# Empirical Ontology of Scientific Uncertainty: Expression of Uncertainty in Food Risk Analysis

Sandrine Blanchemanche*, Akos Rona-Tas**,
Antoine Cornuéjols***, Antonin Duroy***, Christine Martin***

*Met@risk, INRA, Paris (France)
sandrine.blanchemanche@paris.inra.fr
http://www.une-page.html
** University of California, San Diego (USA)
aronatas@ucsd.edu
http://www.une-autre-page.html
*** AgroParisTech-INRA UMR-518 MIA
16, rue Claude Bernard – F_75231 Paris Cedex 05 (France)
antoine.cornuejols,christine.martin@agroparistech.fr

## Abstract

We inductively devised two ontologies to describe scientific uncertainty in food safety risk assessments. We ask three questions. 1. Can we use Machine Learning to assist with coding complex documents such as food safety risk assessments? 2. Can we assess the quality of the ontologies we devised using ML? 3. And, finally, does the quality of our ontologies depend on social factors? We found that Machine Learning can do surprisingly well identifying complex meanings and probably can be helpful making suggestions to human coders. Our ML experiments show that our ontologies do enable a fairly consistent practice of coding. And finally, we found some evidence that institutional factors do influence how well predictable our ontology of uncertainty is.

## 1. Introduction

> "When you have got the clear support of the scientists that deal with these matters […] there is no need to be worried, and I can say perfectly honestly, that I shall go on eating beef as my children will go on eating beef, because there is no need to worry." John Grummer

On May 16, 1990, John Gummer, British Minister of Agriculture invited the media to a food fare during mounting public concern about a mysterious disease attacking cattle. There was little doubt, that bovine spongiform encephalopathy (BSE), also known as mad cow disease, was devastating cattle in the British countryside, but whether it was a threat to other livestock and humans, was less clear. In 1988, the government appointed a working group of scientists, to answer this question. The Southwood Working Party rendered its confident conclusions in February 1989, that BSE poses no threat to other species (Millstone and van Zwanenberg 2001). More than a year later, to put all doubts to rest, Gummer, stood in front of a food tent with his daughter. He handed a freshly prepared hamburger to her to prove British beef was harmless. The little girl found the burger too hot, and thus to make his point, Gummer had to devour the beef patty in the bun himself.

The "clear support of the scientists that deal with these matters," however, was less than unanimous. Many scientists voiced doubt about Southwood's conclusion and pointed to already existing evidence and theory suggesting BSE could easily jump species. The author of one such theory, postulating that cellular protein can fold into infectious prions a mechanism common to

BSE and its human variant, the Creuzfeldt-Jacob Disease, Stanley Prusiner, got the Nobel prize for medicine in 1997.

While the subsequent parliamentary inquiry and scrutiny by independent scholars did not find clear cases of falsifying or distorting evidence, they did find that Southwood did not acknowledge uncertainties in existing research and presented a much more self-confident conclusion than was warranted (Phillips et al 2000). It was also clear that the Working Party was responding to political fears that British meat production may sustain extensive damage if British beef is pronounced unsafe. By not hiding uncertainties in the data and not articulating them properly, the Working Party was able to bend its findings to suit the political ends of the government. [1]

In the wake of the BSE debacle, various international agencies began to emphasize the importance of expressing scientific uncertainty in food safety risk assessment documents. Some, like WHO, EFSA, US EPA, US OMB, issued guidelines working towards a system that both identifies the type of uncertainty scientist perceive and the extent to which our knowledge is uncertain in a particular respect. 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 was adopted systematically by scientific panels.

Our approach to scientific uncertainty is not normative, but empirical and comparative. We want to describe 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 between 2000 and 2010. We investigate two main, distinct areas of food hazards in the food risk field: contaminants and biohazards.[2] As the two fields draw on different subdisciplines, they differ in the way they make use

[1] In an ironic twist, in 2003, the French government wanted to its scientists at AFSSA, to conclude that British beef was more dangerous than French beef, and it was dismayed when AFSSA concluded that there was too much uncertainty to compare the two. (van Zwanenberg and Millstone 2005, p.271).

[2] 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 BSE.

of various scientific methodologies (experiments, observational studies, statistical analyses, analytic modeling etc.) and thus may have different understandings of scientific uncertainties.

