Article

Sociological Methods & Research
1-34

DOI: 10.1177/0049124117729701
journals.sagepub.com/home/smr
SAGE

Enlisting Supervised Machine Learning in Mapping Scientific Uncertainty Expressed in Food Risk Analysis

Akos Rona-Tas[1], Antoine Cornuéjols[2], Sandrine Blanchemanche[3], Antonin Duroy[4] and Christine Martin[2]

Abstract

Recently, both sociology of science and policy research have shown increased interest in scientific uncertainty. To contribute to these debates and create an empirical measure of scientific uncertainty, we inductively devised two systems of classification or ontologies to describe scientific uncertainty in a large corpus of food safety risk assessments with the help of machine learning (ML). We ask three questions: (1) Can we use ML to assist with coding complex documents such as food safety risk assessments on a difficult topic like scientific uncertainty? (2) Can we assess using ML the quality of the ontologies we devised? (3) And, finally, does the quality of our ontologies depend on social factors? We found that ML can do surprisingly well in its simplest form identifying complex meanings, and it does not benefit

[1] Department of Sociology, University of California, San Diego, La Jolla, CA, USA
[2] Département Modélisation Mathématique, AgroParisTech—INRA, Informatique et Physique, Paris, France
[3] Met@risk, INRA, Paris, France
[4] FlameFy, Paris, France

Corresponding Author:
Akos Rona-Tas, Department of Sociology, University of California, San Diego, 488 Social Sciences Building, 9500 Gilman Drive, La Jolla, CA 92093, USA.
Email: aronatas@ucsd.edu

from adding certain types of complexity to the analysis. Our ML experiments show that in one ontology which is a simple typology, against expectations, semantic opposites attract each other and support the taxonomic structure of the other. And finally, we found some evidence that institutional factors do influence how well our taxonomy of uncertainty performs, but its ability to capture meaning does not vary greatly across the time, institutional context, and cultures we investigated.

Keywords

scientific uncertainty, content analysis, machine learning, ontology, food safety

With the growing availability of digitized, unstructured textual information, social scientists are looking for new ways to analyze content. There are two problems that need to be solved simultaneously: how to create a new system of classification and, once we have such a system, how to sort text correctly into those categories (Bail 2014; DiMaggio 2015; Wagner-Pacific, Mohr, and Breiger 2015). The first problem is the problem of induction, the second is the problem of sorting. In this article, we are going to present a project classifying scientific texts on food safety, where our focus is scientific uncertainty expressed by the expert authors of those documents. Our ultimate substantive goal is to contribute to an empirical foundation for the study of uncertainty and ignorance in science. In this article, we offer a general methodology of coding large corpora of text through the cooperation of humans and machines. We show in a nontechnical way that machine learning (ML) can be a robust and useful aid in building and using classification schemes, we call ontologies. Yet, we emphasize that method cannot be divorced from content. Using ML is no substitute for human understanding of the nature of the text corpus to be analyzed. The analysis must be guided by the nature of the question asked and the kind of text available. In our version, ML does not negate but enhances qualitative approaches.

Induction and Sorting

There are two approaches that developed in recent literature to sorting texts with the help of computers. Both, in different ways, combine human knowledge with machine algorithms. One is based on machines processing and

classifying text on the basis of certain algorithms, the other works through a more explicit cooperation of humans and machines (Grimmer and Stewart 2013; Jain, Duin, and Mao 2000). The first, called "unsupervised" learning, does not start with a set of categories or an ontology. It is an exploratory method that constructs the ontology in an inductive manner by generating groupings of similar texts (Aggarwal and Zhai 2012:77-128; Grimmer and King 2010). It uses the relative frequency of words and combination of words to create those clusters. Topic modeling extracts multiple themes from documents sorting them into multiple categories (Blei 2012). However, despite its autonomy in induction, the unsupervised method cannot bypass human input either. Humans must select and prepare the text for analysis, but more importantly, unsupervised methods rely very heavily on human input on the back end. The meaning of the clusters requires human interpretation and the interpretation needs validation. Do the results emerging from the ontology predict outcomes outside the world of the text corpus that it ought to? For instance, can we map changes in the relative frequencies of various categories generated by the unsupervised process onto a flow of events in a meaningful fashion (as in DiMaggio, Nag, and Blei 2013)? Prediction validates meaning, and unsupervised methods depend on smart tests to persuade us to accept a certain interpretation of the meaning of its findings especially if those are not trivial (Grimmer and Stewart 2013; Jelveh, Kogut, and Naidu 2015).

Supervised methods, our focus here, on the other hand are used to sort cases into categories created by humans (Hillard, Purpura, and Wilkinson 2007; Nardulli, Althaus, and Hayes 2015). In this methodology, after selecting the corpus, one begins with an ontology or coding scheme produced or adopted by the researcher, one that is backed up by human reasoning, and then the machine is trained to reproduce the choices the coders would make if they were to work correctly, consistently following the inherent logic of the system.

Supervised methods require heavier human input at the front end in the form of a coding scheme, a set of categories and then humans must hand-sort some text. Then, the task of the machine is simply to scale up the human effort and sort much more text that humans possibly can or help the human coders with suggestions speeding up the process. The machine at its best does as well as the best human coder. In supervised learning, the machine is the student, and humans are its teachers.

In our research on scientific uncertainty, we (humans) devised two ontologies or classification systems, hand-coded texts using these coding schemes, and then used supervised learning to teach machines to find texts that conform to those systems of classification. At the same time, it was not

just the machine that learned to mimic our ontology, but we used the machine's learning process to find out about our own coding scheme. In other words, we used ML to interrogate our ontology and assist us with checking our induction.

Unsupervised methods have proved to be impractical in our research because of the nature of the question we wanted to ask and the kind of documents we were working with. We were interested in uncertainty, which was only one, limited aspect of these documents. Moreover, risk assessments are not primarily about uncertainty. They are about the food risks, the scientific research, and evidence available about them, and ultimately, they need to establish the amount of the contaminant or the biohazard that can be ingested without falling ill and the ways of avoiding exposure above a certain level. Two documents on the same or similar hazards may seem very similar to any unsupervised algorithm, but they may be very different in terms of how much and what kind of uncertainty they express. Had these texts been only about uncertainty on a topic, or had each document included a special section devoted exclusively to uncertainties, a section that we could have easily detached and analyzed separately, unsupervised methods might have brought some interesting results.

Ontology, Coding Scheme, and Sorting

This article traverses various fields each with its own language and vocabulary. While we use terms like coding schemes from the sociological tradition of content analysis (Krippendorff 2013), we borrow the word ontology from computer science (Staab and Studer 2009). Ontology is "an explicit specification of a conceptualization" (Gruber 1993:199), and it is domain-specific.[1] In full-fledged ontologies, concepts and their relationships in a particular domain (e.g., scientific uncertainty) are organized in a system that is an abstract representation of a world with a certain purpose and at a specific level of granularity. Ontologies are powerful because they can clarify and—to some degree—automate various cognitive processes that manipulate meaning (Madsen and Thomsen 2004). Ontologies are also meant to be portable and accommodate a wide range of contexts and users.[2]

Ontologies as classification systems are used as coding schemes. We will use these terms interchangeably, but the word ontology forces us to think about our categories as interrelated and connected often in very intricate ways. Ontologies can depict complex relationships among their concepts; at other times, ontologies are not more than a few bins to sort things into, the

former often referred to as heavy and the latter as lightweight ontology (Gómez-Pérez, Fernández-López, and Corcho 2004:8). In this article, we will offer two ontologies: One is very simple even for a lightweight ontology, 5 categories with little structure which organize expressions of scientific uncertainty along five dimensions, and one that is larger and hierarchical and sorts uncertainty statements into 27 categories arranged in a tree structure. We call the first one a typology and the second a taxonomy.

The Problem of Scientific Uncertainty in Food Safety Science

Scientific uncertainty is both a theoretical problem and, as science has been employed in regulatory decisions, a critical policy issue. In sociology of science, sociological study of the relationship between knowledge and uncertainty goes back at least to the work of Robert Merton (1957:417) who wanted to understand the role of ignorance in the construction of scientific knowledge. Recently, a large literature emerged on ignorance (Frickel and Vincent 2007; Gross and McGoey 2015; Merton 1987; Smithson 1989), issues of uncertainty (Gillund et al. 2008; Winkler 1996), negative knowledge (Cetina 1999), nonknowledge (Böschen et al. 2006, 2010; Kaempner et al. 2011), nescience (Croissant 2014; Gaudet 2013; Gross 2007, 2012), constituting a small subfield called agnotology (Proctor and Schiebinger 2008). This literature focused on how ignorance is constructed as opposed to exist as "state of nature," simply as lack of knowledge, either is highly conceptual or builds on qualitative case studies, and our contribution here is to devise a measurement of uncertainty in one particular context.

