A Method to Deal with Social Bias and Desirability in Ethical Requirements

Claudia Negri-Ribalta[1]

[1]Centre de Recherche en Informatique, Paris I Panthéon-Sorbonne Université, 90 Rue de Tolbiac, 75013, Paris, France

Abstract

[Context] Ethical requirements are a growing area of importance for software. Yet, when dealing with these types of requirements from users, the subject might not always give an accurate representation of their requirements, due to social factors. **[Question/problem]** Which methods can help mitigate social desirability when dealing with ethical requirements? **[Principal ideas/results]** This article proposes the usage of factorial survey experiments (FSE) to work around social desirability when working with ethical requirements. FSE works with vignettes, which the RE practitioner presents to the subject, and experimentally varies them to comprehend how the subject reacts to different stimuli. It enables quantitative analysis of requirements and their specifications, adding explainability and transparency to the RE process. **[Contribution]** This article describes how to use FSE for ethical requirements, and its advantages. We also give an example of application, for which we share preliminary results. Our work opens the discussion for a possible framework using FSE for ethical requirements.

Keywords

Requirements, Ethics, Methodology

1. Introduction

Discussion on the relationship of technology with ethics isn't new [1]. Recently, there's been growing consideration of the ethical aspects of information systems (IS) as an area of research, ranging from privacy concerns to online gambling [2, 3]. One question is how to include ethical requirements from early stages in the software development [2, 3].

However, discussing ethics in software development isn't simple. As [1] have suggested, engineers seem to think that ethical issues aren't part of their job, and it is mostly the responsibility of regulations, and non-engineers. Also, different stakeholders might have different ethical paradigms, which can give rise to tensions between values [2] and some questions might leads to answers that can have social-desirability bias[1] .

In this paper, we discuss FSE as a method of allowing transparent and accountable system design, particularly when dealing with ethical requirements. It proposes the usage of FSE, a well-adopted tool from sociology and other social science used to study items such as beliefs,

In: J. Fischbach, N. Condori-Fernández, J. Doerr, M. Ruiz, J.-P. Steghöfer, L. Pasquale, A. Zisman, R. Guizzardi, J. Horkoff, A. Perini, A. Susi, M. Daneva, A. Herrmann, K. Schneider, P. Mennig, F. Dalpiaz, D. Dell'Anna, S. Kopczyńska, L. Montgomery, A. G. Darby, and P. Sawyer (eds.): Joint Proceedings of REFSQ-2022 Workshops, Doctoral Symposium, and Poster & Tools Track, Birmingham, UK, 21-03-2022, published at http://ceur-ws.org

claudia-sofia.negri-ribalta@uni-paris1.fr (C. Negri-Ribalta)

[1]Social desirability bias occurs when a subject feels (and consequently alters their response) that a specific answer is expected or more socially accepted than the others [4].

intentions, perceptions, or elements that can have socially desirable answers. The experimental nature of FSE allows the RE practitioner to statistically analyze how different specifications interact with each requirement, without necessarily developing the software. This type of analysis can help the RE practitioner explaining a system design, in a transparent and accountable manner, while also keeping trace of the decision of why such design was chosen.

The article is divided as follows: section *Related work* reviews previous work, section *Research Method* introduces the reader to FSE and explains how FSE can be used for RE. In section *Use case*, we share a practical example of FSE on studying trust ($TR), open-source ($OS), and data protection ($DP) requirements in COVID-19 contact tracing apps. The papers shares future work and conclusion in *Future work and conclusion.*

2. Related work

There is growing literature that highlights the importance to include ethical concerns into software design, from early stages [2, 5, 1, 3]. According to [1], ethical requirements, or ethical designing, can be understood as:

- the process by which an IS is developed by taking into consideration elements that can be related to moral norms/values; OR
- when IS have consequences beyond the group that developed it; OR
- it is "related to visions about how to live the ethically good life and virtues involved in that" [1]

[1] discussed how goal-based RE can be insufficient for ethical requirements, and recognizes that given that not all requirements can be met, trade-offs have to be done, which is an ethical choice. [2] presents how Value Sensitive Design (VSD) isn't a well-integrated concept in the RE community, compared to HCI, and proposes a methodology based on VSD to elicit ethical requirements. [5] propose an iterative framework on ethical requirements The authors of [3] study the relationship between ethical requirements and addictive technology and highlight that future work should focus on all the aspects of the RE process and ethical requirements.