We coded the documents by human experts, then tested these ontologies using machine learning. Our main objective was to answer three questions. 1. Can we use Machine Learning to assist with coding complex documents such as food safety risk assessments? Is ML doing a reasonable job overall in coding sentences? 2. Can we assess the quality of the ontologies we devised using ML? 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 learning, institutions and culture?

In the rest of the paper, we first describe the two ontologies, and our data. Then we explain the use of machine learning and the choices we made to conduct our experiments. We pose the three sets of questions, propose hypotheses and discuss the empirical results.

## 2. The Two Ontologies

To map scientific uncertainty, we developed two complementary ontologies, in an inductive and iterative process. Ontology is "an explicit specification of a conceptualization" (Gruber 1993:199). In full-fledged ontologies, concepts and their relationships 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. 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 developed two systems of classification or simple ontologies. The first, a simpler ontology is designed to capture the nature of the judgment the scientists make about the uncertainty of their conclusion. This ontology is a typology, a multidimensional way of classifying the expert's judgment of the evidence. The second, a hierarchical system gauges the content of uncertainty. 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 successive levels of generality.

For both ontologies, our smallest coded unit is the sentence (our data point). Categories can be attached to one or more consecutive sentences. One sentence can contain multiple expressions of

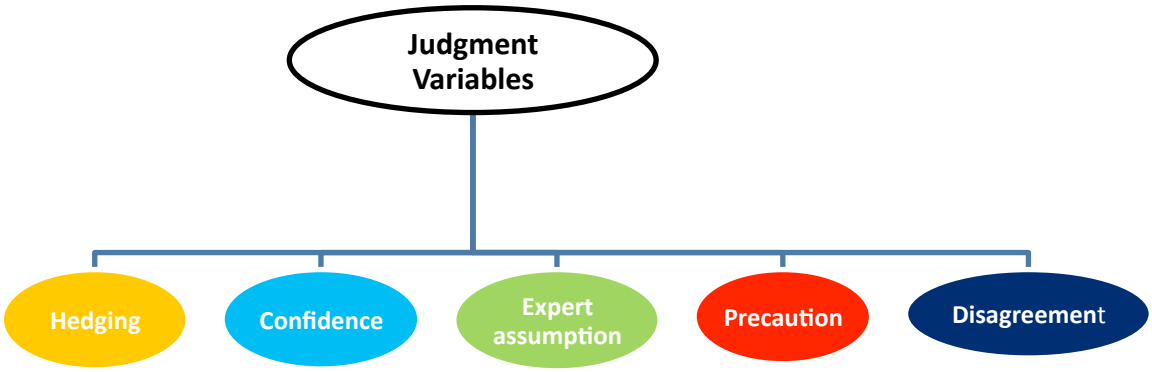
uncertainty and can be sorted into multiple categories.[3] We also refer to categories as "variables" because their values vary from sentence to sentence.[4] Consequently, we talk about judgment variables (JVs) and uncertainty variables (UVs) when we talk about categories of the first and the second ontology.

As we were interested in the final verdict of the experts, just as policy makers 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.

### *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), two (confidence and expert assumption) communicate the opposite.

**Figure 1. The structure of the ontology of uncertainty based on judgment**




*Hedging* is a way of indicating that experts have doubt about or a lack of total commitment to a proposition they present. There is a large literature on hedging. Hedging, a way of making things

[3] When a sentence contains multiple expressions of uncertainty, each is represented by different phrases or clauses. Therefore, in principle, we could break up those sentences and pick out the words relevant to each.

[4] Apart from the presence or absence of the expression of a category in a sentence, we also coded the nature and intensity of the expression, data we will analyze later.

fuzzier (Lakoff 1972),[5] expresses a "lack of complete commitment to the truth value of an accompanying proposition" (Hyland 1998:1). 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 the sentence, albeit the causes may be explained elsewhere in the text.[6] 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, Vazquez 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, 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 that it was demonstrated that the statistical models cannot answer the crucial question. In such cases, there is uncertainty and confidence is to emphasize that it is there.

Our third category is *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.

---

[5] 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.

[6] 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.

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 US that developed mostly in the context of genetically modified organisms (Lynch and Vogel 2001, Hammitt et al. 2005).