Scientific uncertainty as a general concept is a critical issue in the realm of policymaking as well. From the effects of smoking and the necessity of immunization campaigns to the causes of global warming, the certainty of scientific knowledge has been the focus of vigorous debate for policymakers and other public and private stakeholders. In food safety science, concerns about the certainty of scientific knowledge became amplified in the 1990s, in the wake of the food safety crisis created by bovine spongiform encephalopathy (BSE), commonly known as "mad cow disease." This debacle involved British scientists, who for a while misrepresented the certainty of what then was known about the relationship between BSE infecting cattle and the Creutzfeldt–Jacob disease afflicting humans (Phillips, Bridgeman, and Ferguson-Smith 2000; Millstone and van Zwanenberg 2001; van Zwanenberg and Millstone 2005). As a result, various international agencies began to emphasize the importance of expressing scientific uncertainty in food safety

risk assessment documents. Some, like World Health Organization (WHO), the European Food Safety Authority (EFSA), the U.S. Environment Protection Agency (U.S. EPA), and the U.S. Office of Management and Budget (U.S. OMB), issued guidelines working toward a system that both identifies the type of uncertainty scientist perceive and the extent to which our knowledge is uncertain in a particular respect (EFSA 2006; U.S. EPA 1993, 2000; U.S. OMB 2006; WHO 2008). These agencies recognized that their decision makers have to understand the nature of the weakness in the evidence that the experts present and must have a clear sense of how much confidence they can place in various scientific findings in order to take the best decision. Scholars studying science itself also took up the issue of scientific uncertainty and formulated various normative frameworks to guide experts in future risk assessment reports. So far, none of these frameworks were adopted systematically by scientific panels.

Our approach to scientific uncertainty is not normative but empirical and comparative. We want to describe, measure, and understand how scientists express uncertainty in scientific reports assessing food risk. In our larger project, we look at English language risk assessment documents produced for food safety regulators in the United States and the European Union (EU) between 2000 and 2010. We investigate two main, distinct areas of food hazards in the food risk field: contaminants and biohazards.[3] As the two fields draw on different subdisciplines, they differ in the way they make use of various scientific methodologies (experiments, observational studies, statistical analyses, analytic modeling, etc.) and thus may have different understandings of scientific uncertainties.

The documents were coded independently by two human experts, reconciling disagreements by a third, and then we tested these ontologies using ML. Our main objective was to answer three questions: (1) Can we use ML to assist with coding complex documents such as food safety risk assessments on a difficult topic like scientific uncertainty? Is ML doing a reasonable job overall in coding sentences? (2) Can we assess using ML the quality of the ontologies we devised? If ML is doing a reasonable job coding sentences, can we test various logical and semantic properties of our ontologies? (3) And, finally, does the quality of our ontologies depend on social factors? Is performance of our ontologies related to external, social forces such as institutional learning, institutions, and culture?

In the rest of the article, we first describe the two ontologies and our data.[4] Then, we explain the use of ML and the choices we made to conduct our experiments. We pose the three sets of questions, state some simple propositions, and discuss the empirical results.

The Two Ontologies

To map scientific uncertainty, we developed two complementary ontologies in an inductive and iterative process. We set out to develop a conceptualization of uncertainty in scientific documents, to identify textual expressions of uncertainty, and then to sort and analyze documents according to the amount and type of uncertainty they voice.

We began with the general literature on scientific uncertainty. We tried to adapt these categories to the documents. In the process, categories disappeared, others got consolidated, and new ones emerged. In the end, we constructed two systems of classification or ontologies. The first, a simpler ontology, is designed to capture the nature of the judgment the scientists make about the uncertainty of their conclusion about the state of knowledge on a particular food hazard. This multidimensional typology classifies the experts' judgment of the evidence and focuses on their way of expressing their state of mind. The second, a hierarchical system, gauges the content of uncertainty they communicate. The categories identify the problems that give rise to uncertainty about our current state of knowledge as perceived by the authors. It is a taxonomy because the categories are arranged in a genealogical hierarchy, where "ancestry" can be seen as higher levels of generality.

For both ontologies, our context unit is the entire document (the coders read the entire report to fully understand the conclusion), and the coding or recording unit is the sentence or a few related, consecutive sentences forming a passage expressing a particular type of uncertainty (our data point). One sentence or passage can contain multiple expressions of uncertainty and can be sorted into multiple categories.[5] We also refer to categories as "variables" because their values vary from sentence to sentence.[6] Consequently, we talk about judgment variables (JVs) and uncertainty variables (UVs) when we talk about categories of the first and the second ontology.

The scientific documents were official reports ordered by the main federal food safety agencies in the EU (EFSA) and in the United States (U.S. FDA, USDA, and U.S. EPA). As we were interested in the final verdict of the experts, just as policymakers are, we coded only summaries and conclusions of each document to capture the uncertainties that the experts thought remained after they reviewed the available research on the topic. In a few documents, there was a special section devoted exclusively to uncertainty. There we coded that section as well.

Judgment Typology

Our first ontology was designed to capture various aspects of the judgment of the experts in their conclusions. It describes how the panel judges the weight of the evidence, and it follows more closely the language they use to do so. This ontology consists of five categories. They are conceptually distinct. Three of them express uncertainty (*hedging*, *precaution*, and *disagreement*) and two (*confidence* and *expert assumption*) communicate the opposite.

Hedging is a way of indicating that experts have doubt about or a lack of total commitment to a proposition they present. Hedging, a way of making things fuzzier (Lakoff 1972),[7] expresses a "lack of complete commitment to the truth value of an accompanying proposition" (Hyland 1996:1). It suggests that the speaker is not committed entirely to a proposition because he or she is uncertain about the truth of its content. The hedge signals this uncertainty without laying out its causes in detail there, in that sentence, albeit the causes may be explained elsewhere in the text.[8] Hedging ill serves risk managers because it makes the topic of interest less clear. To identify hedging, we ask the question: "Can the proposition be restated in such a way that it is not changed but that the author's commitment to it is greater than at present? If yes, then the proposition is hedged." (Crompton 1997:281). For instance, dropping "likely to be" in the sentence: "The . . . panel concluded that . . . the risk is likely to be conservative . . . " would make it more definite.

Our second category is *confidence*. Here we wanted to capture the opposite of uncertainty, an emphatic commitment to a proposition. Often referred to as boosters, expressions of certainty, assurance, and conviction expressions of confidence provide a crucial clue for risk managers (Myers 1989; Vázquez and Giner 2009) and play an important role in persuasion in risk assessments. They stress finality and absence of doubt. While there are many words that are commonly used as boosters (e.g., undoubtedly, clearly, well known, demonstrate, and proven), whether they express confidence in the relevant scientific knowledge can be judged only from the wider context. Experts, for instance, can be confident that no good data are available on a topic or report and that it was demonstrated that the statistical models cannot answer the crucial question. They are confident about their own assessment of the evidence but not the strength of the evidence to settle an issue. In such cases, there is uncertainty, and confidence is to emphasize that it is there.

Our third category is labeled *expert assumption*. This is another form of confidence. The expert is aware that studies or models make certain assumptions about the world. These assumptions are not directly supported by evidence, but according to the expert, this does not pose any problem. These

are the best assumptions an expert can make or, at least, these are not assumptions that the report questions.

We coded *precaution* as our fourth variable. Precaution is a way of dealing with uncertainty. Making conservative assumptions or building conclusions around "worst-case scenarios" is a way of creating certainty where data and models fail to provide it. There is a large literature on the precautionary principle in food safety and the presumed differences in precaution between the EU and the United States that developed mostly in the context of genetically modified organisms (Hammitt et al. 2005; Lynch and Vogel 2001).

Our final category is *disagreement*. Disagreement is a staple of science, but here we are interested in only disagreements that the report treats as unresolved. This happens either when experts on the panel find unanimously that contradicting evidence on the topic is equally strong or when the panel splits, and some members disagree with others and voice dissent.

