Explaining complex machine learning platforms to members of the general public

Rachel Eardley[a], Ewan Soubutts[a] , Amid Ayobi[a], Rachael Gooberman-Hill[a] and Aisling O'Kane[a]

[a] *University of Bristol, Beacon House, Queens Road, Bristol, U.K.*

Abstract

In this workshop paper we present an overview of our research into understanding how to explain complex machine learning (ML) health platforms to members of the general public who might benefit from them, specifically those who have Type 2 Diabetes (T2D). The availability of home health sensor technology is increasing; however, it is unclear how to explain these platforms to potential users so that they can make an 'informed decision' on the adoption of that platform within their home. Through a user-centered-design approach, we have completed a case study with three studies that have (1) Given an overview of a complex ML platform, that of SPHERE; (2) Identified how the participants would like us to explain this content and (3) Created and validated an explanation document that presents, at a high-level the SPHERE platform. We present our findings on the priority of understanding how and why the platform can help them over the technical detail of the platform itself.

Keywords 1

Explanations, Machine Learning, Digital Health, Informed decision, Home health, Complex platforms, Design.

CEUR-WS.org/Vol-2903/IUI21WS-TExSS-6.pdf

1. INTRODUCTION

In many parts of our daily lives, Artificial Intelligence (AI) and Machine Learning (ML) have become ubiquitous in assisting our decision making, e.g., suggesting films to watch on Netflix [1], suggesting purchases online or people to 'follow' on social media. Similar technologies are also increasingly common in specialist areas such as healthcare, in particular clinical support tools [23], used to support clinician and/or patient decision-making about their condition and the risks and benefits of potential treatments. However, when it comes to more critical factors such as our health and wellbeing, many would argue that those who are receiving and those who are providing healthcare, should be made aware of the reasonings behind those decisions [1,7,9,15]. In order to bridge the lack of understanding, we look to Explainable AI (XAI), an area of study that challenges different disciplines ('developers', 'theorists', 'ethicists' etc.) to make transparent the decisions that the AI and ML algorithms make. This is particularly important for those who are receiving and those who are providing healthcare to understand what the system is doing, for example to justify the clinical results given, correct errors, improve medical algorithms or to highlight a new discovery [1,7,15].

In the domain of healthcare, Holzinger et al [2] states that there is a growing need for AI systems that are 'trustworthy, transparent, interpretable and explainable', and there is evidence to benefit the use of clinical AI systems, for instance predicting the risks of

Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA
EMAIL: rachel@racheleardley.net (A. 1); e.soubutts@bristol.ac.uk (A. 2); amid.ayobi@bristol.ac.uk (A. 3); r.gooberman-hill@bristol.ac.uk (A. 4); a.okane@bristol.ac.uk (A. 5)

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)
CEUR Workshop Proceedings (CEUR-WS.org)

hospital readmission for pneumonia patients or spotting bone fractures [6,20]. However, there is also an opportunity for AI to contribute to healthcare outside clinical settings, for instance supporting individuals with chronic illnesses who manage their own conditions at home, a more common trend with today's increasing healthcare costs [4]. Ballegaard et al [2] argues that healthcare is not just about keeping individuals healthy but allowing them to continue to live a sustainable and independent lives. With this in mind, we look to ML/AI platforms such as SPHERE (sensor platform for healthcare in a residential environment) which uses ML to algorithmically interpret data based on the individual's patterns of living at home [22]. How though, do we gain sufficiently informed consent from the public install such complex ML platforms within their homes?

In the medical field, there is a legal and ethical requirement for the patient and clinician to go through a process of 'informed consent' [8,13,17], where the patient presented with the benefits, risks and any alternatives to their treatment makes a decision [3,8]. For ML platforms, there is also an ethical process that includes explaining the benefits, risks, limitations and the data used for potential translation of the ML algorithms [1,14]. To make an 'informed decision' around the adoption of a complex platform, an individual needs to have enough knowledge to think critically about the processes that the platform implements or supports [11,12]. As with informed consent in medical care, for an individual to make an informed decision around the adoption of a complex platform, a process needs to occur that supports the explanation of both the platform's risks and benefits. When and how does this informed decision process occur for home health technology?