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, we began with the general literature on scientific uncertainty (Morgan and Henrion 1990, Hattis and Burmaster 1994, 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, (Grandjean & Budz Jorgensen 2007, Kang & Kodell et al. 2000, Nautta 2000, Dorne and Renwick 2005, and Kroes et al. 2002). Beside this literature, we drew upon two main institutional documents: the opinion of the Scientific Committee of EFSA entitled Uncertainties in Dietary Exposure Assessment (EFSA 2006) and the WHO Draft guidance document on Characterizing and Communicating Uncertainty in Exposure Assessment (WHO 2007). We simplified and adapted the basic structure of these classification systems through a series of test coding of European, US 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.

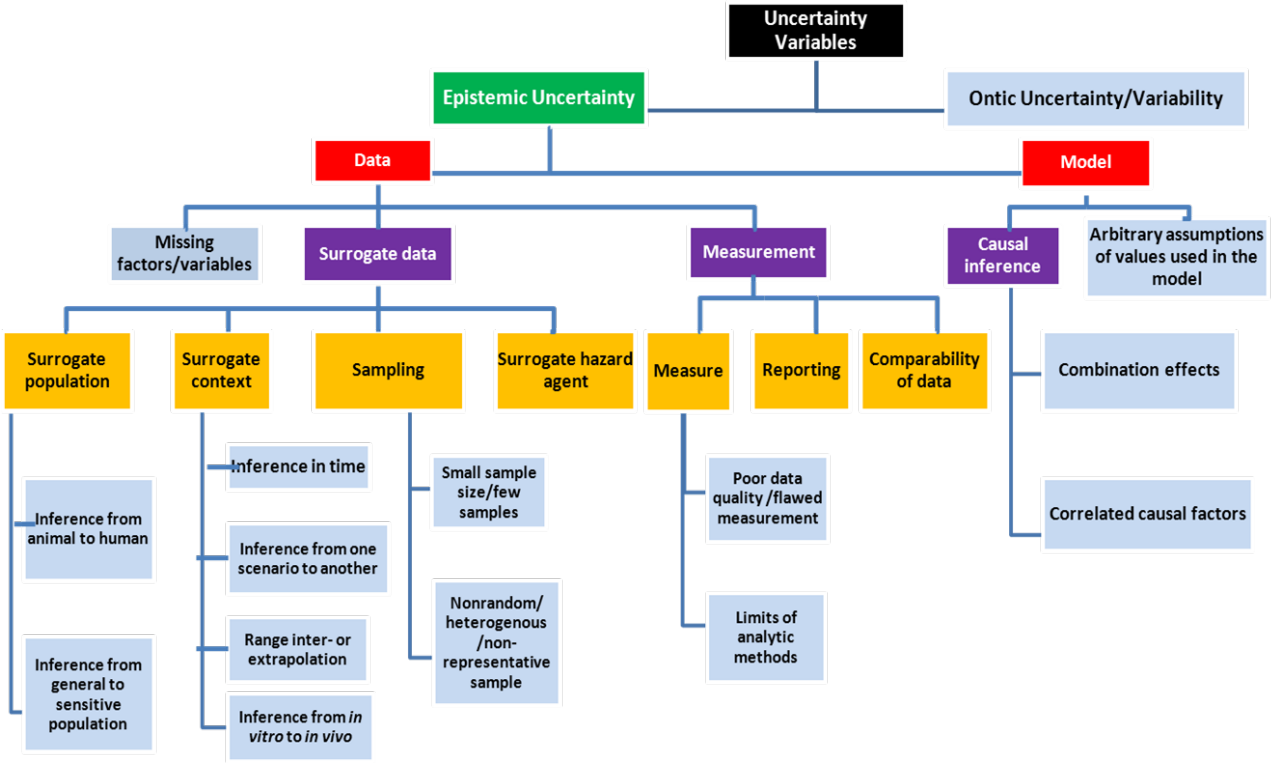
**Table 1. Decision Tree for Uncertainty Taxonomy Coding**

