From Uncertainty to Semantics in Self-reported Data: an Empirical Analysis

Salvatore F. Pileggi[0000−0001−9722−2205] and Gnana Bharathy[0000−0001−8384−9509]

School of Computer Science, Faculty of Engineering and IT
University of Technology Sydney (Australia)
{SalvatoreFlavio.Pileggi, Gnana.Bharathy}@uts.edu.au

Abstract. Survey data often presents uncertainty because of missing values or situations that do not allow the measure of a given variable, for instance inability/reluctance to answer. On the other side, the sensitivity of questions may affect the quality of data, as well as its reliability and interpretation. Intuitively, uncertainty in such kind of data could be related some kind of criticality, more concretely to topic sensitivity in this specific case. This paper reports an empirical study conducted on a subset of the World Values Survey (WVS) aimed at the assessment of the relationship between uncertainty and topic sensitivity in survey data. The experiment shows a fundamental convergence and, although results cannot be generalised because of the limited number of experiments conducted, it establishes the fundamentals for a more systematic approach in the context of the current technological landscape, which offers the capabilities to enable human-centric and fully automated solutions. Last but not least, the critical analysis looking at current limitations has defined a roadmap to further enhance the proposed method aiming at a broader and more consolidated experimental and validation framework.

Keywords: Knowledge Engineering · Data Engineering · Data Quality · Sensitive Data · Uncertainty · Self-reported Data · Human-centric AI

1 Introduction

Uncertainty in its different forms (e.g. aleatory and epistemic [12]) is inherent in a computational world [4]. While potentially any kind of data may present some degree of uncertainty, in a data-intensive society uncertainty tends to be associated mostly with the complexity of a given domain and related applications, such as, among very many, Big Data Analytics [8] and business [2]/economic [6] complexity.

While apparently less critical, also survey data may present uncertainty. That is normally because of missing data and is related more or less directly to the inability/reluctance of participants to answer. A well-designed survey should offer the possibility to disclose such situations in context (e.g. "*Prefer not to answer*"), as well as such situations should be transparently reported with proper codes in the resulting dataset to contribute to assure a proper quality standard.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

On the other side, questions on sensitive topics are relatively common and often critical [19] as it is widely accepted that the sensitivity of questions may affect the quality of data, as well as its reliability and interpretation. However, the intrinsic complexity of topic sensitivity has not been widely studied across the different dimensions and variety of possible contexts [17]. Literature proposes some works on specific methods to address sensitive topics (e.g. [16]).

The huge body of knowledge in literature on uncertainty addresses multiple perspectives [12], from a more conceptual/philosophic view to solutions across the different domains. However, at the very best of our knowledge, there is a fundamental lack of study on the relationship between uncertainty and topic sensitivity in survey data. An interesting study on the topic links sensitive topics to resulting measures or estimations [18], discussing the facto a conceptual relationship.

In order to contribute to bridge such a research gap, this study empirically approaches the issue looking at the following research goals:

RG-1. Empirical assessment of the relationship between uncertainty and topic sensitivity in survey data.

RG-2. Systematic context-specific topic sensitivity assessment.

RG-3. Establish the fundamentals for a systematic validation of the relationship between uncertainty and topic sensitivity in survey data.

The current approach intrinsically relies on the potential of the modern AI technology. However, the focus is on systematic, transparent and reliable human-centric AI solutions [11], which may eventually be customised and tuned by human inputs within explainable environments [21].

Structure of the paper. The core part of the paper is structured in three sections that respectively address the methodological aspects (Section 2), an empirical analysis of a relevant case study (Section 3) and a critical discussion of the results looking at main limitations and future research (Section 4).

2 Methodology and Approach

A workflow-based representation of the adopted methodology is proposed in Fig. 1. As shown, the input for the iterative process is survey data, meaning a dataset composed of a number of survey questions and their measure.

As a very first step, questions are processed to extract the main topic. In the context of this work, such a step is not required as the considered case study [13] includes a conceptualisation. More in general, topics can be extracted from input questions by adopting automatic tools (e.g. BERT [10] based). The output for this step is an aligned set of relevant concepts [7], each one associated with one or more questions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

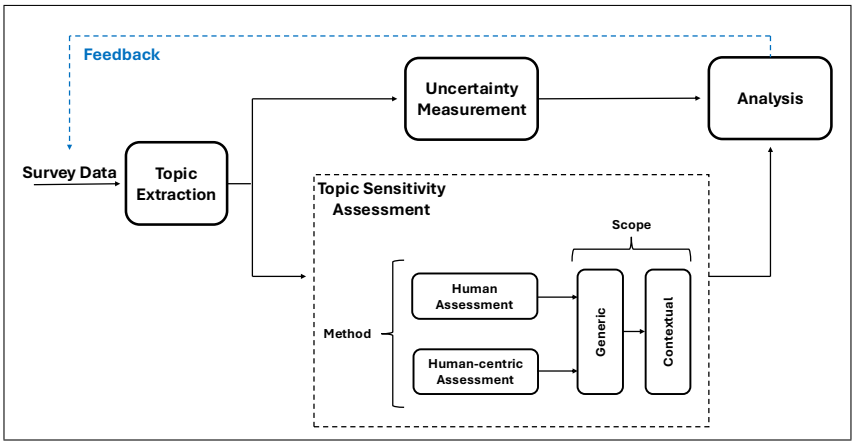


Fig. 1: Methodology overview.