Uncertainty Taxonomy: An Ontology Based on the Source of Uncertainty

To build our second ontology focusing on *content* of uncertainty, we began with the general literature on scientific uncertainty (Hattis and Burmaster 1994; Morgan and Henrion 1990; Pate-Cornell 1996; van Asslet and Rotmans 2002; van der Sluijs et al. 2005; Walker et al. 2003) and papers addressing uncertainties in the different disciplines involved in the food risk assessment process such as epidemiology, microbiology, toxicology, and exposure assessment (Dorne and Renwick 2005; Grandjean and Budtz-Jorgensen 2007; Kang, Kodell, and Chen 2000; Kroes et al. 2002; Nautta 2000). Beside this literature, we drew upon two main institutional documents: the opinion of the Scientific Committee of EFSA (2006) entitled uncertainties in dietary exposure assessment and the WHO draft guidance document on characterizing and communicating uncertainty in exposure assessment (WHO 2008). We simplified and adapted the basic structure of these classification systems through a series of test coding of European, U.S., and international food safety risk assessments arriving at a 28-item hierarchical ontology defined by a decision tree. As one moves down the tree, one gets to more specific content. The coder had to code at the most specific (lowest) level possible. Here we follow a basic insight by Merton that uncertainty in science involves the process of moving from unspecified to specified ignorance (Einstein and Infield 1936; Merton, 1957, 1987; see also Popper [1934] 1959).

The tree was a decision tool to aid our coders. Sentences coded at branches, rather than leafs or terminal nodes (at the right column in Figure 1 or the light-blue label in Figure 3), were unspecified at a lower level.

Coding sentences for content that can be quite complex raises the problem of context much more so than the categories of our first ontology. The meaning of sentences is often influenced by text that is not adjacent. Comprehending the source of a particular uncertainty often required following a long exposition in the body of the report that the coder read but did not code. In fact, while we annotated sentences, here it would be more accurate to say that we were classifying the entire document and flagged the sentences (or passages) that provided the best clue.

Our ontology begins with the common, philosophical distinction between epistemic and ontic uncertainty (also known as natural variability). Epistemic uncertainty is the kind that points to missing or incomplete information. Ontic uncertainty is the inherent, random variation among cases that no further research can reduce. Epistemic uncertainty then is divided into problems that relate to data and those that relate to the model that we use to understand data. Each, in turn, is subdivided into lower level, more specific categories.

Data problem can be that some specific data (factors/variables) are simply missing. This is, however, rarely where the report stops. In the absence of good data, it reaches for surrogate data that are not exactly what we want but with some inference are useful or data we want but measured imperfectly. Surrogate data can be inadequate because we have the wrong population, the wrong context, the wrong hazard, or an imperfect sample. Measurement can be faulty because the measurement was taken incorrectly, it was not properly reported, or reported in a way that creates problems of comparison.

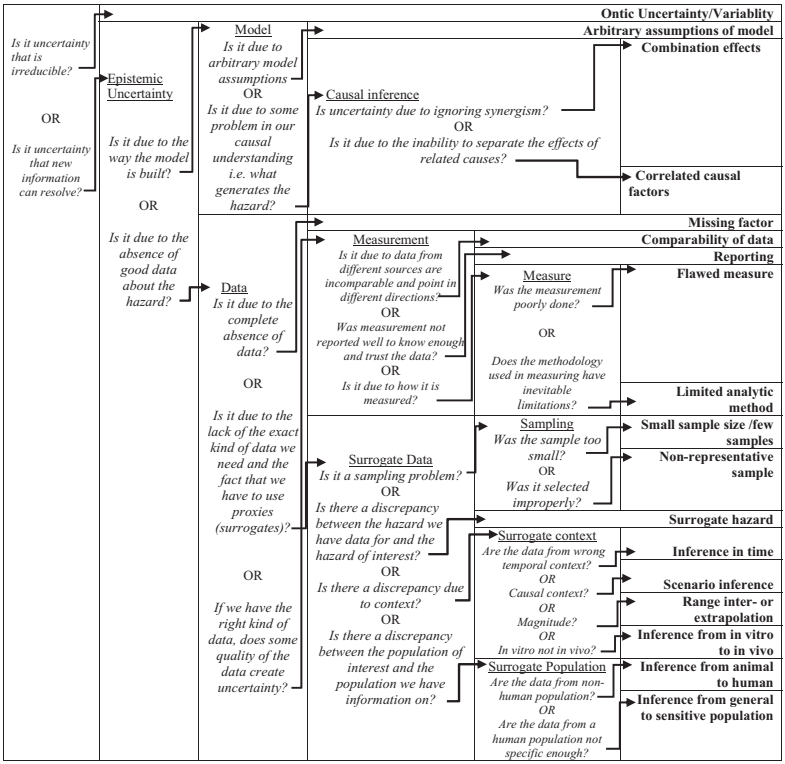
For models, we distinguish between causal and other, formalized models. Causal inference problems are further specified.

This ontology is built as a hierarchy from the most general down to the more specific, but another way of thinking of this tree structure is that the categories are organized in groups of similar content, whereby "children" of the same "parent" show more family resemblance than "children" of different "parents" or "grandparents."

Data (The Documents)

The corpora of text we coded were 115 official risk assessment documents produced by the EFSA in Europe, and the three U.S. federal agencies primarily responsible for food safety across the Atlantic, U.S. FDA, USDA, and

Figure 1. Decision Tree for Uncertainty Taxonomy Coding.



Note. If the answer is "yes" to the question, the arrow points to the next, more specific category.

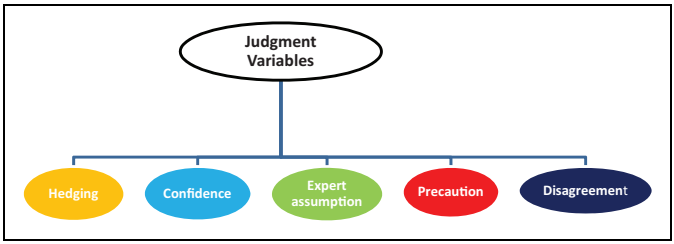


Figure 2. The structure of the ontology of uncertainty based on judgment of scientific experts.

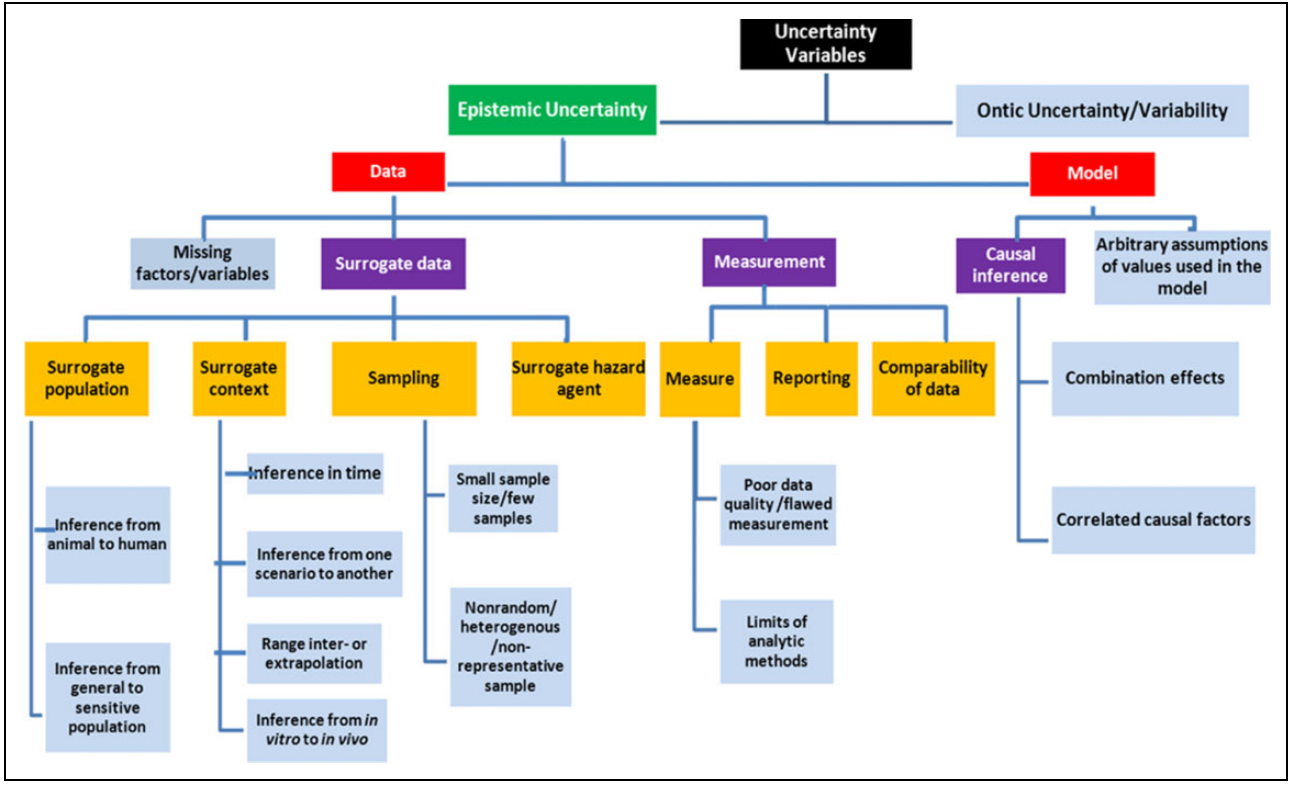


Figure 3. The structure of the ontology of uncertainty based on content.