<table>
<tr><td rowspan="18">Is it uncertainty that is irreducible?<br>OR<br>Is it that new information can resolve</td><td colspan="5">Ontic Uncertainty/Variablity</td></tr>
<tr><td rowspan="17">Epistemic Uncertainty<br><br>Is it due to the absence of good data about the hazard?<br>OR<br>Is it due to the way the model is built?</td><td rowspan="3">Model<br>Is it due to arbitrary model assumptions?<br>OR<br>Is it due to some problem in our causal understanding i.e. what generates the hazard?</td><td colspan="3">Arbitrary assumptions of model</td></tr>
<tr><td colspan="2" rowspan="2">Causal inference<br>Is uncertainty due to ignoring synergism (combination effects)?<br>OR<br>Is it due to the inability to separate the effects of related causes?</td><td>Combination effects</td></tr>
<tr><td>Correlated causal factors</td></tr>
<tr><td rowspan="15">Data<br>Is it due to the complete absence of data<br>OR<br>Is it due to the lack of the exact kind of data we need and the fact that we have to use proxies (surrogates)?<br>OR<br>If we have the right kind of data, does some quality of the data create uncertainty ?</td><td colspan="3">Missing factor</td></tr>
<tr><td rowspan="14">Measurement<br>Data from different sources are incomparable and they point in different directions.<br>OR<br>We don't know enough about how it was measured to trust the data.<br>OR<br>Is it due to how it is measured?</td><td colspan="2">Comparability of data</td></tr>
<tr><td colspan="2">Reporting</td></tr>
<tr><td rowspan="2">Measure<br>Was the measurement poorly done?<br>OR<br>Does the methodology used in measuring have inevitable limitations?</td><td>Flawed measure</td></tr>
<tr><td>Limited analytic method</td></tr>
<tr><td rowspan="6">Surrogate Data<br>Is it a sampling problem?<br>OR<br>Is there a discrepancy between the hazard we have data for and the hazard of interest?<br>OR<br>Is there a discrepancy due to context?<br>OR<br>Is there a discrepancy between the population of interest and the population we have information on?</td><td rowspan="2">Sampling<br>Was the sample too small<br>OR<br>Was it selected improperly</td><td>Small sample size /few samples</td></tr>
<tr><td>Non-representative sample</td></tr>
<tr><td colspan="2">Surrogate hazard</td></tr>
<tr><td rowspan="4">Surrogate context<br>Are the data from the wrong context?</td><td>Inference in time</td></tr>
<tr><td>Scenario inference</td></tr>
<tr><td>Range inter- or extrapolation</td></tr>
<tr><td>Inference from in vitro to in vivo</td></tr>
<tr><td rowspan="2">Surrogate Population<br>Are the data from the wrong population</td><td>Inference from animal to human</td></tr>
<tr><td>Inference from general to sensitive population</td></tr>
</table>

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 Table 1 or the light blue label in Figure 2), were either

unspecified at a lower level or were specified but the specification was so rare that it did not deserve a separate category at the next level.

**Figure 2. The structure of the ontology of uncertainty based on content**




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 that provided the best clue.

Our ontology begins with the common 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."

## 3. The documents

The text corpora of text we coded were 115 official risk assessment documents produced by the European Food Safety Administration (EFSA) in Europe, and the three U.S. federal agencies primarily responsible for food safety across the Atlantic, the Food and Drug Administration (FDA), the U.S. Department of Agriculture (USDA) and the Environmental Protection Agency (EPA) between 2000 and 2010.

The documents range from one to over 600 pages with average of around 70 pages. They are written at the official request of the authorities responsible for managing the risk, by a panel of scientists. [7]

## 4. Using Machine Learning

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

[7] 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. Therefore, our actual unit is the hazard-document.

similar to the one used by the human coders. For instance, the system should be able to code a text input as 'coded' vs. 'non coded' 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 European Union. Or that the rules change over time.

In the estimation, we used naïve Bayesian classifier, Support Vector Machines (SVM), *k* Nearest Neighbor and Decision Tree algorithms (see Flach 2012). The best overall learners were Naïve Bayes and SVM.

We started with the bag-of-words approach. We broke the sentences up into words. We eliminated "stop words," i.e. words that were too common to be helpful such as "the," "a," "is" etc. Then we stemmed the words thus erasing the difference between, for instance, learning, learned, learns etc. 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 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. E.g. 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 multi-class 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 multi-class ones. In our experiments, we used the all-versus-all technique (see Aly 2005).

One problem when learning to separate data from different categories is that their frequency can be significantly dissimilar (e.g. 1 sentence falls under the '*Measure*' category, while 71 fall under the '*Missing factors / variable* category). If, for instance, one category is ten times more highly represented than another one, a simple majority rule will yield a 90% 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 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 and another technique must 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 and made up an artificial data point by retaining for each feature either the one encountered in both actuals if they agreed, or by randomly choosing a value of one actual if they disagreed on this feature.

The data then was split into two halves. One half was the "training set," where the algorithm calculated the best fitting model, then it 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 five-fold cross-validation technique (see Japkowicz et al. 2011).