To understand how we should explain complex ML/AI platforms to members of the general public, we conducted a case study that focused on the SPHERE platform and members of the general public with Type 2 Diabetes (T2D), where most of the care takes place outside clinical settings [19]. Using a user-centered-design methodology in creating an explanation document to aid informed consent, we gained insight into users' interpretation of the 'informed decision' process of adopting the complex platform within their homes. What we found is that even though the document explained the complex ML/AI platform in a manner that was understandable to our participants and that they could see the SPHERE platforms benefits, they were more focused on the purpose of the technology, questioning why and how the platform could help them as individuals with T2D.

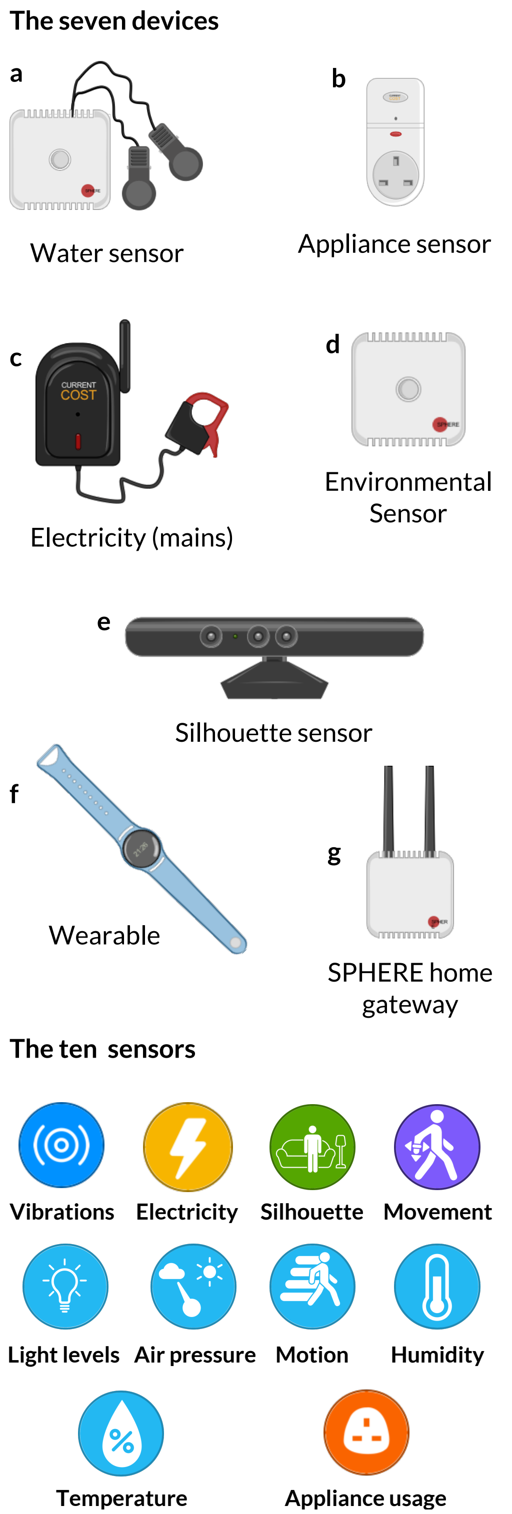

Figure 1: Hardware and networks – the hardware devices of the platform and sensors

2. Defining the Explanation

Using a user-centered-design methodology to define the explanation of the SPHERE platform, we first completed semi-structured interviews with eight members of the SPHERE team who had built and maintained the system. After this, we ran a second study which presented alternative designs about the platform's hardware (figure 2a-c), the ground truthing of the data (figure 2d-f) and the ML process unsupervised learning (figure 2g-i) to nine people with Type 2 diabetes and members of their households who might also have to live with this domestic health technology. From the findings of these two studies, we created an explanation document (figure 4) that presents and explains the SPHERE platform to the general public who had T2D. Finally, we ran a validation study that reviewed how the explanation document was used in an onboarding/set-up session with technicians and how the SPHERE system and the document was interpreted and understood.

2.1. Understanding the platform

Our first challenge was to understand what SPHERE was capable of, its processes, hardware and ML/AI requirements. With this aim in mind, we conducted semi-structured interviews with eight out of eleven of the team members. The team members had been working on the project from two to six years and had mixed roles within SPHERE (2 x Deployment technicians, 3 x ML experts, 1 x Hardware engineer, 1 x Researcher and 1 x Community liaison).

