User Involvement in the Design of ML-Infused Systems

Amela, Karahasanović
SINTEF Digital
Amela@sintef.no

Erik, G., Nilsson
SINTEF Digital
Erik.G.Nilsson@sintef.no

Giorgio, Grani
SINTEF Digital
Giorgio.Grani@sintef.no

Kjell Fredrik, Pettersen
SINTEF Digital
KjellFredrik.Pettersen@sintef.no

Leo, Karabeg
SINTEF Digital
Leo.Karabeg@sintef.no

Patrick, Schittekat
SINTEF Digital
Patrick.Schittekat@sintef.no

ABSTRACT

Advances in machine learning (ML) open up possibilities for better supporting the decision making that occurs in high-stakes domains such as air traffic management (ATM). The success of such decision-making systems highly depends upon end users' involvement in their development process. However, most designers face challenges with finding appropriate ways of doing this. This paper presents our ongoing work to investigate design practices by reporting lessons learned from user involvement in the development of an ML-infused ATM decision support system. To explore if and how UX design methods need to be refined when working with ML as a design material, we conducted an online study with domain experts consisting of three iterations. The paper reports the main challenges we faced and our actions to overcome them. Our results can be useful to other designers working with ML-infused systems.

CCS CONCEPTS

• **Human-centered computing, Human computer interaction (HCI), Machine learning**;

KEYWORDS

Design, ML systems, User involvement

ACM Reference Format:
Amela, Karahasanović, Erik, G., Nilsson, Giorgio, Grani, Kjell Fredrik, Pettersen, Leo, Karabeg, and Patrick, Schittekat. 2021. User Involvement in the Design of ML-Infused Systems. In *CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter (CHI Greece 2021), November 25–27, 2021, Online (Athens, Greece), Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3489410.3489421

1 INTRODUCTION

The design of artificial intelligence (AI) infused systems in general and ML-infused systems in particular is receiving increased attention from the Human Computer Interaction (HCI) community [3, 4, 14, 25]. The design challenges range from trust and explainability to ethical issues. They include challenges with understanding AI capabilities and collaborating with AI engineers throughout the design process [18]. AI has thus been acknowledged as a new design material [14] that requires appropriate design methods, tools and processes.

End users' involvement in development of decision-support systems in high-stake domains is a prerequisite of their success. How to involve users in the development of interactive systems has been well understood [26]. However, there is lack of knowledge on *how to involve end users in the development of ML-infused algorithms.*

This paper presents our ongoing investigation of design practices when working with an ML-infused ATM decision support system. We extended a standard human-centred design process for interactive systems, as described by ISO9241-210 [12], to satisfy the specific needs of ML-infused systems. As the results of such systems depend on the data, their mathematical logic and the ways people interact with the decisions and suggestions of the systems [13], we explicitly added evaluation and development of the ML algorithm and their outcomes, as well as the underlying data in the human-centred design process. To support these evaluation and development activities we developed a prototype (DAC-P) allowing the end users to evaluate the outcome of the algorithms.

To explore usefulness of such process we conducted a study consisting of three iterations, each with three domain experts (experienced air traffic supervisors and flow managers). All sessions were conducted as Microsoft Teams meetings and recorded. During these sessions, the experts were asked to judge the outcomes of different algorithms that were presented to them through a prototype specially developed for this purpose. We analysed the recordings, notes and chats after each iteration to improve both the process and the prototype and to tailor them to the main purpose in this development phase: training the algorithm.

The rest of the paper is organised as follows. Section 2 describes the research context and Section 3 describes our research method. Section 4 presents our main findings. Section 5 concludes and presents future work.

2 RESEARCH CONTEXT

We selected an ATM decision support system, called DAC-FLOW, for our study of end users' involvement in the development of ML-infused systems in high-stakes domains. En-route Air Traffic Controllers (ATCOs) are responsible for the safe and effective guidance of aircrafts. They control flights in their assigned airspace volume, known as a sector. Airspace is divided into several non-overlapping sectors, each controlled by an ATCO. While airspace configurations are traditionally static, the dynamic airspace configuration (DAC) concept, developed in Single European Sky ATM

This is the author accepted version of an article published in
CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter
http://dx.doi.org/10.1145/3489410.3489421

Research (SESAR) Solution PJ.08-01, dynamically adapts the sectorisation to changes in traffic patterns and the amount of traffic, improving thus efficiency [2, 6, 10].