The conceptualization is followed by two independent processes, which aim to (*i*) quantify data uncertainty (Section 3.1) and (*ii*) assess topic sensitivity (Section 3.2) respectively. Uncertainty is semantically associated with missing values or situations that do not allow the measure of a given variable (for instance inability/reluctance to answer), while topic sensitivity is a much more abstracted concept as per common meaning. To remark that uncertainty in this specific case can be measured according to objective criteria; on the contrary, the assessment of sensitivity is contextual and, in general terms, subject to bias and multiple interpretations. Therefore, topic sensitivity is assessed by adopting different methods (human and human-centric) with a progressive scope refinement, from generic to contextual.

Finally, the output of the mentioned processes are object of analysis to estimate the potential convergence between uncertainty and topic sensitivity.

3 Empirical Analysis: a Case Study

The empirical analysis object of this paper is performed on a subset of the *World Values Survey (WVS)* [9]. It is characterised by a relatively low dimensionality and includes a conceptualization [13].

A reduced dimensionality (16 features) is more suitable to this initial study, as it allows a more intelligible framework of analysis. Moreover, a consolidated conceptualization contributes in a determinant way to effectively design and incorporate systematic methods in scope.

3.1 Uncertainty Analysis

In the context of this specific case study, uncertainty is associated with missing values or values that do not allow the measure of a given variable.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

The *World Values Survey (WVS)* [9], which underpins the case study object of analysis, has been designed according to high-quality standards and, indeed, specific codes are available to flag the different situations that may lead to such computational uncertainty. They include *Don't know* (code = -1), *No answer/refused* (code = -2), *Not applicable* (value code = -3), *Missing* (code = -5), in addition to the value -4 which, at the best of authors understanding, is not associated to any specific meaning. In general terms, positive values are associated with computationally valid answers.

As such a fine-grained classification is not relevant for the conducted study, main statistics (reported in Fig. 2) simply refer to the uncertainty resulting by the combination of all mentioned codes. The considered dataset is composed of 94278 rows, including a 77.23% of complete rows and a 22.77% with at least one negative code (uncertainty). The breackdown by feature in the same figure shows a significantly higher amount of uncertainty for the variable *Q36*.

From a code perspective, to remark that there is no entry with code -3 (associated with *Not applicable*); on the other side, a high number of rows present at least one variable with code -1 (*Don't know*) or -4 code (unknown meaning), 13.1% and 11.8% respectively; finally, a relatively small number of rows (4.1% and 2.6%) is associated with some code -2 or -5.

An overview of uncertainty (scaled in the range 0-3) is proposed in Fig. 3a, while Fig. 3b presents the same view excluding variables with an amount of uncertainty significantly higher than others (*Q36* in this specific case).

In order to provide a consistent overview of the uncertainty in presence of those numerical patters, a semi-quantitative approach is adopted: given a range of values x_i and a range for feature scaling $[u_{min}, u_{max}]$, the higher value of the range (u_{max}) is reserved to values significantly higher than others ($x_k = u_{max}$, $\forall k, i : x_k >> x_i$), while all other values are scaled assuming a range $[u_{min}, u_{max} - 1]$. Such an approximation is acceptable in this specific context as the original numerical patterns are still present in the figure but they are mitigated. In this specific case, the variable *Q36* is associated with 3 and all other features are scaled between 0 and 2 (Fig 4).

3.2 Topic Sensitivity Assessment

This sub-section deals with topic sensitivity and its assessment. Such an assessment is addressed by adopting different methods and assumes a different scope - i.e. generic and contextual.

Human assessment In general terms, the assessment of the sensitivity of a given topic depends on several factors, including, among others, phrasing, context, audience, respondent, as well as cultural, social and political environment. Additionally, it may be intrinsically subjective, if not biased, and hard to generalise.

In this sense, in addition to evidence-based research, a collaborative approach aimed at establishing shared views on the model of shared meaning-making [1]