By interviewing these team members with a diverse range of roles within SPHERE, we were able to gain an overview of all aspects of the complex platform. We conducted the interviews individually within a university-based meeting room, audio-recorded and then transcribed verbatim. Using affinity diagramming and a bottom-up approach we created a total of 681 post-it notes (Machine Learning x 245, Research x 63, Community Engagement x 68, Hardware x 100 and Deployment Technician x 205). Once the five job roles (deployment technicians, machine learning, research, hardware and community liaison) had been initially coded into themes, the post-it notes were organized by the first author into 35 further themes that were then broken down into three overarching themes. These overarching themes were (1) Hardware and Network; (2) Installation, Training and Data gathering; (3) Machine learning and Data visualization. We then transferred these themes into a Microsoft Word document. At that stage, the first author merged any duplicated content. We then asked the eight core team members who took part in the interviews to review the document to confirm the draft document was technically correct.

These three overarching themes helped us define the platform, for example, capturing seven sensor devices (Figure 1a-g) and ten individual sensors (Figure 1) with technical and positioning limitations. We also captured the installation process where the deployment technicians will visit a participant's home four times (survey, installation, maintenance and removal) and that the data collected is saved on a hard disk within the participants home and with their permission and processed through supervised and unsupervised machine learning.

2.2. Understanding the interpretations

Once we had gained an understanding of the complex platform, our next challenge was to define how to present the information to our participants. For this study we focused on one area of each of the overarching themes: For Hardware & network we selected the most technically complex sensor, the 'environmental sensor' (figure 2a-c), for Installation, training & data collection, we selected 'ground truth' (figure 2d-f) as this process informs the ML algorithms. For Machine learning & data visualization, we selected 'Unsupervised learning' (figure 2g-i) as this is the more speculative form of ML. Through a design workshop with six participants (three university researchers and three members of a community engagement charity), we focused on the 'environmental sensor' (figure 2a-c) and created three alternative designs that presented the platforms information at different technical levels, detail, approaches to language and visual elements. We then, used these design decisions to create three alternative designs for the further two areas of the platform, 'ground truth' (figure 2d-f) and 'unsupervised learning' (figure 2g-i).

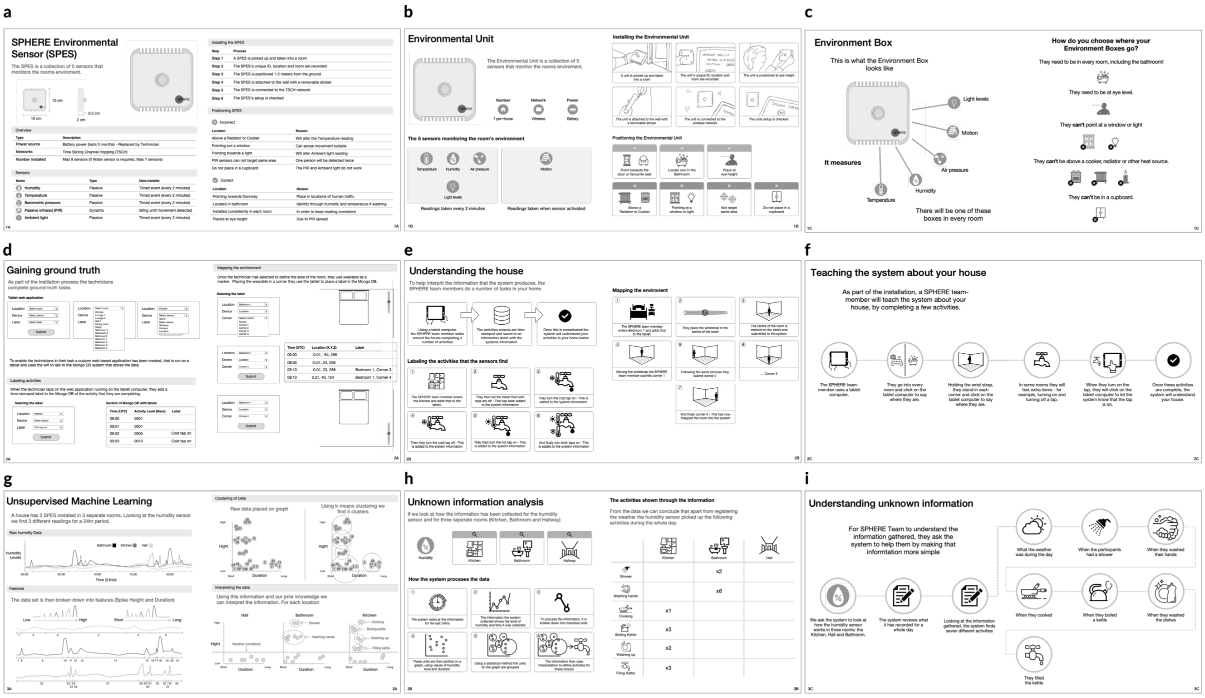

Figure 2: The three alternative designs for the three areas of the SPHERE platform