The dynamic airspace configuration problem (DAC) has gained an increased attention in recent years, although still being a fairly unexploited. We may found a basic initial formalization of the problem in [17], where they studied several issues related to the dynamic environment and the reasons behind the need for such a system. From a computational point of view, most of the work has been done by applying classical optimization techniques. For instance, the work in [1] presents a procedure based on fixed posting areas to transition from a configuration to the other. In [24], a more traditional combinatorial approach is proposed, by seeing DAC as a weighted-graph model, allowing for graph-based algorithms. The challenging nature of DAC makes it suitable also for genetic methods as in [16]. The inclusion of machine learning in airspace usage has been tackled in [8], but despite this, only a limited number of papers can be found on the subject. For instance, in [21] there is a tentative to find a solution through approximate dynamic programming, then extended in [20] to include learning agents.

Without the aim of completeness, we suggest [23] and [15] as reviews covering optimization techniques adopted to solve DAC.

DAC-FLOW is a decision support system under development that supports flow managers and supervisors when selecting airspace sectorisation that will be used by ATCOs. DAC-FLOW was selected because of its complexity and the highly specialised expertise of its end users. The role of ML and how it works in DAC-FLOW is much more difficult to comprehend than, for example, the use of ML for image recognition. Recognising an image presenting a cat, explaining how it works, and explaining problems that might occur seems to be relatively well understood, at least at a high level. Identifying and explaining a good workable sectorisation for a given air traffic is far more complex. We therefore expected that exploration of design practices with such a system would provide useful learning.

DAC-P is a web-based prototype, developed with the goal of visualising results from our ML-based dynamic airspace configuration algorithm (DACA) and enabling the participants to provide feedback informing the further development of this algorithm. In ongoing R&D work in SESAR Solution PJ.09 [7], our DAC algorithm will be an important part of a system supporting the interplay between flow managers, supervisors and ATCOs in a realistic context.

DACA is a two-phase algorithm that dynamically generates sectors for air traffic control. The core of the procedure is a Deep Neural Network [11] that estimates the workload function. Several definitions of workload can be applied, and organisations around the world, being affected by both regulatory rules and the software they adopted, tend to use different formulas.

The first phase for DACA is training the neural architecture, in which the dataset, composed of tuples of trajectories, sectors and computed workloads, is obtained by simulation. The training phase is performed offline with respect to the dynamic computations, but the model can be updated asynchronously. In the second phase, the trained model becomes part of the objective function of the optimisation engine. Particularly, the final objective is to optimise the workload balance among the sectors, penalised quadratically. The optimisation is performed using standard first-order methods

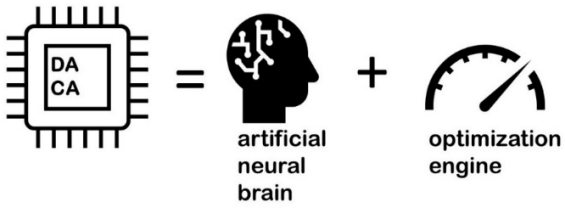

Figure 1: Components of the DAC Algorithm

[5], with conditions guaranteeing the improvement of the objective function at each step. Finally, since the procedure is dynamic, updated trajectories and the older configuration are taken into account at each subsequent optimisation, both to facilitate real-time traffic response and smoothness.

3 METHOD

To explore design practices, we conducted an online study consisting of three iterations. Each iteration was conducted with three domain experts (the intended end users of the system under development), one facilitator and several observers (four to six developers of the prototype and developers of the ML algorithm). The facilitator introduced and led the experiment, assisted the participants and collected scores and other feedback from the participants during and after each scenario and session. Iterations were organised as Microsoft Teams meetings and were recorded. One of the observers (an HCI expert) took notes during the sessions, including the scores and comments. In addition, all observers took their own notes and conducted internal meetings for observers only that followed each session with the participants, during which they summarised the notes in a common document. Some points of improvement regarding the design process were identified after the first and second iterations and followed by appropriate action.