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

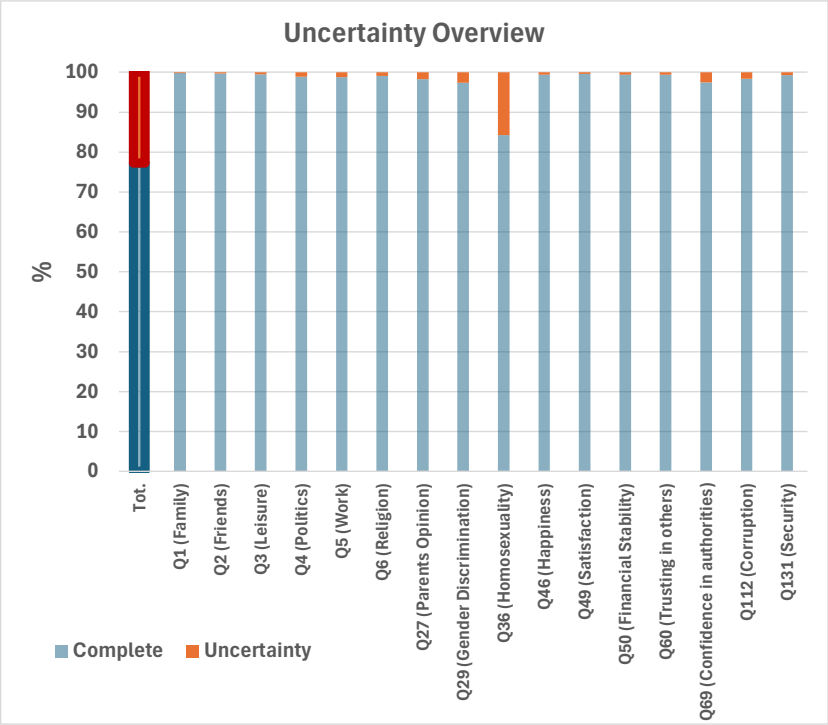


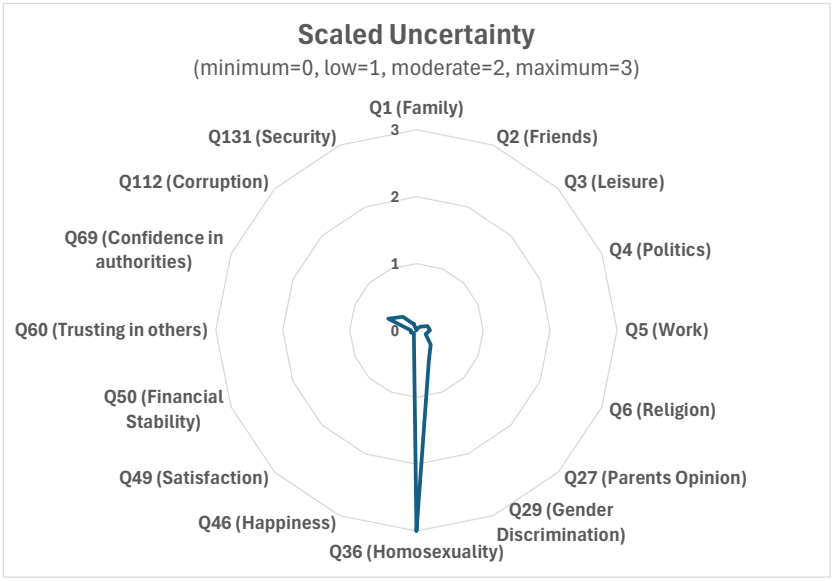
Fig. 2: Uncertainty overview. The first bar (*Tot.*) reports the number of rows with uncertainty; the other bars provides a view by feature.

may positively contribute. However, a collaborative method should be properly designed according to a number of principles to address the inherent sociocultural complexity. It may be hard to enable in practice.

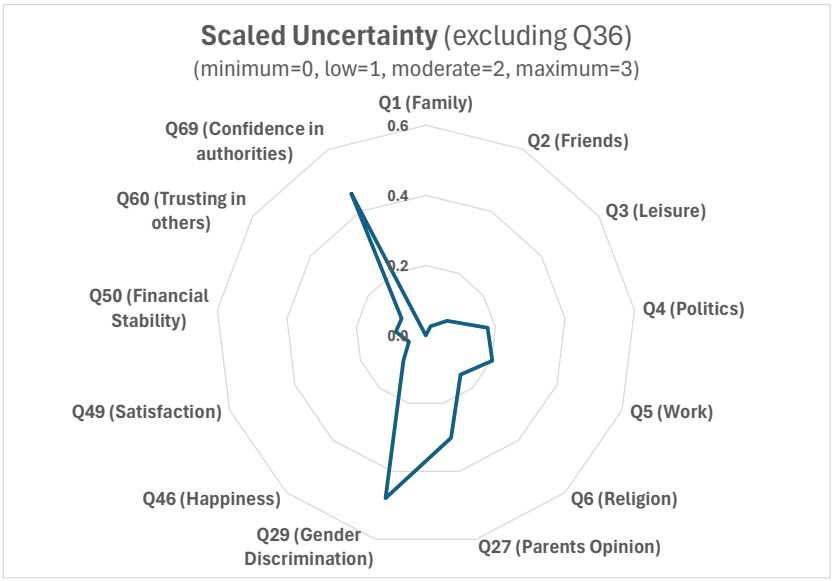
In this work, a human assessment of topic sensitivity is performed according to a simplified approach that relies on common knowledge. Under the assumption that any topic my be potentially sensitive depending on the context, looking at the topics in the case study from a generic perspective, *Politics*, *Religion*, *Gender Discrimination*, *Homosexuality*, *Confidence in authorities* and *Corruption* are definitely critical, with *Work* that may present some intrinsic sensitivity in multiple context. Looking more specifically at the nature of the WVS and the actual questions (contextual assessment), the number of high-sensitive topics is probably lower as questions on politics, work and religion aim to generically measure their relevance as a value for respondents. An overview is reported if Fig. 5a.