We presented these nine designed documents (figure 2) to nine participants who either had T2D or lived with someone who did. The nine participants (five female, four male) were aged between 25 to 74, with a varying education level ranging from that of entry-level to PhD. Six participants had T2D, and three participants lived with someone who did. All participants owned a smartphone, four participants had an IoT device such as Amazon Alexa or Google Home. Two participants (AD2 and AD6) had weather stations at home and due to this had prior knowledge of sensors and their capabilities. The Environmental Sensors were presented first with the alternative designs alternated (using the Latin square method), then the Ground Truth and finally Unsupervised Learning.

2.2.1. Overview of findings

For all three areas (environmental sensor, ground truth and unsupervised learning), the participants considered the alternative design with the most technical information and detail to be far too complex, scary or off putting. The participants additionally preferred the language as used in the simpler design alternatives as it used common language an non-technical words.

For the environmental sensor (figure 2a-c), the participants requested that the image of the sensor be the version from figure 2c, with the sensor measurements as in figure 2a in both centimeters and inches. They requested an understanding of where the position of the sensors within the home, however, they did not like the list in figure 2a or the storyboard in figure 2b as they provided unnecessary information (the deployment technician would fit the sensor). They preferred the more structural visual approach to the rules of the sensor placement as in figure 2b and requested more of a description of what each sensor did.

With the ‘ground truth’ (figure 2d-f) the participants considered the simpler version (figure 2f) to be just enough information and were positive with the storyboard flow. The other two alternatives (figure 2d and 2e) were both thought of as too much information and not relevant to the participants as the deployment technician would complete the process.

Finally, for ‘unsupervised learning’ the participants were confused by the charts and graphs considering figure 2i as the better description with a few changes. These changes included the change of an icon so that it fits the descriptive text better and combining the whole of figure 2i with the righthand side of figure h, here showing the participant how the ‘unsupervised machine learning’ works and showing the results in an understandable chart.

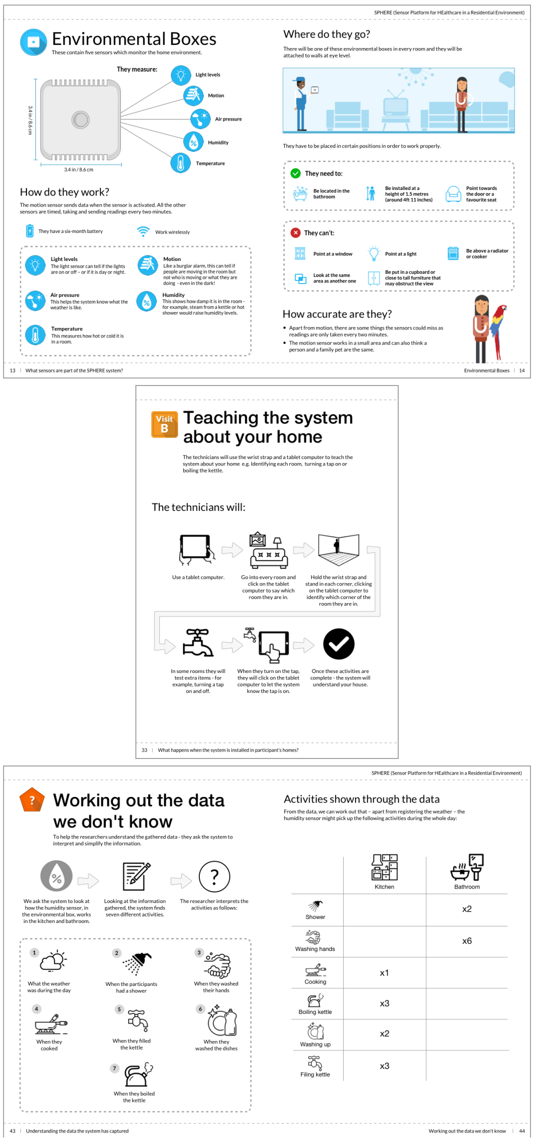

Figure 3: The updated designs showing the platforms content specified by participants from the second study (a) environmental sensor, (b) ground truth and (c) unsupervised learning

2.2.2. Final designs as specified by the participants

Using this feedback, we then updated the page designs (figure 3) to match the participants preferences. For the environmental sensor (figure 3a), we created an illustration to present the sensor placement location and added information about the sensor's limitations as suggested by Cai et al [5]. The 'ground truth' we merged the content that was over two pages in figure 2f to just one page in figure 3c. For 'unsupervised learning', as requested by the participants we merged figure 2h and 2i to highlight the process of collecting and presenting that data. From these final designs, we updated the visual design style and created a number of templates that we used for all similar items (e.g. the SPHERE sensors).