After the third iteration, the overall approach and its usefulness were discussed by the observers and the facilitator. The collected data (observer notes, recorded sessions and chats) were processed by a quick content analysis by the first author of the paper. Open coding was applied. The findings were then discussed and complemented by the other observers. The analysis focused on process deviations, misunderstandings and needed clarifications, the maturity of the prototype and the usefulness of the collected feedback.

The participants' feedback on the quality of the presented sectorisations (the outcome of the ML-infused algorithm) and the feedback on the functionality of the prototype from the recorded sessions, chat and the observers' notes were analysed by the developers and HCI experts and used in further development. The details of the study and the prototype we used are given in the following subsections.

3.1 Participants

There were three participants in the experiment (one male and two females). They are all experienced supervisors and/or flow managers working for a national air navigation service provider in Europe (ENAV, Italy). Recruitment of the participants was done internally by ENAV. As the design solutions evaluated by the experts

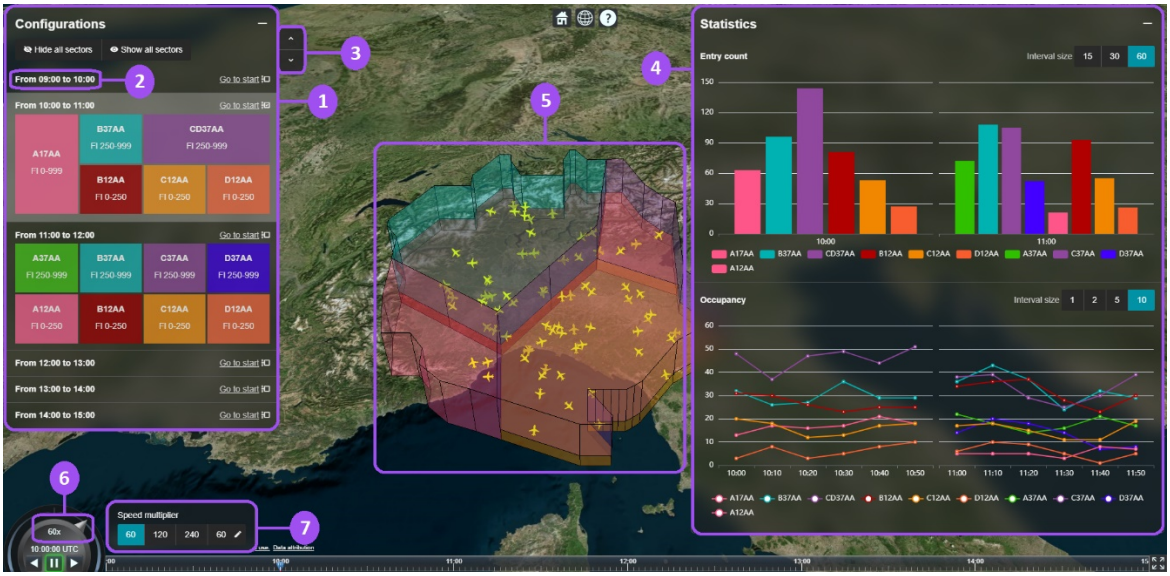

Figure 2: DAC-P – the main functionality

were the outcomes of an ML-infused algorithm (the DAC algorithm) and not the user interface of DAC-FLOW, the same experts were used in all three iterations.

3.2 Study design

The participants, the set-up, protocol, air traffic and data collection methods used were the same or very similar in all three iterations. The three iterations differed according to (1) the DAC algorithm, (2) the web-based user interface of DAC-P, and (3) the scenarios used. In the study, a scenario was a sequence (between two and six) of sector configurations, which were evaluated together. The DAC algorithm gradually evolved and was enhanced through the three iterations. Its development was based on both feedback from the domain experts given in the previous iterations and the ML algorithms' ability to learn over time.