The proposed analysis is evidently qualitative and, indeed, tends to be "radical" by focusing on the identification of the most critical topics.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9



(a)



(b)

Fig. 3: Uncertainty overview scaled between 0 and 3. One of the feature (Q36) presents a significantly higher value.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

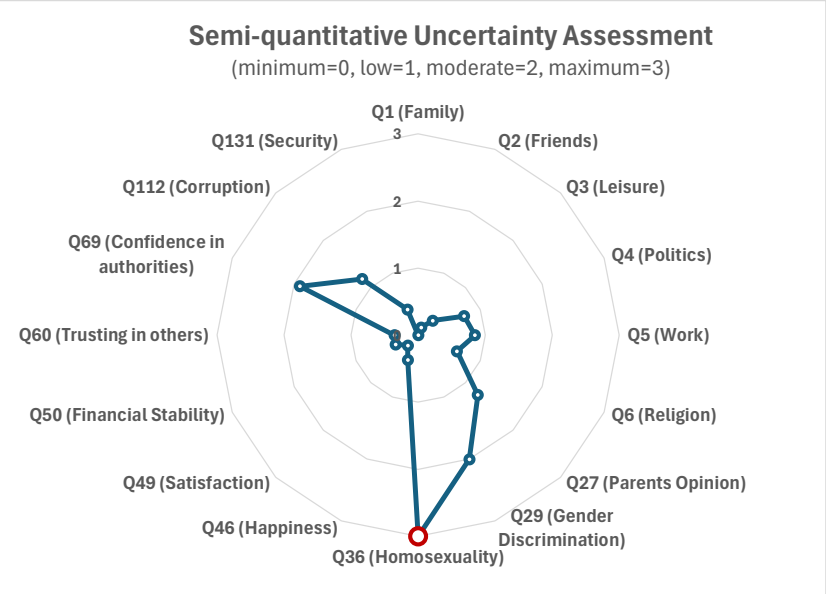


Fig. 4: Semi-quantitative view of uncertainty.

Human-centric approach In a context of huge proliferation of AI technology, in this specific case, a simplified understanding of a human-centric approach [5] assumes a close collaboration between humans and AI to solve a given problem [14]. It may realistically reflect many everyday life situations in the current technological landscape.

More concretely, topic sensitivity has been assessed with the support of WhatsGPT[1] and key human inputs in terms of prompt engineering and refinement of the results.

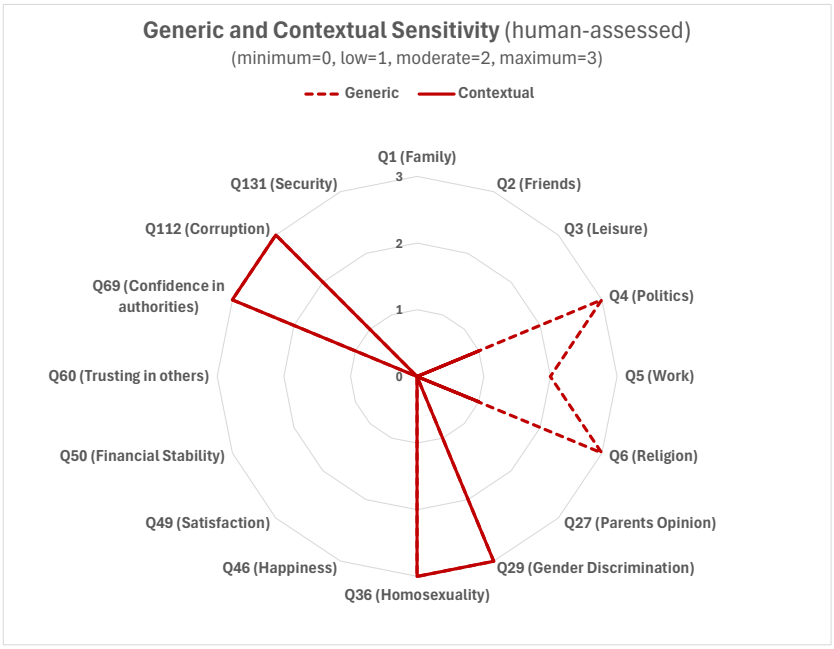
The input for the assessment of the sensitivity of a given topic X is the following query:

```
How sensitive is a question on
X
in a survey
from 0 (minimum) to 3 (maximum)?
```