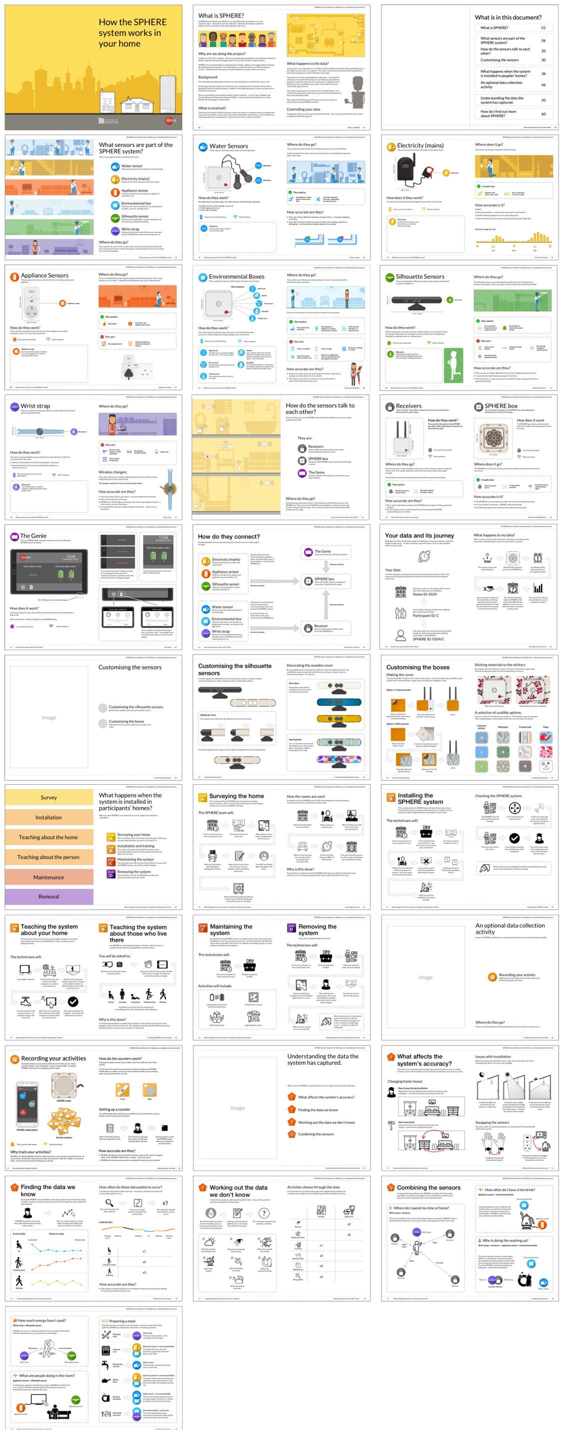

Figure 4: The explanation document used for validation

2.3. Validating the explanation and interpretation

Our next challenge was to validate this explanation document (figure 4) to understand if we had created a translation of the SPHERE platform that potential participants would feel they could use to make an 'informed decision'. Overall, the participants liked the document, all understanding at a high-level the data collected and how that data would be used to identify their daily activity. The participants did ask for a number of updates (e.g. page order, image updates and a reduction of pages within the document) and even though they understood the platform (at a high-level) they wanted to understand why SPHERE was useful to them as individuals with T2D.

3. Next steps

Our next steps are to investigate how we can incorporate the findings from the validation study so that we reduce the number of pages and not just explain the technical aspect of the SPHERE platform but also understand how to explain why this platform would be beneficial to the participants without influencing their decision in consenting to have the platform within their home. Additionally, we wish to investigate the best medium to presenting this content (Paper or video) and understand how this explanation document can work within the first steps of creating a process for the self-installation of the SPHERE platform.

4. Acknowledgements

We would like to thank Sue Mackinnon, Jess Linington, Zoe Banks Gross and Fiona Dowling from Knowle West Media Centre for their support on this project. We would additionally like to thank the SPHERE team members who engaged in this project and for taking their time to explain their work to us.