In this paper, 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.

## 5. Discussion

In this section we will pose three questions, present some hypothesis and discuss our results.

### *Can we use Machine Learning (ML) to assist with coding complex documents?*

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, conclusion and there is also often a

summary in front. Scientists learn how to write risk assessments, what each element should include etc.

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.

### Hypothesis

> H1.1: The task involves complex meanings. The more the method can incorporate complexities, the better the tool is going to be. 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 if 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 errors. Thus, ML learns from human coders, and human coders can learn from ML.

There were 178 mismatched sentences where ML disagreed with human coders. Upon individual inspection, we found that the human coders were correct in 87% and ML was correct in 6% of the cases. For another 5% of the cases both were wrong and for rest the sentence did not express uncertainty (essentially, human error). There were also three cases where the same sentence had more than one code. In 2 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 if 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 small 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 the whether the sentence is coded or not. For judgment variables,

Support Vector Machine (SVM) estimation gave the best results. Eighty-four percent of the sentences were correctly identified as being coded for one of the five categories (Table 2.) Recall was 80.6 percent, ML failed to recognize one fifth of the coded sentences.

Table 2. The success of finding sentences that are coded with the judgment ontology

| SVM | | Predicted | | | |
|---|---|---|---|---|---|
| | | Coded as Judgment | Not Coded | *Recall* | *Nb elmnts* |
| Actual | Coded as Judgment | **378** | 91 | *80.60%* | 469 |
| | Not Coded | 70 | **474** | *87.13%* | 544 |
| | *Precision* | *84.38%* | *83.89%* | | |
| | | | | Total | 1013 |
| | **Overall** | **84.11%** | | | |

For the uncertainty variables, ML was able to correctly identify 79.85% of the sentences (Table 3). Here too, a fifth of the coded sentences went unrecognized.

Table 3. The success of finding sentences that are coded with the uncertainty ontology

| SVM | | Predicted | | | |
|---|---|---|---|---|---|
| | | Coded as Uncertainty | Non Coded | *Recall* | *Nb elmnts* |
| Actual | Coded as Uncertainty | **447** | 108 | *80.54%* | 555 |
| | Non Coded | 109 | **413** | *79.12%* | 522 |
| | *Precision* | *80.40%* | *79.27%* | | |
| | | | | Total | 1077 |
| | **Overall** | **79.85%** | | | |

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,[8] 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 uncertainty variables 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 hypothesis. 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% (Table4). For UVs, it was 78.82% (Table 5). [9]

Table 4

| | | 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%** | | | | | |

[8] http://wordnet.princeton.edu/

[9] We dropped 3 variables from the analysis because we did not have enough observations for model fitting and testing. Those are *Model*, *Sampling* and *Measure*.

**Table 1**

| | | Predicted | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Arbitrary a | Causal infe | Combinati | Comparab | Correlatec | Data | Epistemic | Inference | Inference | Inference | Inference | Inference | Limits of a | Measurem | Missing fa | Non-rando | Ontic unce | Poor data | Range inte | Reporting | Small sam | Surrogate | Surrogate | Surrogate | Surrogate | *Prediction* |
| Actual | Arbitrary assumptions of values used in the mod | **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 scen | 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* |
| | Non-random / heterogeneous / non-representati | 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** | | | | | | | | | | | | | | | | | | | | | | | | | |

### *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]

## Judgment Variables

Our ontology of the Judgment variables 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.

### *Hypotheses*

H2.1: Categories that have opposite connotations will be less likely to be confused. Therefore, *hedging* will be less likely to be confused with *confidence* that 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 assumptions* 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 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.

[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.

Table 6. Confusion among judgment variables

| | | Predicted | | | | |
|---|---|---|---|---|---|---|
| | | Confidence | Disagreement | Expert assumption | Hedged language | Precaution |
| Actual | Confidence | | 0.000 | 0.022 | 0.092 | 0.007 |
| | Disagreement | | | 0.000 | 0.000 | 0.000 |
| | Expert assumption | | | | 0.057 | 0.055 |
| | Hedged language | | | | | 0.070 |
| | Precaution | | | | | |

Table 6 reveals that our hypothesis 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 assumptions* for the same reason: hedging balances certitude. (*Disagreement* is too rare to analyze.)

## Uncertainty Variables

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.

### *Hypotheses*

H2.2a: 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 variables have more similar meanings and thus are easier to confuse.