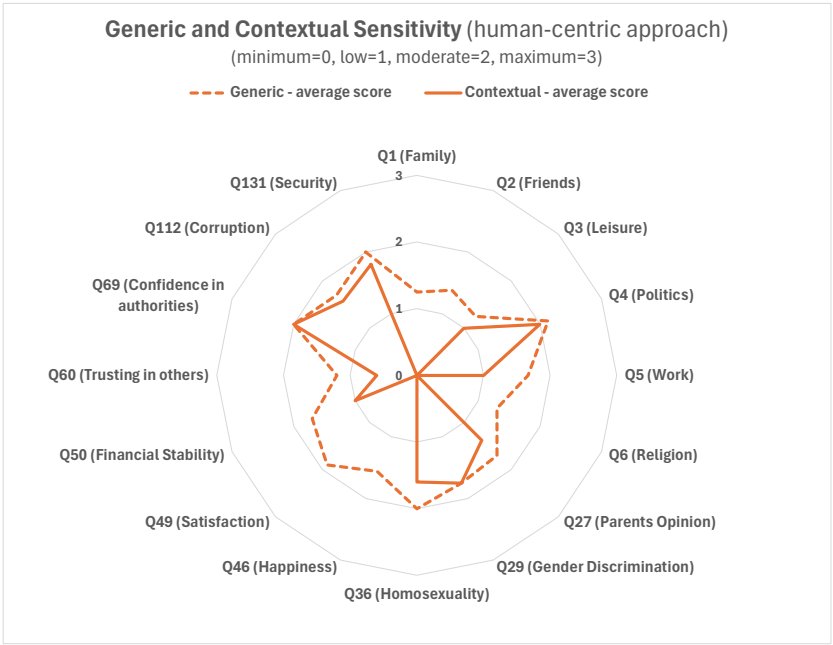
In addition to explanations and other information, the response to such a query includes a variable number of example survey questions on the topic with a related sensitivity score. The average score on the returned examples estimates the generic sensitivity of the topic. The contextual assessment assumes a further level of analysis as a subset of example questions is selected to match the actual case study context. For instance, 5 out of 7 example questions on homosexuality

[1] WhatsGPT - https://www.whatsgpt.me, accessed on 16 January 2025.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9



(a)



(b)

Fig. 5: Sensitivity assessment, including a human assessment and a human-centric approach.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

have been selected looking at the more specific meaning of the question in the dataset, which is on homosexuality acceptance in this specific case.

The human-centric analysis is summarised in Table 1. To note that, due to the generality of the topic and the relatively specific context, a contextual assessment for *Q29* adopting the proposed method was unsuccessful as the example questions are not conceptually aligned with the actual application in the dataset.

A graphical overview is proposed in Fig. 5b. As shown, a human-centric approach reflects a completely different strategy that relies on the identification of a significant range of examples. While this strategy is probably not that effective in its generic version where scores naturally tend to the average, it becomes quantitatively consistent in its contextualised analysis.

	Generic		Contextual		
Variable	**Samples**	**Av.score**	**Samples**	**Av.score**	**Context**
Q1 (Family)	8	1.25	1	0	Value
Q2 (Friends)	8	1.38	1	0	Value
Q3 (Leisure)	8	1.25	1	1	Value
Q4 (Politics)	8	2.13	1	2	Value
Q5 (Work)	9	1.67	1	1	Value
Q6 (Religion)	10	1.3	1	0	Value
Q27 (Parents Opinion)	10	1.7	8	1.38	Value
Q29 (Gender Discrimination)	8	1.75	*0*	*1.75*[a]	Opinion
Q36 (Homosexuality)	7	2	5	1.6	Acceptance
Q46 (Happiness)	9	1.56	2	0	Perception
Q49 (Satisfaction)	10	1.9	1	0	Perception
Q50 (Financial Stability)	10	1.7	2	1	Perception
Q60 (Trusting in others)	10	1.2	5	0.6	Opinion
Q69 (Confidence in authorities)	8	2	1	2	Opinion
Q112 (Corruption)	10	1.7	7	1.57	Perception
Q131 (Security)	10	2	5	1.8	Perception

[a] Same as for generic assessment due to a lack of examples relevant to the specific context.

Table 1: Human-centric assessment.

3.3 From Uncertainty to Semantics

The quantitative estimation of the convergence between uncertainty and sensitivity adopts the *Euclidean Distance*. Assuming two points $p = [p_1, p_2, ..., p_i, ..., p_n]$ and $q = [q_1, q_2, ..., q_i, ..., q_n]$ in the Euclidean n-space, the distance between such points is defined as in Eq. 1.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

Results are reported in Table 2 and point out two clear independent patterns, including (*i*) a significantly lower distance for the contextual approach regardless of the adopted method and (*ii*) human-centric approach slightly more effective to approximate experimental measurements.

A visual representation is provided in Fig. 6 and 7, for the human and the human-centric approach respectively.

Assessment Method	**Assessment Scope**	**Distance**
Human	Generic	4.46
Human	Contextual	2.96
Human-centric	Generic	4.22
Human-centric	Contextual	2.70
Hybrid	Generic	4.82
Hybrid	Contextual	1.94

Table 2: Results summary in terms of Euclidean Distance.

Last but not least, a hybrid approach resulting from the two methods (average values) is proposed. Numerical values are in Table 2 and a visual representation is in Fig. 8. While the generic assessment for the hybrid approach presents the higher distance from experimental data, its fine-tuned version (contextual) outperforms both underlying methods. It provides significant insight for the future evolution of the system and its automation.