DAC-P's web-based user interface evolved quite extensively through the iterations. It was based mainly on feedback from the participants and also the scenarios used. In the first two iterations, the participants were asked to go through up to ten scenarios, each spanning one hour and involving one sectorisation change, with each sector configuration lasting 30 minutes. In the third iteration, the participants were asked to go through two scenarios, each spanning six hours and involving five sectorisation changes, with each sector configuration lasting for one hour. The longer duration of the scenarios caused a major redesign of the sector selection panel (Figure 2, functionality 1), while the longer duration of each configuration, combined with feedback from the participants, caused a major redesign of the statistics panel (Figure 2, functionality 4).
In each iteration, there was one common training session and one experiment session per participant. The training sessions were conducted a few days before the experiment sessions. In the period between the training session and the experiment sessions, the participants could use the DAC-P prototype for individual training on test scenarios. The duration of one experiment session was up to three hours.

3.3 Prototype developed for the study – DAC-P

To evaluate the outcomes of the ML-infused algorithm proposed sectorisations, the experts needed a working prototype that helped them evaluate and compare the basic features of different sectorisations. As the focus of our development work was the algorithm and not the interface, we wanted to limit the functionality of the prototype. When fully developed and evaluated, the algorithm will be integrated into a commercial tool developed by another company. The figure above visualizes the basic functionality of the prototype, as presented in the tutorial we prepared for the study participants.

The main window provided the following functionality to the participants:

1. Sector selection panel – enables the users to select the airspace sectors

2. Configuration time period – shows the time period for each airspace sector configuration

3. Configuration scrolling – presents the configuration that occurs before or after the ones currently shown

4. Statistics panel –shows the basic statistics (entry count and occupancy) for the sectors and the configurations that are currently selected

5. Sector visualisation – the area where the currently selected sectors, with their traffic, are displayed

6. and 7. Visualisation speed – enables the user to change the speed by the speed control (6) or the speed multiplier (7)

It was not a goal of the study to evaluate the usability of the prototype with the functionality just outlined. DAC-P was a means of visualising the results of the DAC algorithm for different scenarios. By presenting such results for a number of scenarios, DAC-P was also a means of illustrating the algorithm.

4 LESSONS LEARNED

The feedback provided by the domain experts on the outcomes of the ML algorithms was valuable to the further development of the algorithms. From workshop sessions with end users and a literature review, a set of objectives were formulated that would allow the algorithm to distinguish between good and bad sector configurations and sector configuration changes. By conducting end-user sessions, we were able to:

1. investigate to determine if objectives were missing
2. elicit the relative importance of each objective
3. estimate the extent to which a new sector configuration needed to improve to warrant the extra workload of a sector configuration change.

The end users evaluated the sector shapes as good and workable. Hence, the objectives guided the algorithm to well-shaped sectors. However, air traffic was sometimes too close to the sector borders, leading to a higher workload. ATCOs needed to interact more with the adjacent sectors to appropriately monitor and adjust flights. From previous workshops and literature review, the "closeness of traffic to sector borders" was not identified as an important enough objective. The end-user sessions revealed this objective: all experts gave the same feedback in scenarios with the traffic close to the boarders.

The end users gave a score for each sector configuration and configuration change. Each configuration and change were produced with different weights attached to each objective. The weights represent the relative importance of the objective. A high score indicates that a combination of weights is preferred by the end user. A lower score is a less preferred weight combination. Surprisingly, the weight preferences of each end user varied widely for some sector configurations and traffic. In future development, it would be helpful to show the results of these sector configurations to all end users in the group to discuss why they differ.

The main findings regarding the design process and end-user involvement in the development of ML-infused tools can be summarised as follows.