This article differentiates from [2], as FSE experimentally varies the vignettes and also assumes that the ethical requirements have been discovered. FSE could help deal with the biases and differences of perception of different stakeholder groups, presented in [5]. In comparison to [3], this article doesn't present a list of categories of ethical requirements, not works with a specific technology. Finally [1] reflects and discusses extensively the relationship technical and ethical requirements, raising research question, but doesn't propose a clear methodology or framework for the topic.

3. Research methodology for ethical requirements

FSE is a research method that is useful to investigate on beliefs, intentions, attitudes, and subjects that have social desirability, among others [6, 7]. It has been used in other areas of research such as sociology, and psychology, among others. In requirement engineering, it has been used by [8, 9].

In brief, this research method prompts the subject to evaluate a set of vignettes that describe a situation around the defined elements of interest. It asks the subject to rate different experimentally varied vignettes, therefore behaving like an experiment (thus having internal validity). By not asking directly to the subject to answer about a subject that might have a social-desirable bias, but rather rating different vignettes (whose factors might have social-desirable bias) it is possible to statistically analyze how the factors behave and, aif there is significant difference between levels. For example, [10] shows how the relationship between sex, education, occupation, citizenship and other variables affect the perception of fair income, through the usage of FSE.

3.1. Factorial Survey Experiments

The first step in FSE is to specify which are the elements of interest of the research - such as privacy - which are labeled as factors by the FSE. For RE, this could be the requirements(s) of the system. The factors should have different levels, that is to say, they have different values or specifications. For example, a researcher might be interested in seeing how the requirements of data protection and security relate to each other, thus investigating which specification fits better for the intentions of a system. It could help with the prioritization of requirements and which specifications suits better for the stakeholders.

Once the researcher defines the factors and their levels they must construct the vignettes. The vignettes describe a situation, either real of hypothetical. There is no one way on how to build vignettes, as these can be images, text, or video. Each vignette uses a specific and unique combination of factors. The size of the vignette universe will depend on the number of factors and their levels. For example, if there are 5 factors with 3 levels, and 1 factor with 4 levels - known as a $3^5 4^1$ - the universe is 972 vignettes. However, it is unreasonable to present this amount of vignettes to a subject to rate them, and thus a sample of vignettes should be selected.

There are different way for selecting a sample of vignettes such as random, efficient-design, blocking, co-founding variables [11, 7, 6]. [7, 11] indicate that random sampling of vignettes has several disadvantages from a statistical point of view, particularly they lose power of explanation. Thus the literature suggests the usage of efficient design (such as D-efficiency) when doing a fractional factorial survey, even when the perfect orthogonality objective is relaxed [7, 11, 10]. This type of design can be provided by specialized tools, such as R or STATA. In R, there are specialized packages such as AlgDesign that have specialized functions for selecting the vignette sample. It is also possible to present one vignette per subject, however this type of approach can affect the statistical power of the model as it is no longer an experimental survey [7].

The survey can be shared through different means, such as in-person or online. The decision of which strategy to follow will depend on a case-by-case. The data can be analyzed using different statistical models, such as mixed multilevel, ANOVA, Tobit models, etc [7, 10]. It is important to note that if a subject rates several vignettes, the data isn't independent and the subject's ID adds error to the model [7].

The data analysis can happen at different levels: the vignette (Level 1 - L1) and respondent-specific (Level 2 - L2). L1 allows to identify how the subject reacted to the different stimuli; which is known as within-subject variables. For example, it is possible to analyze how the specifications affect a specific requirement. L2 allows to analyze how the different interest groups or control variables (such as gender or age) rated the vignettes, and analyze how the

groups differentiate; this is the between-subject variable. Different statistical software offers options to carry out different types of statistical analysis and models for this type of data. For example, the R software provides the "Ordinal" package, that offers the clmm (Cummulative Link Models) function that fits a mixed multilevel, which can be used to analyze the data.