4 Current Limitations & Outline of Future Research

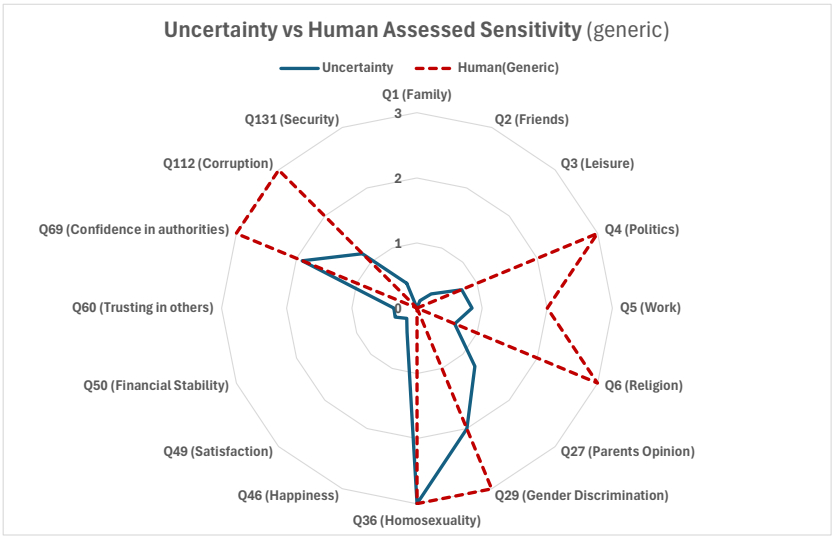
The empirical assessment has provided valuable insight, in line with the predefined research goals. The results achieved establish a grounding framework to enable in fact a systematic evidence-based approach.

However, such a result should be considered looking at current limitations to define a roadmap for future research:

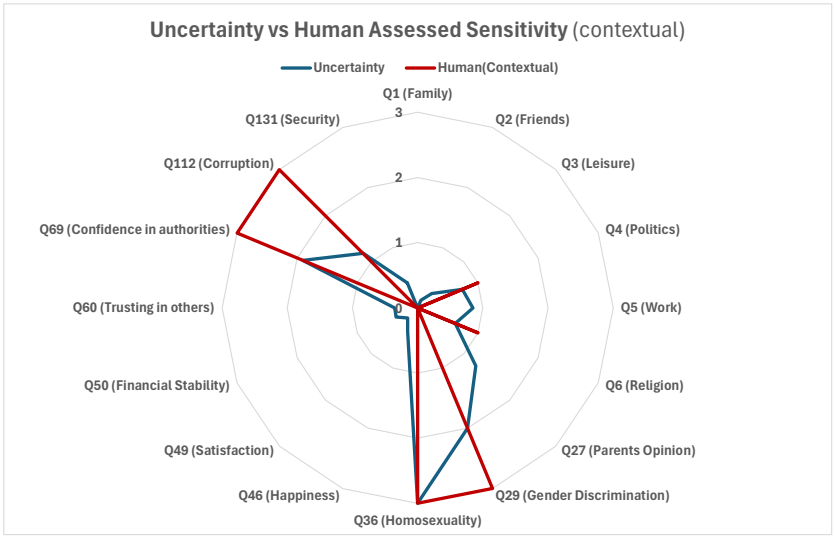
- *Scale*. While the conducted experiment involves a large dataset, the analysis has been conducted at a reduced dimensionality. The current focus on topics over specific question intrinsically enable a scalable framework of analysis.

- *Data source*. The experiment is currently related to one single dataset. Extending the empirical assessment to multiple datasets in a diverse and multidisciplinary context is a key factor to consolidate a systematic and generic approach, as well as to identify peculiarities and possible limitations.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9



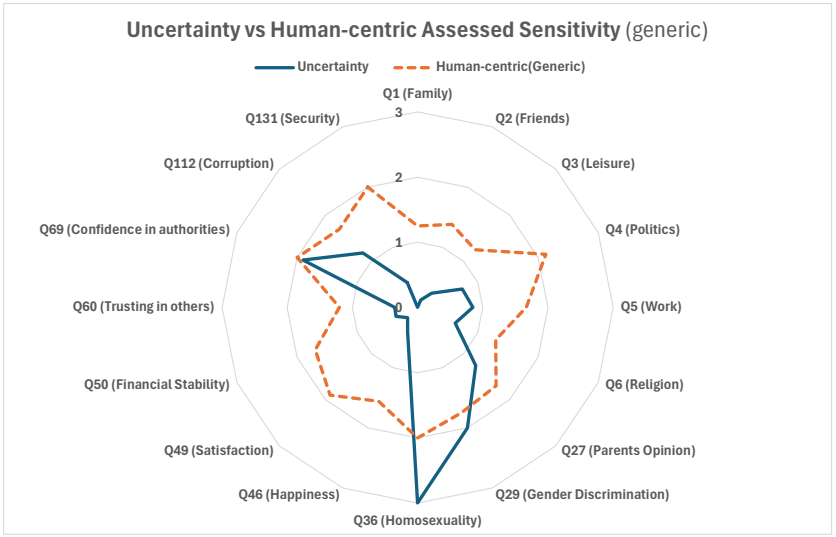
(a)