The importance of explaining the role of ML and how it works. In the first iteration, we gave the participant a very brief explanation of the role of ML in the algorithm. Their feedback and the questions they asked helped us realise that this needed to be explained better. A presentation explaining the architecture of the algorithm and the role of ML was included in the training sessions in the second and third iterations. We noted that the participants were used to evaluating and working with rule-based systems; thus, it was not easy for them to change their mindsets. Examples should be developed that pedagogically explain ML and how it works in domains other than image recognition.

The clear communication of the goal of user involvement. Although we communicated to the participants that the object of the evaluation was the algorithm and not the prototype, a significant amount of the received feedback was related to the user interface of DAC-P. On several occasions, we gently pushed the participants back on track when they started explaining the details of an optimal tool from their perspective. As we had enough time, we did not interrupt their explanations, but rather reminded them that we needed a score for the outcome of the algorithm. Such feedback was, however, effective, both for identifying the appropriate maturity level of the prototype needed to evaluate the algorithm and for further developing the tool. When planning the evaluation sessions, one should set aside enough time to allow the participants to reflect on issues other than those that are the main objective of the study.

Difficulties to find the appropriate maturity level of the prototype. Although the end users were involved in the specification of the prototype's functionality, the prototype was extended in two iterations, based on the feedback of the participants. On one hand, the prototype should be as simple as possible at this early stage of development; on the other hand, it should support the participants when they make judgements about the quality of the solutions presented by ML. In our case – a dynamic airspace configuration and presentation of sector shapes – we needed to present some additional information and allow navigation between sectors in a user-friendly way. To alleviate this problem, one should apply an iterative development approach and plan enough resources for several iterations.

5 CONCLUDING REMARKS AND FUTURE WORK

Advances in machine learning (ML) open a whole new spectrum of possibilities for decision making in high-stake domains [19] but also pose new challenges to designers of such systems. This paper presented our ongoing investigation on user-involvement in the design of ML-infused algorithms for an ATM decision support system. We extended a standard human-centred design process for interactive systems by activities explicitly addressing evaluation and development of the ML-infused algorithm. We also developed a prototype specially tailored for the evaluation of the algorithm. The process and the prototype enabled the domain experts to provide valuable feedback to the development of the algorithms.

Working with ML as the design material in the ATM domain has not been an easy task. Although, one cannot generalise based on the results of one case, we hope that our experience and lessons learned might help both researchers and designers working in this area. We plan to continue this research by exploring other ways of user-involvement in the evaluation of ML-infused algorithms. Further, we plan to explore how to better integrate design of the algorithmic part of the system with the design of the overall system. Different ways of explaining the capabilities of the ML-infused algorithms and usefulness of the techniques under development in the emerging field of explainable AI [3, 9, 22] will also be explored.

ACKNOWLEDGMENTS

This project has received funding from the SESAR Joint Undertaking (JU) under grant agreement No 874463. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the SESAR JU members other than the Union. We thank the participants in our study for their valuable contribution.

REFERENCES

[1] Alexander Klein, Mark D. Rodgers, and Hong Kaing, 2008. Dynamic FPAs: A new method for dynamic airspace configuration. In *2008 Integrated Communications, Navigation and Surveillance Conference.* IEEE.

[2] Amela Karahasanovic, Erik G. Nilsson, Patrick Schittekat, Vetle Volden-Freberg, Morten Smedsrud, Patrizia Criscuolo, and Giuseppe Esposito, 2019. Supporting

Air Traffic Controllers During Sector Configuration Changes in Dynamic Air Space Configuration. In *SESAR Innovation Days. ISSN 0770-1268.* SJU.

[3] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli, 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada2018), Association for Computing Machinery, New York, NY, USA, Paper 582. DOI= https://doi.org/10.1145/3173574.3174156.

[4] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholoas Diakopoulos, 2016. Grand challenges for HCI researchers. *interactions 23*, 5 (September + October 2016), 24-25. DOI= https://doi.org/10.1145/2977645.

[5] Dimitri P. Bertsekas, 1997. Nonlinear programming. *Journal of the Operational Research Society 48*, 3, 334-334.

[6] E-OCVM, 2010. European Operational Concept Validation Methodology - 3.0

[7] Elodie Bastie *et al.* 2021. *SESAR Solution PJ.09-W2-44 SPR-INTEROP/OSED.* EUROCONTROL.