U.S. EPA between 2000 and 2010. Because a risk assessment document often covers several hazards, each with its own scientific research and uncertainties, our record was the document of a specific hazard, and reports covering more than one hazard were coded as if they were separate documents effectively separating its content as if it were separate documents. Therefore, our actual unit is the hazard document.

The documents range from one to over 600 pages with average of around 70 pages. They are official risk assessment reports written by a panel of scientists.

Using ML

One straightforward strategy for the construction of an automated coding system for these documents is to use supervised learning techniques. Supervised learning aims at finding rules starting from a set of training instances. In our case, the training instances are the sentences (or small set of sentences) and their associated labels given by human coders. The goal is to extract rules that would allow a system to automatically label new sentences (or set of sentences) drawn from documents similar to the one used by the human coders. In our case, the system should be able to code a text input as "coded" versus "noncoded" and, if coded, as one of the categories present in the ontologies. Furthermore, if the learning system is trained from instances drawn from different contexts, differences in the learned rules could be enlightening about differences between these contexts. For instance, it could appear that the rules learned from American documents differ somewhat from the rules learned with documents from the EU because the documents are produced by a different institutional contexts. Or that the rules of reporting may change over time.

In the estimation, we used naive Bayesian classifier, support vector machines (SVMs), *k*-nearest neighbor, and decision tree algorithms (see Flach 2012). The best overall learners were naive Bayes and SVM.

We started with the bag-of-words approach. We broke the sentences up into words. We eliminated "stop words," that is, words that were too common to be helpful such as "the," "a," "is," and so on. Then, we stemmed the words thus erasing the difference between, for instance, learning, learned, learns, and so on. In our experiments, we removed 146 stop words and used the Porter stemmer (see Perkins 2010; Porter 1980), yielding a final vocabulary of size 2,530.

At this point, one can adopt one of several existing numeric coding strategies. For instance, in the bag-of-words approach, each word present in a

sentence can be associated with a 1 in the input vector and all the other, absent, words with a 0. Or one can count the number of times this word is present. For example, the sentence "In addition to the limitations al-ready listed, there are also limitations introduced by the methods used to analyze data inputs to the risk assessment" would become "addit limit alreadi list limit introduc method analyz data input risk assess" after stop words have been removed and after stemming. Then, it can be coded either with a 1 for the feature "limit" or with a 2 since this stem appears twice in this sentence. Other coding techniques involve the computation of relative frequencies (e.g., tfidf for "term frequency/inverse document frequency"). Here we report experiments with the 0/1 coding method.

Some algorithmic learners are, at the core, designed to separate two classes. If one wants then to learn multiclass classification, for instance, separating the three class "*missing variables*," "*surrogate context*," and "*sampling*," some scheme must be devised to turn two-class classification rules into multiclass ones. In our experiments, we used the all-versus-all technique (see Aly 2005).

It is a paradox that supervised learning is there to let us code very large amounts of text but ML itself works well only when a large amount of text for training and testing is already human coded. One problem when learning to separate data from different categories is that their frequency can be significantly dissimilar (e.g., in our data, one sentence falls under the "*measure*" category, while 71 fall under the "*missing factors/variable* category"). If, for instance, one category is 10 times more highly represented than another one, a simple majority rule will yield a 90 percent successful prediction rate without any learning taking place. In order to circumvent this opportunistic but uninformative behavior, one method is to balance class sizes. When the hand-coded data are plentiful, it is sufficient to sample the overrepresented classes in order to reduce their size to that of the least represented one. In our experiments, however, data are already rather scarce as human coding is costly (having to read long and difficult texts to affix the right code) and another technique had to be used. For each underrepresented class, we chose to generate artificial data points (sentences) by mixing the characteristics of existing data points from this class. Specifically, we randomly drew two actual data points (say, two different sentences coded to the same class or UV) and made up an artificial data point (an artificial sentence or set of words) by retaining for each feature either the one encountered in both actuals if they agreed (common words) or by randomly choosing a value of one actual if they disagreed on this feature.

The data then were split into two halves. One half was the training set, where the algorithm calculated the best fitting model, then it was tested on the other half, the test set, to see how well it can reproduce human coding. In order to measure the prediction performance, we used a fivefold cross-validation technique (see Japkowicz and Shah 2011).

In this article, we will use recall, precision, and overall predictive power to describe our results. Recall is the percentage of the observations (sentences) in category *k* predicted as *k* by ML. These are the correct predictions as percentage of cases actually in that category. Precision is the percentage of cases predicted as being in *k* actually being in *k*. These are the correct predictions as a percentage of predicting that category. The overall predictive power is the correct predictions as a percentage of all the cases. Later, we will also introduce a simple index to measure pairwise confusion.

Discussion

In this section, we will pose three questions, translate them to empirical propositions, and discuss our results.

Can We Use ML to Assist With Coding Complex Documents on a Difficult Topic Like Scientific Uncertainty?

Is ML doing a reasonable job overall in coding sentences? What features of the text do we need to consider to maximize predictions? How are false positives (sentences that are incorrectly put in a category) and false negatives (sentences that are incorrectly left out of a category) distributed?

If ML is able to recognize and classify sentences with little error, it can be useful to aid human coders. Coding risk assessment documents is, in certain ways, an easier task than coding other types of texts such as blogs, novels, or electronic mail. Scientific texts use a more standardized vocabulary than most other documents, and they put a premium on clarity and explicit expression. Risk assessments often follow a common format: introduction, hazard identification, dose–response assessment, exposure assessment and risk characterization, and conclusion, and there is also often a summary in front. Scientists learn how to write risk assessments, what each element should include, and so on.

Yet, coding expressions of uncertainty involves finding a set of meanings that are quite complex. A single sentence or a set of contiguous sentences may not carry the entire meaning, but the uncertainty is signaled by reference to earlier parts of the text. Context can modify the meanings of entire statements.

Proposition 1.1: The more the method can incorporate complexities, the better the tool is going to be.

The task involves complex meanings. Adding hypernyms, considerations for context should improve our predictions over using just "bag of words."

Cross-checking. Before we evaluated the performance of ML, we looked at the errors ML made and went back to each case to see whether it was the algorithm that erred or the human coder. One of the ways ML can help with coding is by calling attention to human mistakes. Thus, ML learns from human coders, and human coders can learn from ML. This is especially important when the coding scheme is not a simple classification task where a few preestablished categories have to be applied but a complex system that is inductively developed and implemented in the same process.

There were 178 mismatched sentences where ML disagreed with human coders. We reinspected each case individually, and we agreed with the human coders in 87 percent and with ML in 6 percent of the cases. For another 5 percent of the cases, both should have chosen a different UV, and for rest, the sentences did not express uncertainty (essentially, human error). There were also three cases where the same sentence had more than one code. In two of those cases, the machine correctly attached one of the codes. We corrected the human mistakes, and the final results were obtained on the corrected set.

Results

Is ML doing a reasonable job overall? We looked at this in two steps. The first step was to find whether a sentence was an expression of uncertainty or judgment of ANY kind. Here we care more about recall than precision, as it is better for the coders to get false positives (wrong suggestions for coding), which the coders can simply override by looking at a smaller set of ML suggestions, than false negatives (no suggestion where there should be one), which force coders to scrutinize the entire document if they are to correct them. The second step was to find the sentence's code in an ontology, given that it was an uncertainty expression.

Step 1. We began with the simplest approach, bag of words, using just the stems of words in each sentence to predict whether the sentence is coded or not. For JVs, SVM estimation gave the best results. Eighty-four percent of the sentences were correctly identified as being coded

Table 1. The Success of Finding Sentences That Are Coded With the Judgment Ontology.

Support Vector Machine		Predicted			
		Coded as Judgment	Noncoded	Recall	No. of Elements
Actual	Coded as judgment	**378**	91	*80.60%*	469
	Not coded	70	**474**	*87.13%*	544
	Precision	*84.38%*	*83.89%*		
				Total	1,013
	Overall	**84.11%**			

Table 2. The Success of Finding Sentences That Are Coded With the Uncertainty Ontology.-

Support Vector Machine		Predicted			
		Coded as Uncertainty	Noncoded	Recall	No. of Elements
Actual	Coded as Uncertainty	**447**	108	*80.54%*	555
	Non Coded	109	**413**	*79.12%*	522
	Precision	*80.40%*	*79.27%*		
				Total	1,077
	Overall	**79.85%**			

for one of the five categories (Table 1). Recall was 80.6 percent, ML failed to recognize one-fifth of the coded sentences.