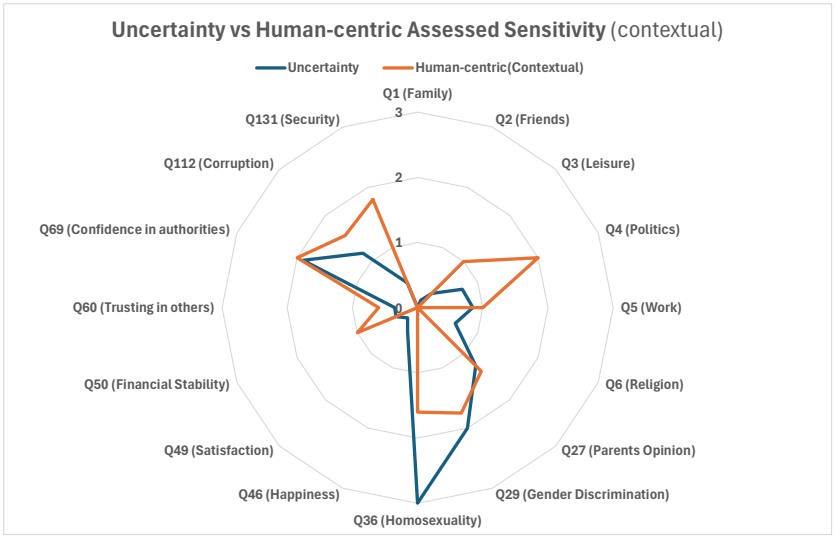
(b)

Fig. 6: Uncertainty and human-assessed sensitivity, including a generic and contextual analysis.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9



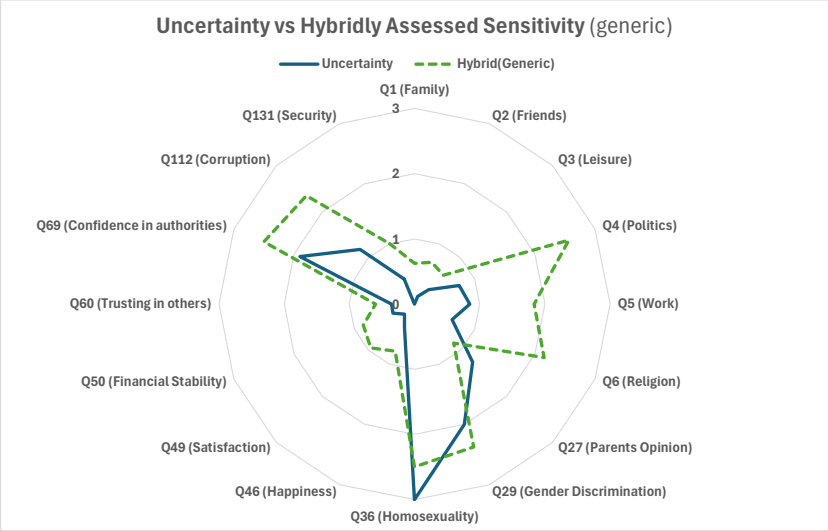
(a)



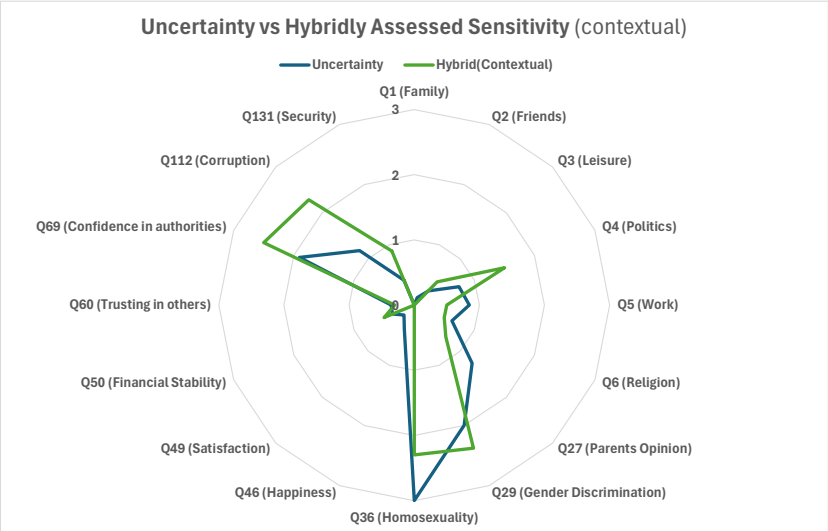
(b)

Fig. 7: Uncertainty and sensitivity (assessed according to a human-centric approach). It includes generic and contextual analysis.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9



(a)



(b)

Fig. 8: Uncertainty and sensitivity assessed according to a hybrid approach.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

- *Bias.* As extensively discussed in the paper, there is an intrinsic risk of bias to assess the sensitivity of a given topic. Such a quantification inherently presents a certain degree of complexity, as it may depend on different contextual factors. The simplified approach adopted in this paper should be further enhanced through more consistent mitigation strategies (for both human and AI bias [15]) based on formal analysis frameworks.