This work was completed through the SPHERE Next Steps Project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R005273/1.

5. References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Stinne Aaløkke Ballegaard, Thomas Riisgaard Hansen, and Morten Kyng. 2008. Healthcare in Everyday Life - Designing Healthcare Services for Daily Life. 1807–1816.

[3] M Brezis, … S Israel - … Journal for Quality in, and undefined 2008. Quality of informed consent for invasive procedures. *academic.oup.com*. Retrieved December 15, 2020 from https://academic.oup.com/intqhc/article-abstract/20/5/352/1794518

[4] Alison Burrows and Ian Craddock. 2014. SPHERE: Meaningful and Inclusive Sensor-Based Home Healthcare.

[5] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello Ai": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW. https://doi.org/10.1145/3359206

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*: 1721–1730. https://doi.org/10.1145/2783258.2788613

[7] Liya Ding. 2018. Human Knowledge in Constructing AI Systems — Neural Logic Networks Approach towards an Explainable AI. *Procedia Computer Science* 126: 1561–1570. https://doi.org/10.1016/j.procs.2018.08.129

[8] Johanna Glaser, Sarah Nouri, Alicia Fernandez, Rebecca L. Sudore, Dean Schillinger, Michele Klein-Fedyshin, and Yael Schenker. 2020. Interventions to Improve Patient Comprehension in Informed Consent for Medical and Surgical Procedures: An Updated

Systematic Review. *Medical Decision Making 40*, 119–143. https://doi.org/10.1177/0272989X19896348
[9] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? Ml: 1–28. https://doi.org/10.3109/14015439.2012.660499
[10] Alexandra Kirsch. 2018. Explain to whom? Putting the user in the center of explainable AI. *CEUR Workshop Proceedings* 2071. https://doi.org/10.1016/j.juro.2013.04.049
[11] Emily R Lai. 2011. *Critical Thinking: A Literature Review Research Report*. Retrieved December 15, 2020 from http://www.pearsonassessments.com/research.
[12] Susan Lechelt, Yvonne Rogers, and Nicolai Marquardt. 2020. Coming to your senses: Promoting critical thinking about sensors through playful interaction in classrooms. *Proceedings of the Interaction Design and Children Conference, IDC 2020*: 11–22. https://doi.org/10.1145/3392063.3394401
[13] Roger G. Lemaire. 2006. Informed consent - A contemporary myth? *Journal of Bone and Joint Surgery - Series B* 88, 1: 2–7. https://doi.org/10.1302/0301-620X.88B1.16435
[14] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. *IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.1016/j.jsams.2012.02.003
[15] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. *Stakeholders in Explainable AI.*
[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778
[17] Yael Schenker, Alicia Fernandez, and Rebecca Sudore. 2011. Interventions to Improve Patient Comprehension in Informed Consent for Medical and Surgical Procedures: A Systematic Review. *journals.sagepub.com* 31, 1: 151–173. https://doi.org/10.1177/0272989X10364247
[18] Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne Biundo. 2012. Making Hybrid Plans More Clear to Human Users - A Formal Approach for Generating Sound Explanations. *International Conference on Automated Planning and Scheduling*: 225–233. Retrieved from https://www.aaai.org/ocs/index.php/ICAPS/ICAPS12/paper/viewPaper/4691
[19] Diabetes UK. 2020. No Title. https://www.diabetes.org.uk/type-2-diabetes].
[20] Rebecca Voelker. 2018. Diagnosing Fractures With AI. *JAMA* 320, 1: 23. https://doi.org/10.1001/jama.2018.8565
[21] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. *IEEE Conference on Computatonal Intelligence and Games, CIG* 2018-Augus. https://doi.org/10.1109/CIG.2018.8490433
[22] Ni Zhu, Tom Diethe, Massimo Camplani, Lili Tao, Alison Burrows, Niall Twomey, Dritan Kaleshi, Majid Mirmehdi, Peter Flach, and Ian Craddock. 2015. Bridging e-Health and the Internet of Things: The SPHERE Project. *IEEE Intelligent Systems* 30, 4: 39–46. https://doi.org/10.1109/MIS.2015.57
[23] How Machine Learning is Transforming Clinical Decision Support Tools. Retrieved December 14, 2020 from https://healthitanalytics.com/features/how-machine-learning-is-transforming-clinical-decision-support-tools