H2.2b: However, we may find the opposite: confusions increase with distance. Variables that we find farther in the ontology are more likely to be confused. This could happen because small differences require more explicit demarcations that hinge on a single (or very few) trait(s). E.g., once we agree that this is a sampling problem, we can more easily decide if it is the size or the representativeness that is the problem (or both). What is harder to agree on is whether it is

[12] Mushin calls these "uncertainty sandwiches." (Mushin 2001)

primarily a sampling problem or a case of a variable missing or causal problem. Larger distances hinge on multiple traits which involve multiple and often contradictory differences and similarities.

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 5, 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, 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) or *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 7 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 non-random / non-representative sample.*

[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 UVs (animal, human, inter, species and intra, species, variability) are on the list but they are not the most frequent ones.

Table 7. The pairs most confused

| Pair | | Distance | Confusion |
|---|---|---|---|
| *Missing factors/variables* | *Limits of analytical methods* | 4 | 0.122 |
| *Missing factors/variables* | *Inference from one scenario to another* | 4 | 0.081 |
| *Non-random / non-representative sample* | *Missing factors/variables* | 4 | 0.079 |
| *Limits of analytical methods* | *Comparability of data* | 3 | 0.075 |
| *Ontic uncertainty / Variability* | *Limits of analytical methods* | 6 | 0.075 |
| *Inference from one scenario to an another* | *Arbitrary assumptions of values used in the model* | 6 | 0.071 |
| *Missing factors/variables* | *Data* | 1 | 0.063 |
| *Limits of analytical methods* | *Epistemic uncertainty* | 4 | 0.061 |
| *Missing factors/variables* | *Arbitrary assumptions of values used in the model* | 4 | 0.058 |
| *Non-random / non-representative sample* | *Comparability of data* | 5 | 0.058 |
| *Ontic uncertainty / Variability* | *Comparability of data* | 5 | 0.057 |
| *Ontic uncertainty / Variability* | *Epistemic uncertainty* | 2 | 0.057 |
| *Missing factors/variables* | *Comparability of data* | 3 | 0.056 |

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 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 H2.2a.

### *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 social forces? 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.

[14] More common variables tend to have higher recall error and higher pairwise confusion rates.

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 US documents to search for larger institutional differences. And, finally, we also contrasted documents discussing contaminants and biohazards to see if different scientific fields may articulate uncertainty differently. We concentrated here only the uncertainty ontology.

**Hypotheses:**

H3.1: 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 then 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.

H3.2: The EU documents will be better predicted overall by ML than those from the US. This is an institutional theory. Because in the EU risk assessment in food safety is more centralized than in the US and the RAs are more standardized, Eurpoean 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 US, there are three principal agencies in charge of this topic (FDA, USDA, 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 US 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.

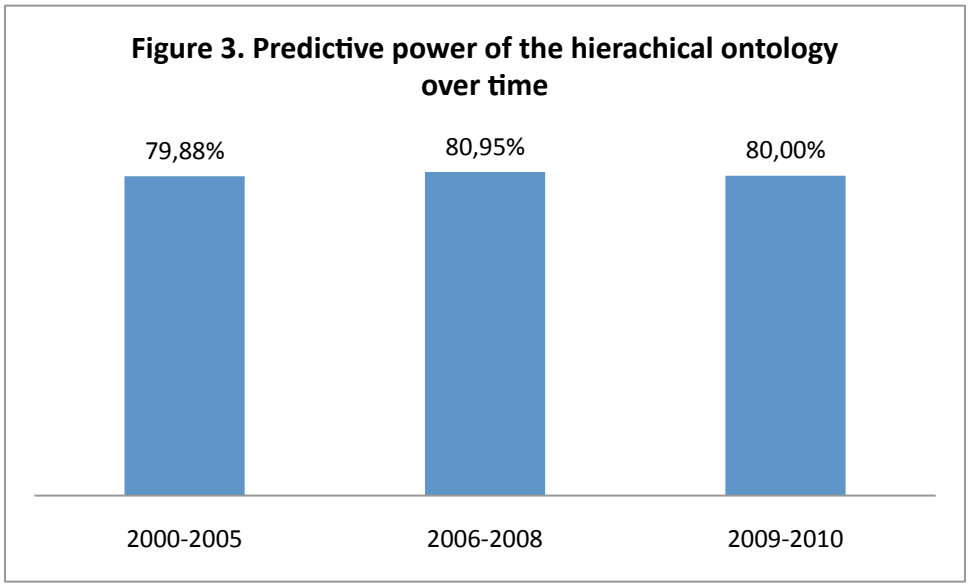
H3.3: There will be a pattern of prediction that will be different in the world of contaminants and biohazards. This is a theory about different scientific cultures driven by their subject matters and histories. We do see that the relative frequency of various uncertainties differ across the two worlds. E.g., contaminant research use 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*.