- *Interaction with AI.* The human-centric approach is valuable in the context of this work and, more in general, critical in the modern technological landscape. In this specific case, a consistent approach requires a more sophisticated interaction model [3] to further enhance the reliability and effectiveness of systematic solutions. Additionally, the stability of the output and the impact of different tools as a function of the input (e.g. Prompt Engineering [20]) should be carefully assessed.

- *Assessment metrics.* In generic terms, the assessment metrics are simple to prioritize intelligible analysis and interpretations. While this is an unquestionable advantage, additional metrics are required to support the evolution of the system toward automation.

- *Automated approach.* More in general, the synthesis and the validation of fully-automated solutions require additional experimentation and testing, in line with the previous discussion points.

5 Conclusions

This empirical work has provided valuable insight to assess the relationship between uncertainty and topic sensitivity in survey data, showing a relatively clear convergence. Although the current experimentation (limited to one single case study) doesn't allow the generalisation of the results, it establishes the fundamentals for a more systematic approach in the context of the current technological landscape, which offers the capabilities to enable human-centric and fully automated solutions. Additionally, the critical analysis looking at current limitations has defined a roadmap to further enhance the proposed method aiming at a broader and more consolidated experimental and validation framework.

References

1. Aldemir, T., Borge, M., Soto, J.: Shared meaning-making in online intergroup discussions around sensitive topics. International Journal of Computer-Supported Collaborative Learning **17**(3), 361–396 (2022)
2. Altig, D., Barrero, J.M., Bloom, N., Davis, S.J., Meyer, B., Parker, N.: Surveying business uncertainty. Journal of Econometrics **231**(1), 282–303 (2022)
3. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al.: Guidelines for human-ai interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems. pp. 1–13 (2019)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9

4. Bossaerts, P., Yadav, N., Murawski, C.: Uncertainty and computational complexity. Philosophical Transactions of the Royal Society B **374**(1766), 20180138 (2019)
5. Bryson, J.J., Theodorou, A.: How society can maintain human-centric artificial intelligence. Human-centered digitalization and services pp. 305–323 (2019)
6. Cascaldi-Garcia, D., Sarisoy, C., Londono, J.M., Sun, B., Datta, D.D., Ferreira, T., Grishchenko, O., Jahan-Parvar, M.R., Loria, F., Ma, S., et al.: What is certain about uncertainty? Journal of Economic Literature **61**(2), 624–654 (2023)
7. Chuang, J., Gupta, S., Manning, C., Heer, J.: Topic model diagnostics: Assessing domain relevance via topical alignment. In: International conference on machine learning. pp. 612–620. PMLR (2013)
8. Hariri, R.H., Fredericks, E.M., Bowers, K.M.: Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big data **6**(1), 1–16 (2019)
9. JD Systems Institute & WVSA: European Values Study and World Values Survey: Joint EVS/WVS 2017-2022 Dataset (Joint EVS/WVS). doi:10.14281/18241.21 (2022), Dataset Version 4.0.0
10. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1. Minneapolis, Minnesota (2019)
11. Lepri, B., Oliver, N., Pentland, A.: Ethical machines: The human-centric use of artificial intelligence. IScience **24**(3) (2021)
12. Li, Y., Chen, J., Feng, L.: Dealing with uncertainty: A survey of theories and practices. IEEE Transactions on Knowledge and Data Engineering **25**(11), 2463–2482 (2012)
13. Pileggi, S.F.: A hybrid approach to analysing large scale surveys: individual values, opinions and perceptions. SN Social Sciences **4**(8), 144 (2024)
14. Pileggi, S.F.: Ontology in hybrid intelligence: A concise literature review. Future Internet **16**(8), 1–19 (2024)
15. Roselli, D., Matthews, J., Talagala, N.: Managing bias in ai. In: Companion proceedings of the 2019 world wide web conference. pp. 539–544 (2019)
16. Rosenbaum, A., Rabenhorst, M.M., Reddy, M.K., Fleming, M.T., Howells, N.L.: A comparison of methods for collecting self-report data on sensitive topics. Violence and Victims **21**(4), 461–471 (2006)
17. Roster, C.A., Albaum, G., Smith, S.M.: Topic sensitivity and internet survey design: A cross-cultural/national study. Journal of Marketing Theory and Practice **22**(1), 91–102 (2014)
18. Tourangeau, R., Groves, R.M., Redline, C.D.: Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error. Public Opinion Quarterly **74**(3), 413–432 (2010)
19. Tourangeau, R., Yan, T.: Sensitive questions in surveys. Psychological bulletin **133**(5), 859 (2007)
20. Velásquez-Henao, J.D., Franco-Cardona, C.J., Cadavid-Higuita, L.: Prompt engineering: a methodology for optimizing interactions with ai-language models in the field of engineering. Dyna **90**(SPE230), 9–17 (2023)
21. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable ai: A brief survey on history, research areas, approaches and challenges. In: Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8. pp. 563–574. Springer (2019)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_9