3.2. FSE and ethical requirements - How to use them

FSE seems to be particularly interesting for ethical requirements. As stated in section 3.1, FSE has a longstanding tradition of being used on topics that can have a socially desirable answer, attitudes, or judgments.

A similar case can be built for requirements, particularly ethical. In RE, FSE has been used to study the relationship between security requirements and the perception of risk on users [9]. From sociology, [10] shares that although most subjects would have answered that they believe in equality of gender, the data obtained by FSE shows the contrary. Similarly, software development could take a similar stance when dealing with topics or concepts such as gender issues, privacy, fairness.

By describing a set of realistic scenarios and experimentally varying them, the RE practitioner can analyze the subjects' attitude or reactions to the stimuli [6]. As a consequence of the experimental variation of the factors, FSE reflects the subjects' reaction to different specifications (the stimuli) without necessarily developing the whole system. Furthermore, given its experimental and survey characteristics (if correctly designed, the internal and external validity) the RE process can be replicated. This method may allow the RE practitioner to explain the prioritization of a requirement, beyond ethical topics. As such, FSE gives accountability to the RE process, as it helps the RE practitioner explain the rationale of a design.

4. Use Case - Using FSE for COVID-19 contact tracing apps

We enquired about the willingness, from a user's perspective, to download a COVID-19 contact tracing app, using FSE research method. The requirements were chosen using review of the literature available from September 2020 to May 2021. Parameters, inspired from literature review, were as follows.

- App provider: Government, private company, university and any combination of these (7 possible values) ($TR)
- Data Protection: high, basic, low ($DP)
- Open Source: open source, proprietary code ($OS)

Given the universe of vignettes (42) and taking in consideration fatigue effects from subjects, the universe was divided in decks divided BY the $OS factors. In other words, subjects would receive a sample (or treatment) containing either open-source or proprietary code vignettes. Therefore $TR and $DP are within-subject variables, while $OS is a between-subject variable.

There were 2 decks of 21 vignettes, which can still be considered a big sample. Thus, we modified the graphical interface and the deck was presented as 4 vignettes, each being divided into 7 sub-sections, following a "table-like feeling". This was based on feedback received from

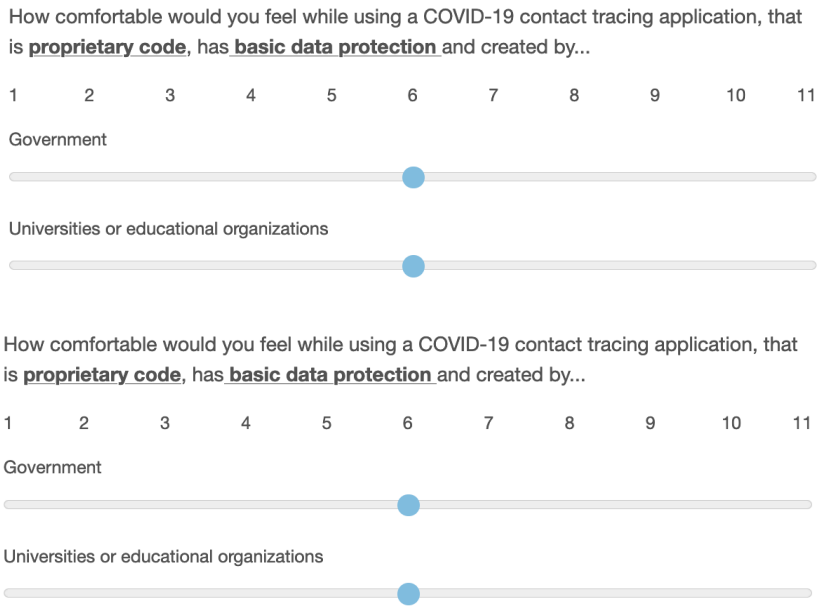

Figure 1: Screenshot of two vignettes presented in the Qualtrics Platform for the COVID-19 FSE. For the sake of conciseness, the screenshots only show the first two choices of each vignette

the testing phase. It used an 11 point Likert scale, following the advice to avoid censored responses [7].