For the UVs, ML was able to correctly identify 79.85 percent of the sentences (Table 2). Here too, a fifth of the coded sentences went unrecognized.

While ML is far from perfect, the results using the most primitive method are surprisingly good. To further improve predictions, we tried to add more complexity.

One idea was to recognize that what matters for assigning a sentence to a category is not so much the words in the sentence but the meaning of the word which can be expressed by synonyms that should be treated as the same word, even though they "look" different. We tried to use WordNet (available at http://wordnet.princeton.edu/), a large lexical database that groups nouns,

verbs, adjectives, and adverbs into sets of cognitive synonyms (synsets). Using hypernyms, more general words covering a set of more specific words (like using the word "fish" for "tuna," "sardine," or "swordfish"), we could not improve our prediction. In fact, the proportion of correctly predicted UVs fell by about 15 percent.

We also tried word sets assuming that the joint presence of certain words may make a difference, and we modeled context by looking at the position of a sentence in the conclusion. Neither of these improved our predictions. So far, we have not seen any improvement by adding complexity, but we are not ready to reject our first proposition. We will try other methods.

Step 2. How well was the ontology reproduced by ML once we knew the sentence expressed uncertainty? For JVs, the best overall accuracy was 84.1 percent (Table 3). For UVs, it was 78.82 percent (Table 4).[9]

Can We Assess the Quality of the Ontology We Devised Using ML?

Can we say something about the quality and properties of our ontologies using ML? If ML is doing a reasonable job coding sentences, can we test various logical and semantic properties of our ontologies? Where does ML work better and where does it work worse inside the ontologies?

Ontologies should be applicable in a consistent and reproducible manner. Algorithms take consistency and reproducibility to a mechanical extreme. Algorithms can spot human inconsistency. This inconsistency can be some systematic weakness in the ontology that guides coders. By looking at the patterns of errors of classification, what is called "confusion matrices," we can learn about the weaknesses of our ontologies and understand the cognitive process behind coding. [10]

JVs. Our ontology of the JVs is simple. As we did not elaborate the logical connections, we have only semantic relationships among the five variables. Confusion, therefore, will be driven by the compatibility of the connotations of the variables.

Proposition 2.1: Categories that have opposite connotations will be less likely to be confused.

Therefore, we would expect that *hedging* will be less likely to be confused with *confidence* than with *disagreement*. Categories that have similar connotations will be more likely to be confused. Therefore, on the one hand, *hedging* with *disagreement*, and on the other, *confidence* with *expert*

Table 3. Confusion Matrix for Judgment Variables.

		Predicted					
		Confidence	Disagreement	Expert Assumption	Hedged Language	Precaution	Recall
Actual	Confidence	**92**	0	1	11	1	*0.88*
	Disagreement	0	**6**	0	0	0	*1.00*
	Expert assumption	3	0	**89**	10	7	*0.82*
	Hedged language	26	0	12	**273**	20	*0.82*
	Precaution	0	0	1	4	**48**	*0.91*
					Total	604	
	Overall	**84.10%**					

Table 4. Confusion Matrix for Uncertainty Variables.

	Arbitrary Assumptions of Values Used in the Model	Causal Inference	Combination Effects	Comparability of Data	Correlated Causal Factors	Data	Epistemic Uncertainty	Inference from Animal to Human	Inference from General to Sensitive Population	Inference from in vitro to in vivo	Inference from one Scenario to an Another Scenario	Inference in Time	Limits of Analytical Methods	Measurement	Missing Factors/Variables	Nonrandom/Heterogenous/ Nonrepresentative Sample	Ontic Uncertainty/Variability	Poor Data Quality/ Flawed MEASUREMENT	Range Inter-or Extrapolation	Reporting	Small Sample Size/Few Samples	Surrogate Context	Surrogate Data	Surrogate Hazard Agent	Surrogate Population	Prediction
Arbitrary assumptions of values used in the model	**39**	1	0	1	0	1	1	1	0	0	3	0	1	0	2	0	1	0	1	1	0	0	0	0	0	*73.58*
Causal inference	0	**17**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Combination effects	0	0	**19**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Comparability of data	0	0	0	**25**	1	0	0	0	0	0	0	0	3	0	1	2	3	0	0	0	0	0	0	0	0	*71.43*
Correlated causal factors	0	0	0	0	**15**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Data	0	0	0	1	0	**17**	2	0	0	0	1	0	1	0	3	0	0	0	0	0	0	0	0	0	0	*68.00*
Epistemic uncertainty	0	1	0	0	0	0	**25**	0	0	0	0	0	2	0	1	0	2	0	0	0	1	0	0	0	0	*78.13*
Inference from animal to human	0	0	0	0	0	0	0	**21**	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	*87.50*
Inference from general to sensitive population	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Inference from in vitro to in vivo	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Inference from one scenario to an another scenario	2	0	0	1	0	0	1	2	1	0	**26**	0	1	0	3	1	1	0	1	0	0	0	0	0	0	*65.00*
Inference in time	0	0	0	0	0	0	0	0	0	0	0	**19**	0	0	1	0	0	0	0	0	0	0	0	0	0	*95.00*
Limits of analytical methods	2	0	1	2	0	1	2	1	0	0	2	0	**37**	0	5	1	3	0	1	0	0	0	0	0	0	*63.79*
Measurement	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	0	0	*100.00*
Missing factors/variables	3	1	1	3	1	1	1	1	0	0	3	2	6	0	**42**	5	1	0	0	0	0	0	0	0	0	*59.15*
Nonrandom/heterogenous/nonrepresentative sample	0	0	0	2	0	0	0	0	0	0	0	0	2	0	2	**40**	1	0	0	0	0	0	0	0	0	*85.11*
Ontic uncertainty/variability	0	0	0	0	0	0	1	0	0	0	0	0	2	0	2	0	**25**	0	0	1	0	0	0	0	0	*80.65*
Poor data quality/flawed measurement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	*100.00*
Range inter- or extrapolation	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	**22**	0	0	0	0	0	0	*91.67*
Reporting	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**15**	0	0	0	0	0	*100.00*
Small sample size/few samples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	0	0	0	*100.00*
Surrogate context	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**	0	0	0	*100.00*
Surrogate data	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**	0	0	*100.00*
Surrogate hazard agent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	*100.00*
Surrogate population	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	*100.00*
Overall	**78.82**																									

Table 5. Confusion among Judgment Variables.

		Predicted				
		Confidence	Disagreement	Expert Assumption	Hedged Language	Precaution
Actual	Confidence		.000	.022	.092	.007
	Disagreement			.000	.000	.000
	Expert assumption				.057	.055
	Hedged language					.070
	Precaution					

assumption and *precaution* will be more likely to be confused. Some of these confusions will arise from the fact that the same sentence is often coded by more than one of the variables with similar connotations.

Results. We use pairwise confusion to measure the likelihood that category A and B are confused with one another. Confusion is a symmetric measure that is 0 when there is no confusion between a pair of categories and 1 when all cases are misclassified as belonging to the other category.[11]

Pairwise confusion $= (f_{ij} + f_{ji})/(f_{ij} + f_{ji} + f_{ii} + f_{jj})$, where f_{ij} is the number of cases in category i predicted as being in category j.

Table 5 reveals that our proposition is wrong. The most confused variables are *hedging* and *confidence*. This is clearly unexpected. It turns out that the source of the confusion is that we often find both in the same sentence. Hedging clears the sentence for a following confident statement or modulates confident pronouncements later.[12]

Hedging is also confused with *precaution* and *expert assumption* for the same reason: hedging balances certitude. (Disagreement is too rare to analyze.)

UVs. For a hierarchical ontology such as the one we created for uncertainty, we can compare errors in relationship with the distance among concepts in the ontology. We can define distance between two concepts as the number of steps it takes to reach from one concept to another following links in the hierarchy.

Proposition 2.2: In an ontology, confusions increase with closeness.

Variables that we find closer in the ontology (separated by fewer splits in the tree/decisions) are more likely to be confused. Semantically close

Table 6. The Pairs Most Confused.

Pair		Distance	Confusion
Missing factors/variables	Limits of analytical methods	4	.122
Missing factors/variables	Inference from one scenario to another	4	.081
Nonrandom/ nonrepresentative sample	Missing factors/variables	4	.079
Limits of analytical methods	Comparability of data	3	.075
Ontic uncertainty/variability	Limits of analytical methods	6	.075
Inference from one scenario to an another	Arbitrary assumptions of values used in the model	6	.071
Missing factors/variables	Data	1	.063
Limits of analytical methods	Epistemic uncertainty	4	.061
Missing factors/variables	Arbitrary assumptions of values used in the model	4	.058
Nonrandom/ nonrepresentative sample	Comparability of data	5	.058
Ontic uncertainty/variability	Comparability of data	5	.057
Ontic uncertainty/variability	Epistemic uncertainty	2	.057
Missing factors/variables	Comparability of data	3	.056

variables have more similar meanings and thus are easier to confuse. An ontology works better, if confusions increase with closeness, when big, more general analytic distinctions are more clear and reliable.

Results. As shown in Table 4, we found variation in recall among the UVs. The worst fit is *missing factor/variables* and *limits of analytic methods*, followed by s*cenario inference and arbitrary assumptions.* It is also interesting that one would expect some types of uncertainties to be more common and ritualized and thus better predicted. *Inference from animal to human* (interspecies variation) and *ontic uncertainty* are both common and often expressed in a common form.[13]

What can we say about which UVs are the easiest to confuse? Table 6 shows pairs with pairwise confusion larger than .05. Pairwise confusion rates are fairly low. Of the top 13 pairs, *missing factors/variables* is part of six. This seems to be the weakest part of our ontology. Because its root is *epistemic uncertainty*, defined as "uncertainty that can be reduced by additional information," it is easy to see why *missing factors/variables* is easy to

confuse with anything in this branch. This is also the most common type of uncertainty. It is followed by *limits of analytical methods*, the inadequacy of scientific methods to reach a strong conclusion, and *comparability of data*, both are one side of four pairs.

The highest level of confusion is between the pairs of missing factors/variables and limits of analytic methods, inference from one scenario to another and nonrandom/nonrepresentative sample.

To assess the correlation between distance and confusion, we first weighted our data by the numbers of observations and eliminated pairs where one of the variables had fewer than 5 observations or neither had more than 10.[14] The result shows that confusion is weakly but negatively correlated with distance: The closer the two variables are, the more likely they are to be confused.

We also ran a multiple regression of confusion on distance. We controlled for the number of sentences observations each variable in the pair had. We found the same, even stronger result: closeness increases confusion. Thus, we found some evidence for Proposition 2.2a.

Does the Value of Our Ontologies Depend on Social Factors?

Can ML say anything about the circumstances that influence the performance of our ontologies? Is the predictability of the categories a function of the larger social context? As documents are social constructs, the success of classification may not depend only on the powers of ML to find the best algorithm, nor just on the cognitive and logical properties of the ontology, but also on the social process that generated the documents.

We tested three such factors: time, institutions, and scientific cultures. Thus, we compared earlier and later documents to see how learning and increased awareness of the problem may had an effect. We also compared the EU and the U.S. documents to search for larger institutional differences. And, finally, we also contrasted documents discussing contaminants and biohazards to see whether different scientific fields may articulate uncertainty differently. We concentrated here only on the hierarchical uncertainty ontology.

Propositions 3.1: The more recent documents are better predicted overall by ML.

This is a learning theory. Because panels that write risk assessments today are more aware of the importance of explicitly expressing uncertainties than those that wrote them a decade ago, and because panels developed a more

standardized way of talking about uncertainty, we expect predictions improve over time.

Proposition 3.2: The EU documents will be better predicted overall by ML than those from the United States.

This is an institutional theory. Because in the EU risk assessment in food safety is more centralized than in the United States and the RAs are more standardized, European RAs will be easier to machine code. In the EU, the vast majority of food safety risk assessments are written by panels of one agency, EFSA, at the request of the European Commission, that then decides on what, if any, action to take. In the United States, there are three principal agencies in charge of this topic (U.S. FDA, USDA, and U.S. EPA), and each can choose to use its own staff and in-house experts to generate risk assessments or they can outsource it to outside experts or institutions. The relationship between food safety agencies and the food industry also varies across the Atlantic. The U.S. agencies see their role more as balancing between the interests of consumers and producers. In EFSA's role, this balancing is less explicit, and not answerable to any national industry of government, it sees itself more as a neutral scientific enterprise.

Proposition 3.3: There will be a pattern of prediction that will be different in the world of contaminants and biohazards.

This is a theory about different scientific cultures driven by their subject matters and histories. We do see that the relative frequency of various uncertainties differ across the two worlds. For example, contaminant research uses experiments more frequently (animal experiments on biohazards are generally less useful), while biohazard research relies more on epidemiological evidence. As a result, inference from *animal to human* is a bigger concern for contaminants, while *comparability of data* and *causal inference* are more common concern for biohazards. We also found that contaminant RAs tend to be more specific about uncertainty than biohazards that tend to report more *epistemic uncertainty* and *missing variables/factors*. Contaminant research is more likely to point to *ontic uncertainty* and synergistic *combination effect*.

Proposition 3.3a: Overall contaminants will be better predicted because contaminant RAs are more specific and explicit about uncertainties.

This is probably due to historical reasons. Contaminant research can draw on older science (especially since many of the biohazards are novel zoonoses

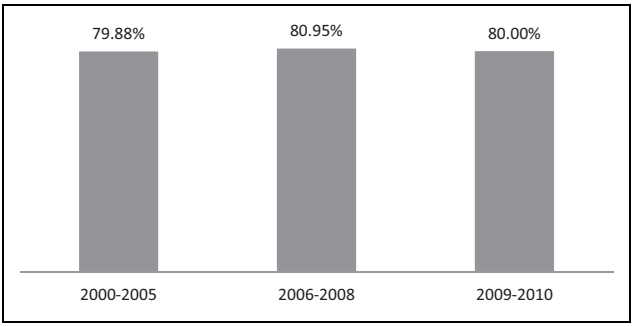


Figure 4. Predictive power of the hierarchical ontology over time.

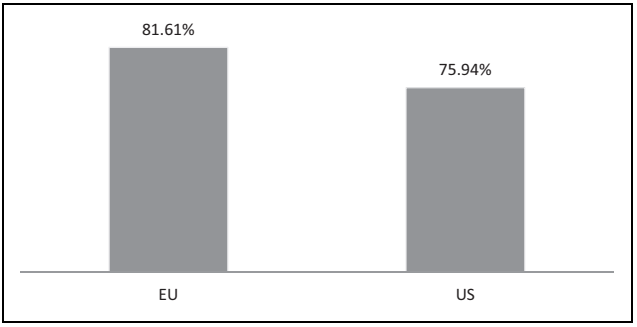


Figure 5. Predictive power of the hierarchical ontology in the European Union and the United States.

[hazards passed from animals to humans] such as BSE, avian flu, or new strains of older viruses or bacteria).

Proposition 3.3b: Each field will be better predicted on issues that are more central to that field because frequent concerns are expressed in a more ritualized form.

Results. The overall fit for earlier RAs is not different than for later RAs. As is seen in Figure 4, there is no learning.[15] The three periods are virtually identical. This rejects Proposition 3.1.

The overall fit for the EU is better than for the United States (Figure 5). The difference is not large, but it supports Proposition 3.2.

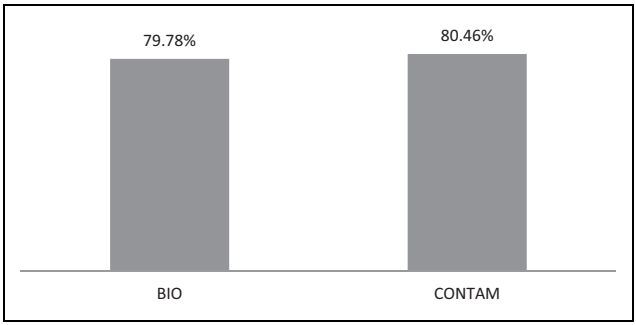


Figure 6. Predictive power of the hierarchical ontology for biohazards and contaminants overall.

The overall fit for contaminants is less than 1 percent better than for biohazards (Figure 6). This is small and not enough to lend support for Proposition 3.3a.

There are some clear differences in the pattern of fit for the two fields of food safety. There are a few UVs where the difference in recall is substantial and there is sufficient number of cases for both fields. In the biohazard field, *epistemic uncertainty, missing factors/variables, limited analytic method, arbitrary assumptions of values used in the model,* and *reporting* are better predicted. The first two have to do with the fact that biohazard deploys these general expression more often.

In the contaminant field, *causal inference*, *comparability of data*, and *combination effect* are predicted better. *Combination effect*, that two health hazards act in concert one amplifying the effect of the other, is a common concern for contaminants, but it is almost absent for biohazards. Therefore, while we see differences between the fields, our Proposition 3.3b is not supported.

Conclusion

In this article, we have demonstrated the value of enlisting the help of ML in creating and deploying an ontology for the content analysis on a complex and important topic, with significant policy implications. Explicit articulation of uncertainty in science, especially in science involved in public policy, improves science because it clarifies what future research is necessary, helps policymakers to evaluate scientific reports, and reminds the public about the limitations of current scientific knowledge.

We have built two complementary ontologies to measure the scientific uncertainty expressed in food safety documents. We have used supervised ML to help with three tasks. First, we want help with coding the thousands of risk assessments in the United States and the EU. Our ontologies were developed for food risk, but it has already been applied to environmental risk and could easily be adopted in other areas. To make document coding easier is a practical matter. In this article, we showed that even with relatively simple methods, ML can do surprisingly well identifying complex meanings and thus can be helpful making suggestions to human coders.

Second, ML can aid us to evaluate our coding practices and our ontologies. We found that our ontologies enable a fairly consistent practice of coding. Evaluating our ontology of judgment, we learned that elements of judgment are often communicated in relationship to one another. In our future work, we will try to exploit these relationships to identify JVs. Assessing our ontology of uncertainty, we found out that the deductive decision-making process, that aids human coders, and that is reflected in the hierarchical structure of our ontology, makes the first large cuts fairly well and confusions tend to emerge between variables that our system defines as being closer.

Third, we wanted to use ML to get insights into the causal processes of making uncertainty explicit. We found some evidence that institutional factors may influence the consistency with which uncertainty is expressed. We also found some indication that scientific cultures by encountering forms of uncertainties with different frequencies articulate uncertainties differently. This warns us that ontology building must be sensitive to the larger social context of the topic it intends to map.

One must know a lot about how texts are structured, produced, and what their contents are before the machine help can be enlisted. One has to understand the nature of the documents and its relationship to the question one is interested in to decide whether unsupervised ML is an option at all. Even if unsupervised techniques are warranted, human judgment plays a role in selecting the corpus and then interpreting and validating the results. With supervised learning, as we have demonstrated, machine and humans work together. After selecting the texts, humans must develop an ontology, the machines can help testing and, if needed, improving the classification scheme. Then, the machine can scale up human coding on its own with humans taking samples to check the quality of the machine's decisions or machines can be used to aid and speed up coding by providing suggestions to human coders.

Acknowledgments

We would like to thank Eve Feinblatt-Mélèze for assisting in the development of the ontologies and Edward Hunter, Kevin Lewis and Juan Pablo Pardo Guerra for their generous comments on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research (HolyRisk Project) was funded by the French National Research Agency (ANR-09-BLAN-0314) and supported by the French National Institute of Agricultural Research (INRA), the Joint Institute of Food Safety and Applied Nutrition (University of Maryland and the U.S. Food and Drug Administration; JIFSAN), the EFSA, the European Commission (EC), the University of California, San Diego, and the French Institute for Research in Computer Science and Automation (INRIA).

Notes

1. This usage of the word "ontology" is very far from its much more general philosophical meaning of the study of nature of being.
2. Our ontology has been adopted in the field of environmental safety.
3. Contaminants are any substance, such as arsenic, cadmium or lead, not intentionally added to food which is present as a result of the production, manufacture, or other steps while holding food or as a result of environmental contamination. Biological hazards include pathogenic viruses, bacteria, and prions that cause bovine spongiform encephalopathy.
4. A detailed description of these ontologies including the codebook with definitions and examples can be found at the website of the HolyRisk Project: http://www6.inra.fr/holyrisk.
5. When a sentence contains multiple expressions of uncertainty, each is represented by different phrases or clauses. Therefore, in principle, at least in some cases we could break up those sentences and pick out the words relevant to each.
6. Apart from the presence or absence of the expression of a category in a sentence, we also coded the intensity of the expression, data we don't present here because in many cases it proved to be impossible to ascertain gradations.
7. Lakoff's original article that set off research on hedges makes the claim that making propositions fuzzier is actually making them more accurate because the

world is fuzzy and truth is a matter of degree. Hedges allow us to move beyond the stark and misconceived binary distinction between truth and untruth.

8. The literature attempts to classify hedges depending on how it deals with uncertainty, whether it serves to protect the author, or whether it just indicates that information is incomplete or that the validity or reliability of the proposition is not fully accepted. We did not make these distinctions.
9. We dropped three variables from the analysis because we did not have enough observations for model fitting and testing. Those are model, sampling, and measure.
10. The errors assume that we correctly decided whether to code a sentence and the calculations are based on sentences correctly identified as uncertainty expressions.
11. When there are only two categories, the overall fit = 1 − confusion.
12. Mushin (2001) calls these "uncertainty sandwiches."
13. We would expect both to be better predicted and they are not. When we look at the most frequent words, the words that one would expect to be most strongly associated with those two uncertainty variables (animal, human, inter, species and intra, species, and variability) are on the list, but they are not the most frequent ones.
14. More common variables tend to have higher recall error and higher pairwise confusion rates.
15. In Figures 4, 5, and 6 the left axis starts at 50 percent.

References

Aggarwal, Charu C. and ChengXiang Zhai, eds. 2012. *Mining Text Data*. New York: Springer.

Aly, Mohamed. 2005. "Survey on Multiclass Classification Methods." Technical Report, California Institute of Technology. Retrieved September 17, 2017. (http://vision.caltech.edu/malaa/publications/aly05multiclass.pdf).

Bail, Christopher. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43:465-82.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55/4:77-84.

Böschen, Stefan, Karen Kastenhofer, Luitgard Marschall, Ina Rust, Jens Soentgen, and Peter Wehling. 2006. "Scientific Cultures of Non-knowledge in the Controversy over Genetically Modified Organisms (GMO): The Cases of Molecular Biology and Ecology." *GAIA-Ecological Perspectives for Science and Society* 15:294-301.

Böschen, Stefan, Karen Kastenhofer, Luitgard Marschall, Ina Rust, Jens Soentgen, and Peter Wehling. 2010. "Scientific Nonknowledge and Its Political Dynamics:

The Cases of Agri-Biotechnology and Mobile Phoning." *Science, Technology & Human Values* 35:783-811.

Cetina, Karin Knorr. 1999. *Epistemic Cultures: How the Sciences Make Sense*. Chicago, IL: University of Chicago Press.

Croissant, Jennifer L. 2014. "Agnotology: Ignorance and Absence or Towards a Sociology of Things That Aren't There." *Social Epistemology* 28:4-25.

Crompton, Peter. 1997. "Hedging in Academic Writing: Some Theoretical Problems." *English for Specific Purposes* 16:271-87.

DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* June–December:1-5.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41:570-606.

Dorne Jean-Lou, C. M. and Andrew G. Renwick. 2005. "The Refinement of Uncertainty/Safety Factors in Risk Assessment by the Incorporation of Data on Toxicokinetic Variability in Humans." *Toxicological Sciences* 86:20-26.

EFSA. 2006. "Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment." *The EFSA Journal* 438:1-54.

Einstein, Albert and Leopold Infield. 1936. *The Evolution of Physics. From Early Concepts to Relativity and Quanta*. London, UK: Cambridge University Press.

Flach, Peter. 2012. *Machine Learning. The Art and Science of Algorithms that Make Sense of Data*. Cambridge, UK: Cambridge University Press.

Frickel, Scott and M. Bess Vincent. 2007. "Hurricane Katrina, Contamination, and the Unintended Organization of Ignorance." *Technology in Society* 29:181-8.

Gaudet, Joanne. 2013. "It Takes Two to Tango: Knowledge Mobilization and Ignorance Mobilization in Science Research and Innovation." *Prometheus* 31:169-87.

Gillund, Frøydis, Kamilla A. Kjølberg, Martin Krayer von Krauss, and Anne I. Myhr. 2008. "Do Uncertainty Analyses Reveal Uncertainties? Using the Introduction of DNA Vaccines to Aquaculture as a Case." *Science of the Total Environment* 407: 185-96.

Gómez-Pérez, Asunción, Mariano Fernández-López, and Oscar Corcho. 2004. *Ontological Engineering—with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. London, UK: Springer Verlag.

Grandjean, Philippe and Esben Budtz-Jorgensen E. 2007. "Total Imprecision of Exposure Biomarkers: Implications for Calculating Exposure Limits." *American Journal of Industrial Medicine* 50:512-9.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21/3:1-31.

Grimmer, Justin and Gary King. 2010. "General Purpose Computer-assisted Clustering and Conceptualization." *PNAS* 108/7:2643-50. Retrieved September 17, 2017. (http://www.pnas.org/content/108/7/2643).

Gross, Matthias and Linsey McGoey, eds. 2015. *Routledge International Handbook of Ignorance Studies*. New York: Routledge.

Gross, Matthias. 2007. "The Unknown in Process Dynamic Connections of Ignorance, Non-knowledge and Related Concepts." *Current Sociology* 55:742-59.

Gross, Matthias. 2012. "'Objective Culture' and the Development of Nonknowledge: Georg Simmel and the Reverse Side of Knowing." *Cultural Sociology* 6:422-37.

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5:199-220.

Hammitt, James K, Jonathan B. Wiener, Brendon Swedlow, Denise Kall, and Zheng Zhou. 2005. "Precautionary Regulation in Europe and the United States: A Quantitative Comparison." *Risk Analysis* 25:1215-28.

Hattis, Dale and David E. Burmaster. 1994. "Assessment of Variability and Uncertainty Distributions in Practical Risk Analyses." *Risk Analysis* 14:713-30.

Hillard, Dustin, Stephen Purpura, and John Wilkinson. 2007. "Computer-Assisted Topic Classification for Mixed Methods Social Science Research." *Journal of Information Technology and Politics* 4:31-46.

Hyland, Ken. 1996. "Writing without Conviction? Hedging in Science Research Articles." *Applied Linguistics* 17:433-54.

Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. 2000. "Statistical Pattern Recognition: A Review." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:4-37.

Japkowicz, Natalie and Mohak Shah. 2011. *Evaluating Learning Algorithms. A Classification Perspective*. Cambridge, UK: Cambridge University Press.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. 2015." Political Language in Economics." Columbia Business School Research Paper No. 14-57. Retrieved September 17, 2017. (http://ssrn.com/abstract=2535453 or http://dx.doi.org/10.2139/ssrn.2535453).

Kang, Seung-Ho, Ralph L. Kodell, and James J. Chen. 2000. "Incorporating Model Uncertainties along with Data Uncertainties in Microbial Risk Assessment." *Regulatory Toxicology and Pharmacology* 32:68-72.

Kempner, Joanna, Jon F. Merz, and Charles L. Bosk. 2011. "Forbidden Knowledge: Public Controversy and the Production of Nonknowledge." *Sociological Forum* 26:475-500.

Krippendorf, Klaus. 2013. *Content Analysis. An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: Sage.

Kroes, Robert, D. Muller, J. Lambe, M. R. H. Lowik, J. van Klaveren, J. Kleiner, R. Massey, S. Mayer, I. Urieta, P. Verger, and A. Visconti. 2002. "Assessment of Intake from the Diet." *Food and Chemical Toxicology* 40:327-85.

Lakoff, George. 1972. "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." *Journal of Philosophical Logic* 2:458-508.

Levin, R., S. O. Hansson, and C. Rudén. 2004. "Indicators of Uncertainty in Chemical Risk Assessments." *Regulatory Toxicology and Pharmacology* 39:33-43.

Lynch, D. and D. Vogel. 2001. *The Regulation of GMOs in Europe and the United States: A Case-Study of Contemporary European Regulatory Politics*. New York: Council on Foreign Relations.

Madsen, Bodil Nistrup and Hanne Erdman Thomsen. 2004. "Ontologies vs. Classification Systems." *Northern European Association for Language Technology Series* 7:27-32.

Merton, Robert K. 1957. *Social Theory and Social Structure*. New York: Free Press.

Merton, Robert K. 1987. "Three Fragments from A Sociologist's Notebooks: Establishing the Phenomenon, Specified Ignorance, and Strategic Research Materials." *Annual review of sociology* 13:1-29.

Millstone, Erik and Patrick van Zwanenberg. 2001. "Politics of Expert Advice: Lessons from the History of the BSE Saga." *Science and Public Policy* 28:99-112.

Morgan, M. Granger and Max Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge, UK: Cambridge University Press.

Mushin, Ilana. 2001. *Evidentiality and Epistemological Stance—Narrative Retelling*. Amsterdam, the Netherlands: John Benjamins.

Myers, Greg. 1989. "The Pragmatics of Politeness in Scientific Articles." *Applied Linguistics* 10: 1-35.

Nardulli Peter, F., Scott L. Althaus, and Matthew Hayes. 2015. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* 45:1-36.

Nautta, Maarten J. 2000. "Separation of Uncertainty and Variability in Quantitative Microbial Risk Assessment Models." *International Journal of Food Microbiology* 57:9-18.

Pate-Cornell, M. Elisabeth. 1996. "Uncertainties in Risk Analysis: Six Levels of Treatment." *Reliability Engineering and System Safety* 54:95-111.

Perkins, Jacob. 2010. *Python Text Processing with NLTK 2.0 Cookbook*. Birmingham, UK: Packt.

Phillips, Lord of Worth Matravers, June Bridgeman, and Malcolm Ferguson-Smith. 2000. *The BSE Inquiry: Report: Evidence and Supporting Papers of the Inquiry into the Emergence and Identification of Bovine Spongiform Encephalopathy*

(BSE) and Variant Creutzfeldt–Jakob Disease (vCJD) and the Action Taken in Response to It Up to 20 March 1996. London, UK: The Stationery Office.

Popper, Karl. [1934] 1959. *The Logic of Scientific Discovery*. London, UK: Hutchinson.

Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 14:130-7.

Proctor, Richard and Londa Schiebinger, eds. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford, CA: Stanford University Press

Smithson, Michael. 1989. *Ignorance and Uncertainty: Emerging Paradigms*. New York: Springer.

Staab, Steffen and Rudi Studer, eds. 2009. *Handbook on Ontologies*. 2nd ed. Heidelberg, Germany: Springer.

U.S. EPA. 1993. *Descriptive Guide to Risk Assessment. Methodologies for Air Toxics. 453-R-93-038*. Washington, DC: U.S. Environmental Protection Agency.

U.S. EPA. 2000. *Supplementary Guidance for Conducting Health Risk Assessment of Chemical Mixtures. 630-R-00-002*. Washington, DC: U.S. Environmental Protection Agency.

U.S. OMB. 2006. *Proposed Risk Assessment Bulletin*. Washington, DC: U.S. Office of Management and Budget.

van Asslet, Majrolein, B. and Jan Rotmans. 2002. "Uncertainty in Integrated Assessment Modelling. From Positivism to Pluralism." *Climatic Change* 54:75-105.

van der Sluijs, Jeroen, P., Matthieu Craye, Silvio Funtowicz, Penny Kloprogge, Jerry Ravetz, and James Risbey. 2005. "Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System." *Risk Analysis* 25:481-92.

van Zwanenberg Patrik and Eric Millstone. 2005. *BSE: Risk, Science and Governance*. Oxford, UK: Oxford University Press.

Vázquez, Ignacio and Diana Giner. 2009. "Writing with Conviction: The Use of Boosters in Modelling Persuasion in Academic Discourses." *Revista Alicantina de Estudios Ingleses* 22:219-37.

Wagner-Pacifici, Robin, John W Mohr, and Ronald L Breiger. 2015. "Ontologies, Methodologies, and New Uses of Big Data in the Social and Cultural Sciences." *Big Data & Society* June–December:1-11.

Walker, Warren, Peter Harremoës, Jan Rotmans, Jeroen van der Sluijs, Majrolein B. A. van Asselt, Peter Jansen, and Martin P. Krayer von Krauss. 2003. "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-based Decision Support." *Journal of Integrated Assessment* 4:5-17.

WHO. 2008. *Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment. World Health Organization/International Program on Chemical Safety*. Geneva, Switzerland: WHO Press.

Winkler, Robert L. 1996. "Uncertainty in Probabilistic Risk Assessment." *Reliability Engineering & System Safety* 54:127-32.

Author Biographies

Akos Rona-Tas is professor of sociology at UC, San Diego. He works on problems of uncertainty in science and the economy. He was the co-director of the HolyRisk project. His latest book is Plastic Money: Constructing Markets for Credit Cards in Eight Postcommunist Countries (Stanford University Press) written with Alya Guseva.

Antoine Cornuéjols is professor of computer science at AgroParisTech. He is associated with the "Modélisation Mathématique, Informatique et Physique" (MMIP) research department, and a corresponding member of the Thème Apprentissage et Optimisation research team at the Laboratoire de Recherche en Informatique at the Université de Paris-Sud, Orsay (France). His fields of research lie in machine learning, data mining and cognitive science.

Sandrine Blanchemanche is a sociologist, and a research scientist at INRA-Paris. She was the co-director of the HolyRisk project. Her research is on risk, uncertainty, and public policy in the area of food science.

Antonin Duroy was a researcher at AgroParisTech. He is currently lead data scientist at FlameFy.

Christine Martin is an associate professor of computer science at AgroParisTech. Her research is in machine learning, data mining and bioinformatics.