and major risks for morbidity and mortality in the modern day. In this sense, the risk discourse reflects changes in threats as well as changes in the potential to mitigate them.

Reference

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Beck, U. (1992). *Risk society: Towards a new modernity.* Sage.

Bhatia, S. (2019) Predicting risk perception: New insights from data science. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2018.3121

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit.* O'Reilly Media.

Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1*(1), 33–47.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bourke, J. (2005). *Fear: A cultural history.* Virago.

Buechel, S., Hellrich, J., & Hahn, U. (2016). Feelings from the past: Adapting affective lexicons for historical emotion analysis. In E. Hinrichs, M. Hinrichs, & T. Trippel (Eds.). *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 54–61). The COLING 2016 Organizing Committee.

Centers for Disease Control and Prevention. (CDC). (2010). *HIV/AIDS surveillance report*. CDC.

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly*, *56*(4), 837–849.

Davies, M., Hegedűs, I., & Fodor, A. (2012, November). The 400 million word corpus of historical American English (1810–2009). *In English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010* (Vol. 325, p. 231). John Benjamins Publishing.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. Behavior research methods, 51(3), 987-1006.

Douglas, M. (1992). *Risk and blame: Essays in cultural theory.* Routledge.

Firth, J. R. (1957). *Papers in linguistics 1934–1951.* Oxford University Press.

Gaissmaier, W., & Gigerenzer, G. (2012). 9/11, Act II: A fine-grained analysis of regional variations in traffic fatalities in the aftermath of the terrorist attacks. *Psychological Science, 23*(12), 1449–1454.

Giddens, A. (2013). *The consequences of modernity.* John Wiley & Sons.

Glassner, B. (2018). *The culture of fear: Why Americans are afraid of the wrong things.* Basic Books.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science, 24*(9):1722–1731.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change.* arXiv. https://arxiv.org/abs/1605.09096

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow.* Random House.

Hills, T. T. (2019). The dark side of information proliferation. *Perspectives on Psychological Science*, *14*(3), 323–330.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition, 143*, 87–92.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2):431–440.

Hills, T. T., Proto, E., Sgroi, D., & Seresinhe, C. I. (2019). Historical analysis of national subjective wellbeing using millions of digitized books. Nature human behaviour, 1-5.

Holzmann, R., & Jørgenson, S. (2000). Social risk management: A new conceptual framework for social protection, and beyond. *International Tax and Public. Finance, 8*(4), 529–556.

Hornik, K., & Grün, B. (2011). Topic models: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30. [not in text]

Ito, T. A., Larsen, J. T., Smith, N. K., Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75*(4), 887–900.

Jagiello, R. D., & Hills, T. T. (2018). Bad news has wings: Dread risk mediates social amplification in risk communication. *Risk Analysis, 38*(10), 2193–2207.

Kasperson, R. E., Renn, O., Slovic, P., Brown, H., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis, 8*(2),177–187.

Li, Y., Engelthaler, T., Siew, C. S., & Hills, T. T. (2019). The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods*, *51,* 1864–1877. https://doi.org/10.3758/s13428-018-1177-6

Lin, Y., Michel, J. B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. *Proceedings of the ACL 2012 System Demonstrations* (pp. 169–174). Association for Computational Linguistics.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1):77–91.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014):176–182.

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Science, 112*(18), 5631–5636.

Oeppen, J., & Vaupel, J. W. (2002). Broken limits to life expectancy. *Science, 296*(5570):1029–1031.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, *10*(10):e0137041.

Pinker, S. (2011). *The better angels of nature: The decline of violence in history and its causes.* Penguin.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology, 143*(5), 2000–2019.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica, 32*(1–2), 122–136.

Rayner, S., & Cantor, R. (1987) How fair is safe enough? The cultural approach to societal technology choice. *Risk Analysis,* 7(1):3–9.

Renn, O. (2014). *Das Risikoparadox: Warum wir uns vor dem Falschen fürchten* [The risk paradox: Why we are afraid of the wrong things]. Fischer Taschenbuchverlag.

Rothman, L. (2016, January 6). *Why Americans are more afraid than they used to be*. TIME. https://time.com/4158007/american-fear-history/

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.

Sandhaus, E. (2008). *The New York Times Annotated Corpus.* Linguistic Data Consortium.

Short, J. F., Jr. (1984). The social fabric at risk: Toward the social transformation of risk analysis. *American Sociological Review, 49*(6):711–725.

Schönemann, P. H. (1966). A generalized solution of the Orthogonal Procrustes problem. *Psychometrika*, *31*(1), 1–10.

Schröder, M. (2018). *Warum es uns noch nie so gut ging und wir trotzdem ständig von Krisen reden* [Why we've never had it so good and we still constantly talk about crises]. Benevento.

Simon, H. (1971). Designing organisations for an information-rich word. In Greenberger, M. (Ed.). *Computers, communications, and the public interest* (pp. 40–41). The Johns Hopkins Press.

Slovic, P. (1987). Perception of risk. *Science, 236*(4799):280–285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1985) Characterizing perceived risk. In Kates, R. W., Hohenemser, C., & Kasperson, J. X. (Eds.). *Perilous progress: Managing the hazards of technology* (pp. 91–125). Westview.

Stearns, P. N. (2012). *American fear: The causes and consequences of high anxiety*. Routledge.

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle.* Cambridge University Press.

Taddy, M. (2012). On estimation and selection for topic models. In T. Sichelman & K. Atkinson, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (pp. 1184–1193). Association for Computing Machinery.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in individualistic words and phrases in American books, 1960–2008. *PLOS ONE, 7*(7):e40181.

Uz, I. (2014). Individualism and first person pronoun use in written texts across languages. *Journal of Cross-Cultural Psychology, 45*(10):1671–1678.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4):1191–1207.

World Health Organization. (2009). *Global health risks: Mortality and burden of disease attributable to selected major risks*. World Health Organization.

Young, M. E., Norman, G. R., & Humphreys, K. R. (2008). Medicine in the popular press: The influence of the media on perceptions of disease. *PLoS One, 3*(10), e3552.