[8] Emre Osman Birdal and Serdar Üzümcü, 2019. Usage of Machine Learning Algorithms in Flexible Use of Airspace Concept. In *2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC).* IEEE.

[9] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu, 2019. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science*, Tang J., Kan M.Y., Zhao D., Li S. and Z. H. Eds. Springer, Cham. DOI= https://doi.org/10.1007/978-3-030-32236-6_51.

[10] Hassini Hind, Amina El Omri, Noreddine Abghour, Khalid Moussaid, and Mohamed Rida, 2018. Dynamic airspace configuration: Review and open research issues. In *2018 4th International Conference on Logistics Operations Management (GOL)*, Le Havre, 1-7. DOI= http://dx.doi.org/10.1109/GOL.2018.8378093.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, 2016. *Deep learning.* MIT press.

[12] ISO, 2010. ISO 9241-210 Ergonomics of human–system interaction — Part 210: Human-centred design for interactive systems

[13] Kartik Hosanagar, 2019. A Human's Guide to Machine Intelligence: How Algorithms Are Shaping Our Lives and How We Can Stay in Control. Viking.

[14] Lars Erik Holmquist, 2017. Intelligence on tap: artificial intelligence as a new design material. *interactions 24*, 4 (July-August 2017), 28–33. DOI= https://doi.org/10.1145/3085571.

[15] Manuel Graña, 2019. Dynamic Airspace Configuration: A Short Review of Computational Approach. In *International Conference on Computational Collective Intelligence.* Springer, Cham.

[16] Marina Sergeeva *et al.* 2017. Dynamic airspace configuration by genetic algorithm. *Journal of traffic and transportation engineering (English edition) 4*, 3, 300-314.

[17] Parimal Kopardekar, Karl Bilimoria, and Banavar Sridhar, 2007. Initial concepts for dynamic airspace configuration. In *7th AIAA ATIO Conf, 2nd CEIAT Int'l Conf on Innov and Integr in Aero Sciences, 17th LTA Systems Tech Conf; followed by 2nd TEOS Forum.* AIAA.

[18] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman., 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), Association for Computing Machinery, New York, NY, USA, 1-13. DOI= https://doi.org/10.1145/3313831.3376301.

[19] Randy Goebel *et al.* 2018. Explainable AI: the new 42. In *Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science*, Holzinger A., Kieseberg P., Tjoa A. and Weippl E Eds. Springer, Cham. DOI= https://doi.org/10.1007/978-3-319-99740-7_21.

[20] Sameer, K., Ganesan, R., and Sherry, L., 2012. Dynamic airspace configuration using approximate dynamic programming: intelligence-based paradigm. *Kulkarni, Sameer, Rajesh Ganesan, and Lance Sherry. "Dynamic airspace configuration using approximate dynamic programming: intelligence-based paradigm." Transportation research record 2266.1 2666*, 1, 31-37.

[21] Sameer Kulkarni, Ganesan Rajesh, and Sherry Lance, 2011. Static sectorization approach to dynamic airspace configuration using approximate dynamic programming. In *2011 Integrated Communications, Navigation, and Surveillance Conference Proceedings.* IEEE.

[22] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein, 2019. *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI.* DARPA.

[23] Shannon Zelinski and Chok Fung Lai, 2011. Comparing methods for dynamic airspace configuration. In *2011 IEEE/AIAA 30th Digital Avionics Systems Conference.* IEEE.

[24] Stephane Martinez *et al.* 2007. A weighted-graph approach for dynamic airspace configuration. In *AIAA guidance, navigation and control conference and exhibit.* AIAA.

[25] Umer Farooq and Jonathan Grudin., 2016. Human-Computer Integration. *interactions 32*, 6 (November-December 2016), 26-32. DOI= https://doi.org/10.1145/3001896.

[26] Yvonne Rogers, Helen Sharp, and Jennifer Preece. 2011. *Interaction Design, Beyong human-computer interaction.* Wiley.