H3.3a: Overall contaminants will be better predicted because contam 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, zoonooses, such as BSE, avian flu or new strains of older viruses or bacteria).

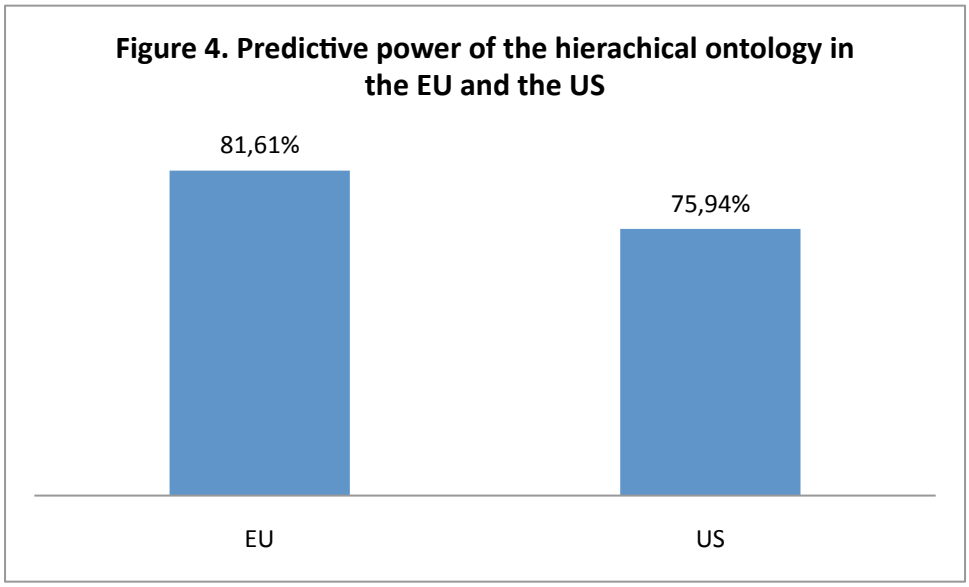
H3.3b: Each field will be better predicted on issues that are more central to that field, because frequent concerns are expressed in a more ritualized form.

## Results

The overall fit for earlier RAs is not different than for later RAs. As is seen in Figure 3, there is no learning.[15] The three periods are virtually identical. This rejects H3.1.

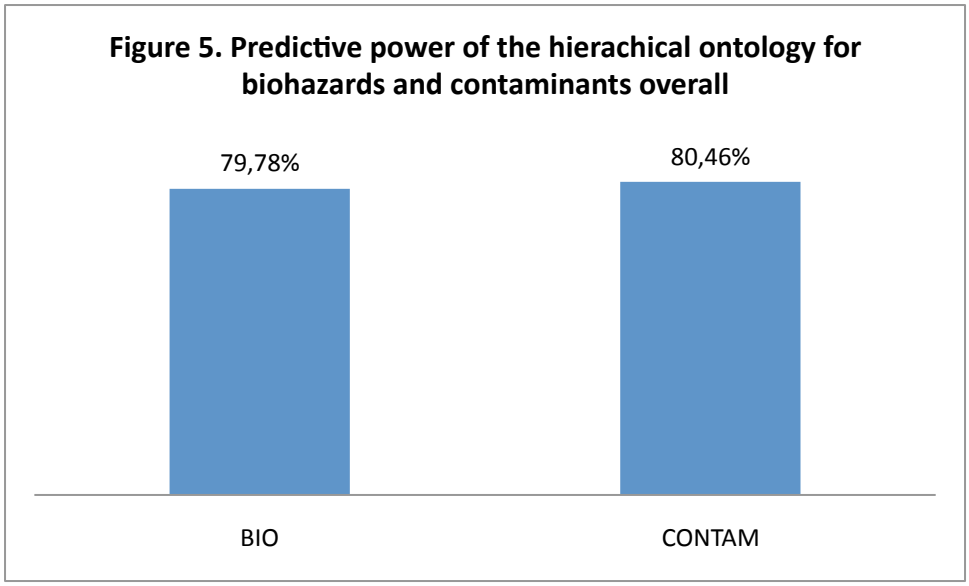



The overall fit for the EU is better than for the US (Figure 4). The difference is not large, but it supports H3.2.