The survey was tested with 80 participants, who answered that they filled out the survey between 4-6 minutes and weren't fatigued. Other feedback received was about wording, the graphical interface as previously stated. Comments were integrated into the final survey version.

The population of interest were french university students. Given that the data gathering phase occurred while national lockdowns were in place, data was gathered from students living in the Paris area. The FSE was answered by 434 subject, and after cleaning the data from rushers, subjects that failed the attention test and those not living in Paris, the final dataset has 415 answers from different subjects.

Our preliminary statistical analysis shows that it is possible to see relationships and differences between each specification and a requirement. In other words, given the experimental variation of the vignettes, it is possible to see how each specification affects each requirement, either negatively or positively. In this use case, the different levels $TR and $DP appear to be significant at the moment of the willingness to download for subjects. In contrast, the levels of $OS don't seem to have a relationship with the willingness to download. The analysis is done in a reproducible and falsifiable way, allowing other RE practitioners to check if they arrive to the same results. These results could help in the development of governance models in COVID-19 contact tracing apps. The results and analysis will be published in future venues, where it will discuss the impact of the details on the design and governance model of COVID-19 contact tracing apps.

5. Future work and conclusion

For future work, a framework for ethical requirements and design, using FSE, can be proposed. Such framework could profit from the usage of chat bots for automatizing the process. Furthermore, the tool could be based on the shiny package from *R*, and help the RE practitioner define the requirement, specifications, optimal design options and analysis. Finally, research could be carried out on the usage of FSE for requirement prioritization, not just for ethical requirements.

This article presents the usage of FSE as a method for RE practitioner to work around the social desirability of ethical requirements, in a transparent, explainable and accountable manner. The RE practitioner must choose the requirements, which describe the system in question. By asking the user to rate the vignettes in an experimentally varied fashion, the RE practitioner gathers data on the users attitude towards different specifications. This can help the RE practitioner for choosing certain design options over others.

Acknowledgments

Thanks to Camille Salinesi, René Noel and Marius Lombard-Platet for their help and comments.

References

[1] I. Van de Poel, Investigating ethical issues in engineering design, Science and engineering ethics 7 (2001).
[2] C. Detweiler, M. Harbers, Value stories: Putting human values into requirements engineering., in: REFSQ Workshops, 2014.
[3] D. Cemiloglu, E. Arden-Close, S. Hodge, T. Kostoulas, R. Ali, M. Catania, Towards ethical requirements for addictive technology: The case of online gambling, in: 2020 1st Workshop on Ethics in Requirements Engineering Research and Practice (REthics), 2020.
[4] P. Grimm, Social desirability bias, Wiley international encyclopedia of marketing (2010).
[5] A. Rashid, K. Moore, C. May-Chahal, R. Chitchyan, Managing emergent ethical concerns for software engineering in society, in: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, 2015.
[6] C. Atzmüller, P. Steiner, Experimental vignette studies in survey research, Methodology: European Journal of Research Methods for The Behavioral and Social Sciences (2010).
[7] K. Auspurg, T. Hinz, Factorial survey experiments, volume 175, Sage Publications, 2014.
[8] J. Bhatia, T. D. Breaux, J. R. Reidenberg, T. B. Norton, A theory of vagueness and privacy risk perception, in: 24th International Requirements Engineering Conference (RE), 2016.
[9] H. Hibshi, T. D. Breaux, S. B. Broomell, Assessment of risk perception in security requirements composition, in: 2015 IEEE 23rd International Requirements Engineering Conference (RE), 2015.
[10] P. Steiner, C. Atzmüller, D. Su, Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap, Journal of Methods and Measurement in the Social Sciences (2016).
[11] T. Baguley, G. Dunham, O. Steer, Statistical modeling of vignette data in psychology (2021).