The overall fit for contaminants is less than 1 percent better than for biohazards (Figure 5). This is small and not enough to lend support for H3.3a.

[15] In Figure 3,4,5 the left axis starts at 50%.




There are some clear differences in the pattern of fit for the two fields of food safety. There are a few uncertainty variables 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 H3.3b is not supported.

## 6. Conclusion

Explicit articulation of uncertainty in science, especially in science involved in public policy, improves science, because it clarifies what future research is necessary, helps policy makers 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 enlisted Machine Learning to help with three tasks. First, we want help with coding the thousands of risk assessments in the US 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 paper, we showed that even with relatively simple methods, Machine Learning 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 judgment variables. 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.

## 7. References


Aly, M. 2005. "Survey on Multiclass Classification Methods." *Neural Networks*, pp.1-9.

Crompton, Peter. 1997. "Hedging in Academic Writing: Some Theoretical Problems." *English for Specific Purposes*, 16/4, pp. 271-287

Dorne J. L. and A.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, pp. 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, pp. 1-54.

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

Grandjean, P. and E. Buz-Jorgensen E. 2007. "Total Imprecision of Exposure Biomarkers: Implications for calculating Exposure Limits." *American Journal of Industrial Medicine*, 50, pp. 512-519.

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition*, 5/2:199-220.

Hammitt, J. K, J. B. Wiener, B. Swedlow, D. Kall and Z. Zhou. 2005. "Precautionary Regulation in Europe and the United States: A Quantitative Comparison." *Risk Analysis*, 25, 1215-1228

Hattis, Dale and David E. Burmaster. 1994. "Assessment of Variability and Uncertainty Distributions in Practical Risk Analyses." *Risk Analysis*, 14/5, 713-730

Hyland, Ken. 1996. "Writing without Conviction? Hedging in Science Research Articles." *Applied Linguistics*, 17/4:433-454

Japkowicz, N. and Shah, M. 2011. *Evaluating Learning Algorithms. A Classification Perspective*. Cambridge University Press.

Kang S., R.L. Kodell, and J.J. Chen. 2000. "Incorporating model uncertainties along with data uncertainties in microbial risk assessment." *Regulatory Toxicology and Pharmacology* 32, pp. 68–72.

Kroes, R., D. Muller, J. Lambe, M.R.H. Lowik, J. van Klaveren, J. Kleiner, R. Massey, S. Mayer, I. Urieta, P. Verger, A. Visconti. 2002. "Assessment of intake from the diet." *Food and Chemical Toxicology* 40, 327–385

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

Levin, R., Hansson, S.O. and Rudén, C. 2004. "Indicators of uncertainty in chemical risk assessments." *Regulatory Toxicology and Pharmacology*, 39, pp. 33-43.

Lynch, D., Vogel, D. 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, Inc

Millstone, Erik and Patrick van Zwanenberg. 2001. "Politics of expert advice: lessions from the history of the BSE saga." *Science and Public Policy*, 28/3:99-112.

Morgan, M.G. and M. 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 and Philadelphia: John Benjamins Publishing Co.

Myers, Greg. 1989. "The pragmatics of politeness in scientific articles." *Applied Linguistics* 10, pp. 1-35.

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

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

Perkins, J. 2010. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing.

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* (The Stationery Office, London).

Porter, M.F. 1980. "An algorithm for suffix stripping." *Program*, 14(3), pp. 130-137

van Asslet, Majrolein B. and Jan Rotmans. 2002. "Uncertainty in Integrated Assessment Modelling. From Positivism to Pluralism." *Climatic Change*, 54, pp. 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*, 25/2, pp. 481-492.

van Zwanenberg P & Millstone E (2005), *BSE: risk, science and governance*, 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, pp. 219-237

Walker, W., Harremoës, P., Rotmans, J., Van der Sluijs, J., Van Asselt, M.B.A, Jansen, P., Krayer von Krauss M.P. 2003. "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support." *J Integr Assess*, 4/1, pp. 5-17.

WHO/IPCS (World Health Organization/International Program on Chemical Safety). 2007. Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment