Design as Common Good

Online Conference | 25-26 March 2021

Legible AI by Design: Design Research to Frame, Design, Empirically Test and Evaluate AI Iconography

Franziska Pilling*[a], Haider Ali Akmal[b],
Adrian Gradinar[c], Joseph Lindley[d], Paul Coulton[e]

[a]Lancaster University, Imagination
[b]Lancaster University, Imagination
[c]Lancaster University, Imagination
[d]Lancaster University, Imagination
*f.pilling@lancaster.ac.uk

Abstract | Artificial Intelligence (AI) is becoming increasingly ubiquitous. Implemented into a wide range of everyday applications from social media, shopping, media recommendations and is increasingly making decisions about whether we are eligible for a loan, health insurance and potentially if we are worth interviewing for a job. This proliferation of AI brings many design challenges regarding bias, transparency, fairness, accountability and trust etc. It has been proposed that these challenges can be addressed by considering user agency, negotiability and legibility as defined by Human Data Interaction (HCD). These concepts are independent and interdependent, and it can be argued, by providing solutions towards legibility, we can also address other considerations such as fairness and accountability. In this design research, we address the challenge of legibility and illustrate how design-led research can deliver practical solutions towards legible AI and provide a platform for discourse towards improving user understanding of AI.

Keywords: Artificial Intelligence, Legibility, Iconography, Digital Workshops, Research through Design

1. Introduction

User adoption of AI, infused into a plethora of products and services operating via the Internet of Things has been expeditious (Lindley et al., 2017). Enabling service providers to monitor in significant detail users behaviour through data (often without explicit consent (Zuboff, 2019)) and subsequently turn this data into decisions and predictions which are increasingly cited as potentially producing harmful results (Angwin, et al., 2016). In an attempt to combat this harm, we have seen a proliferation of frameworks, principles and guidance documents for AI. Of particular note for our research towards legibility is the identified theme of transparency and explainability, which is considered one of the principal challenges needed for AI implementation (Fjeld et al., 2020). Whilst design thinking is cited in many frameworks as a means of potentially addressing AI concerns; it is merely the outlining of problems, rather than providing practical responses. This may be seen as the false promise of design thinking (Kolko, 2018), though in reality, it perhaps reflects the need to articulate better how designers can provide approaches which traverse the current gap between abstract principles and specific implementation. To address this issue, we present research that practically addresses AI legibility through a Research through Design (RtD) enquiry into AI iconography. Taking inspiration from lived experience, with the use of icons to convey effectively important information to a user, we have designed icons to communicate and diffuse the complexity of AI functions to raise user awareness of how AI is operating within the products and services they use. This paper will provide not only the theoretical underpinnings that led to the project and the first designs but also detail the process of iterating the AI icons via a series of workshops using a set of bespoke tools.

The paper proceeds as follows, first by framing AI's relevant pitfalls and the rationale for AI legibility and the role of design towards this end; secondly, a synopsis of designing AI iconography through researching semiotics; thirdly, a summary of empirical testing the prototypical set of AI icons through a series of workshops using bespoke tools; fourthly; an overview of the second iteration of AI icons, designed through analysis of workshop data. In conclusion, we will showcase how a design-led enquiry can respond towards making our relationship with AI more legible and provide a platform for framing the challenge and relevant research landscape for improving user agency.

2. Addressing AI Legibility

An important consideration for this research is to frame what we mean when we say AI and the challenge of legibility. The challenge of AI is socio-technical and therefore requires the complex integration of diverse disciplines, which design is well suited to accomplish. We utilised an RtD approach (Frayling, 1993) as it is generative and geared towards nesting disparate disciplines together (Gaver, 2012) and incorporating various appropriate research ideas, theories and perspectives into design artefacts. Making a user aware, and their interaction with data and AI legible, is a key concern for the field of Human-Computer

Interaction (HCI). Contemporary HCI research is concerned with re-evaluating conventional methods towards designing ‘exploded interfaces’ to provide richer and more tenable inter-relationships with systems as they become ‘smarter’, networked and complex; evolving beyond the traditional duality of interaction between user and computers-as-artefacts (Bowers & Rodden, 1993). The relatively new field of HDI is concerned with recentring the human to explicitly interact with these systems, the data, and the ramifications that transcend from these interactions (Mortier et al., 2015). HDI’s perspective is that data is ontologically malleable and changes depending on the observer. This notion is established via the concept of ‘Boundary Objects’, (Star, 2010) where *things* ‘are both adaptable to different viewpoints and robust enough to maintain identity across them’ (Star & Griesemer, 1989, p. 387). The challenges raised by HDI are organised into three interrelated, though distinct core themes - legibility, agency, and negotiability. Legibility is considered a ‘precursor’ (Mortier et al., 2015) to exercise agency within these systems, where manifestations of agency influence negotiability, enabling a user to build a relationship with those who receive data as means to negotiate how they use data (HDI network, ND).

2.1 The Duality of AI: Mundane vs Sentient Robots and Magic

Commonly and misleadingly AI simultaneously refers to the grand vision of producing a machine with a human level of general intelligence, *as well as* describing a range of real technologies which are in general use today often described as narrow AI. This paradox of misinterpretation between these two divergent, though entangled concepts of AI has been defined as the ‘Definitional Dualism of AI’ (Lindley et al., 2020a).

The *theoretically* straightforward concept of narrow AI (Neural Networks, Expert Systems and Machine Learning) is, in reality, deceptively multifaceted and confused, hindered by the lack of AI legibility and explainability. This misunderstanding is further hampered by the AI found in science-fictions such as the sentient AI cyborg killers in *The Terminator* (1984), and also products falsely claiming to be AI-infused for profitable gains known as AI snake oil. Additionally, AI-infused products are also presented as *magic*, where misleading accounts of AI-technology are deemed as beyond comprehension within the remit of users. Generally, when magic and technology are discussed, Clarke’s third law is often quoted - ‘[a]ny sufficiently advanced technology is indistinguishable from magic’ (Clarke, 1976, p. 21). Clarke’s quote is repeatedly taken out of context; rather, his three laws are meant to express his aspiration for humanities technological endeavours. However, the misperception of technology echoes the statement in Brackett’s short story *The Sorcerer of Rhiannon* - ‘Witchcraft to the ignorant, … simple science to the learned’ (1942). Concerning AI technology, the user is not ignorant by choice, as there is currently very little in the way in which users can legibly understand when AI functions are being performed.

There are more pitfalls that besiege AI, and subsequently, its users and those affected by governing algorithmic decisions. Individually these challenges are too expansive to unpack them all in this paper and go beyond the scope of this research. An important point for this research is that AI reflects the coding and data they are trained on which, are often societally biased and inaccurate (Angwin, et al., 2016; O'Neil, 2016; Suresh & Guttag, 2020). A common misconception is that AI systems are free from human influence and therefore, bias. However, humans are always part of the system, a contagion, if you will, spreading disturbances in several ways from feature extraction, data curation to providing oversight on algorithmic outputs (Lindley & Coulton, 2020). As Turing speculated computers are limited by our instructions, codes and given structures, meaning that they would essentially share and also create blind spots in logic (Turing, 1938). To this end, AI is often cited as a black box (BB). Latour described the notion of Black Boxing as; '[t]he way scientific and technical work is made invisible by its own success' (Latour, 1999, p. 304). This emulates Stahl's findings that '[w]hen a technology is a black box it becomes magical' (Stahl, 1995, p. 252). AI is a BB for both its users and its creators; where some AI-experts state that they are not sure how AI-systems reach an output, as AI based on Machine Learning is coded to exponentially expand through training data and its interrelation with thousands of weights and variables, eventually evolving beyond human intelligibility and accessibility. Additionally, experts cannot explain the algorithms used to create the AI in the first place, where Rahimi (former Google researcher) stated that 'machine learning has become alchemy' (Elish & Boyd, 2018), arguing that even though alchemy 'worked' the foundations of alchemy were formed upon unverifiable and for modern times dubious theories. To this end, our interest in design research is principally concerned in contemporary, functional and practical uses of AI.

Creating transparent AI systems is repeatedly called for to oppose the BB nature of current AI systems as well as the legibility and explainability of these systems. These terms are used almost interchangeably, though they describe subtly different things. Transparency is concerned with how open the data and algorithms are to outsider auditing to be verified or challenged. In comparison, legibility and explainability are similar and focused on how we can make AI systems and their decisions understandable and readable to non-AI experts. Making a system transparent does not equate to making it legible or explainable, where explainability can come from the legibility of a system via more appropriate metaphors and as this research will show - iconography.

3. Researching AI legibility through Design

To address the existing illegibility of AI, we started with a survey of current AI imagery by searching icon and stock image repositories. What we found was that while some icons represent the underlying system such as neural networks (see figure 1a) and some might suggest what it's doing such as face detection (see figure 1b), the vast majority of icons play into AI's definitional dualism of human-like machines (see figure 1c and 1d). With closer inspection, the existing imagery seldom articulates how an AI would function and in what context, or if it did it would raise more questions than answers (e.g. see figure 1a, does this network have three layers, is it adaptive?). Furthermore, no imagery articulated the ramifications or implications of use.

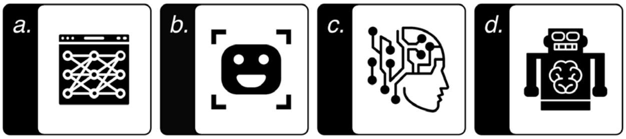


Figure 1. Typical examples of current AI Iconography.

3.1 Establishing the Semiotics of AI

The lack of semantics or communication within current AI imagery indicates there is scope to develop a visual language which would help to enhance AI legibility (Lindley et al., 2020a). Research into the design, theory and effectiveness of icons in the field of HCI is diverse underpinned by the theory of Semiotics, for instance - semiotic analysis for user interfaces (Ferreira et al., 2002), icon taxonomy to categories computer icons (Ma et al., 2015), the advantages and disadvantages of icon based dialogues in HCD (Gittins, 1986), relationships between different presentation modes of graphical icons and user attention (Lin et al., 2016), testing the intuitiveness of icons (Ferreira et al., 2006). Iconography has proven to be a useful tool for encapsulating the complexity of a particular interaction for users so that they know how it works, and thereafter infer implications of said interaction. Influenced by how semiotics can help designers improve their communication (de Souza et al., 2001) and aligning with HCI scholars (Ferreira et al., 2002) we have referred to the renowned theory the Peircean triad (Peirce, 1991). Peirce's model (see figure 2) consists of a triadic relationship; comprising the *representamen* (the symbol used to represent an idea, e.g. a save icon), the *object* (the actual construct being represented, e.g. data or document being *saved*), and the *interpretant* (the logical implication of the sign, e.g. using this icon will save my data). Central to Peirce's thesis is the 'classification of signs' which is based on the relationship between object and representamen; these categories are; *indexical,* signs which refer to the object indirectly, through an association (e.g. smoke signifies fire), *symbolic* signs which have meaning based solely on convention and may be culturally specific, such as alchemy symbols (e.g. a triangle to represent fire); *iconic* signs have a signifier which resembles the signified package (e.g. flames pictorial).

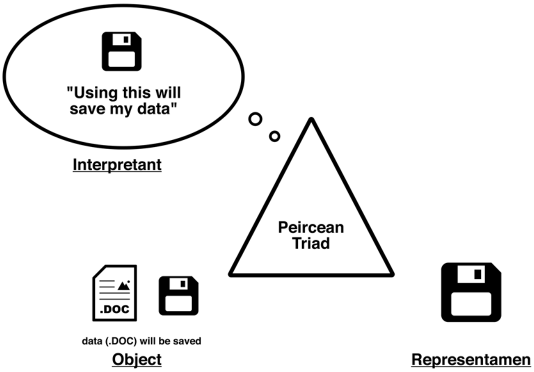


Figure 2. A triadic view of the save icon (the document icon used here will have its own triadic view).

Peirce noted that categories are not mutually exclusive, as most signs contain elements of indexicality, symbolism and iconicity in varying degrees. Taking this theory back to the analysis of the existing AI icons (see figure 1c and 1d) the representamen forces the notion of AI's dualism, which we could argue is 'misleadingly' both categories of symbolism and iconicity. In some cases, the interpretant functions to some degree (see figure 1b facial recognition); however, there is no understanding of the greater AI system (Lindley et al., 2020b). There is clearly a distortion in AI communication where categories are used misleadingly through the lack of conventions for AI and as previously outlined cultural understanding.

Timeframe for designing icons is also an essential consideration as symbols can change meaning over time by appropriation or other means, which pointed towards taking inspiration from an already functional system of icons - laundry care labels. Whilst we may not always take notice of these icons, or indeed always understand their meaning, they provide a means of understanding how we can most easily maintain a working relationship with our clothes standing the test of time. This archetypal iconography system further influenced us into considering that multiple icons could be used together to form a language of interaction, suiting the complexity of the issues that confound AI legibility.

3.2 Generative Designing of AI Iconography

Consolidating and reviewing prior AI research, we identified six key AI factors to package into a system of AI iconography, whereby effectively communicating AI functions and operations for user legibility (see figure 3). These relevant factors for AI legibility are as follows;
1.Presence - denoting that some form of AI processing is happening heeding to the principle

of ‘informed use’ (EPRS, 2020) (see figure 3a). 2.Processing Location – in the cloud, on the edge or elsewhere. The location of processing impacts users perception of accountability (Rader et al., 2018)(see figure 3b). 3. Learning Scope – how does the AI learn or adapt over time, through usage or is it static? Communicating to a user changes and adaptions of an AI system is deemed a fundamental guideline for human-AI interaction (Amershi et al., 2015)(see figure 3c). 4. Data Provenance – What is the source of the training data? Is it proprietary, public or the user? Data quality directly reflects the AI and therefore, its trustworthiness (Arnold et al., 2019)(see figure 3d). 5. Training Data types – what data types are used to train the AI? Visual, audio, location? Similar to data Provence this factor is more granular account on the type of data, which is a crucial element to reduce opacity (Burrell, 2016) increase trust (Arnold et al., 2019) and reduce bias (Angwin, et al., 2016; O’Neil, 2016)(see figure 3e). 6. Intrinsic Labour – is ‘work’ being done for the AI operator. This is more of a philosophical and discursive factor, as it reflects the monetisation of data through the commodification of users and their interactions with AI-infused products and services (Greengard, 2018; Zuboff, 2019).

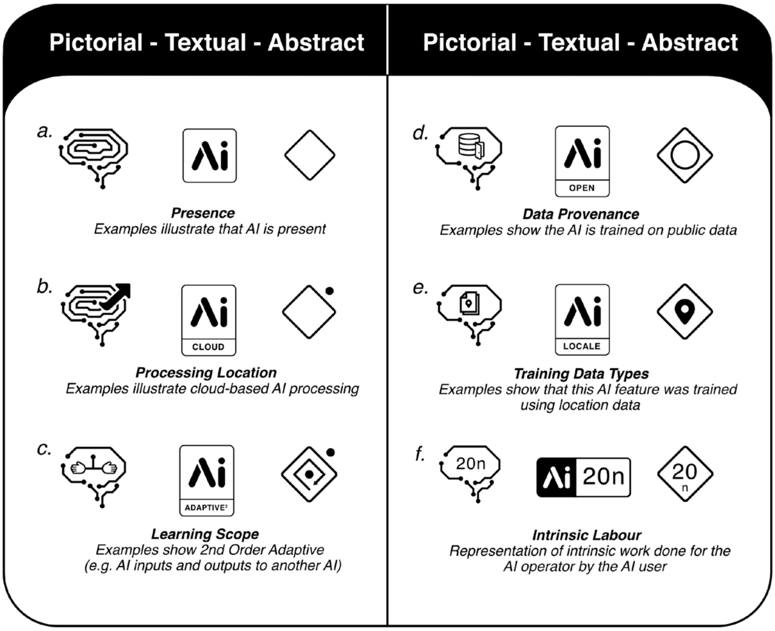

Pictorial - Textual - Abstract
Pictorial - Textual - Abstract
a.
Ai
Presence
Examples illustrate that AI is present
b.
Ai
CLOUD
Processing Location
Examples illustrate cloud-based AI processing
c.
Ai
ADAPTIVE²
Learning Scope
Examples show 2nd Order Adaptive
(e.g. AI inputs and outputs to another AI)
d.
Ai
OPEN
Data Provenance
Examples show the AI is trained on public data
e.
Ai
LOCALE
Training Data Types
Examples show that this AI feature was trained using location data
f.
20n
Ai 20n
20 n
Intrinsic Labour
Representation of intrinsic work done for the AI operator by the AI user

Figure 3. The identified AI factors arranged in different visual styles.

Merging the AI research with semiotics, we designed three different visual styles (see figure 3 Pictorial, Textual and Abstract). The pictorial concept which deliberately utilises the sort of iconography resulting from AI’s dualism. Though despite conforming to the current problem, unexpectedly iconic imagery emerged and established a baseline to move forward. The textual concept explored the use of a brand identity inspired by the symbology employed by trade organisations. However, we theorised that there is a limit to how much textual information can be gleaned in a single instance. The abstract concept as per Peirce's thesis hybridises symbolic, indexical and iconic categories equivalent to laundry labels. We elected to develop the abstract concept further (see figure 4), minimising the problematic aspects of the former two concepts and the flexibility to incorporate iconic imagery into the abstract style (see figure 3e and figure 4 Data Types). As with laundry labels, a degree of a convention is necessary to understand these abstract icons, though once core elements are deciphered, such as triangle denoting learning (see figure 4 Learning scope), readability begins to emerge.

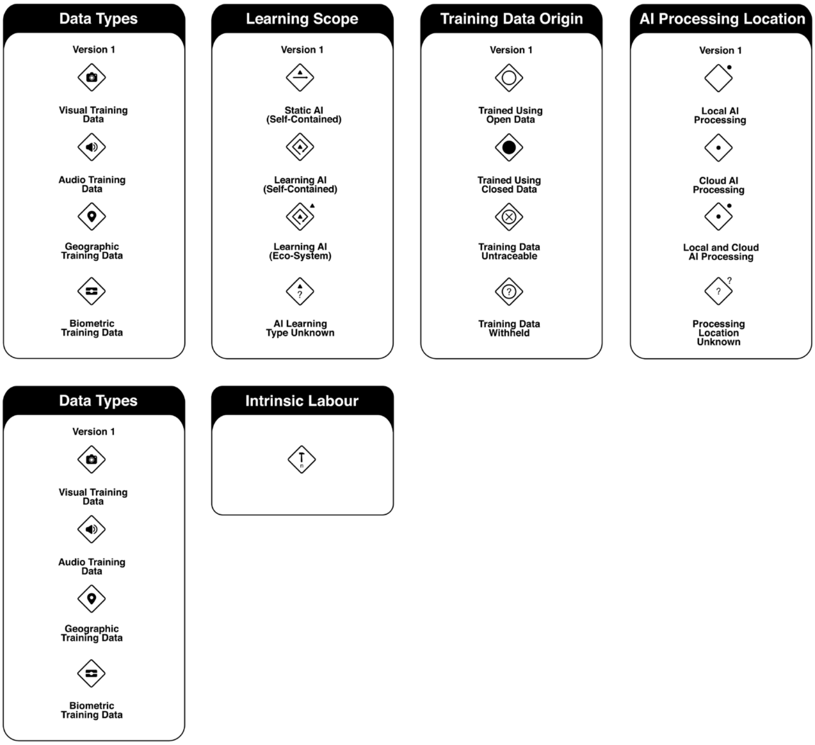


Figure 4. Version 1 iconography.

4. Empirical testing and Co-Designing

The difficulty in predicting how or why an icon may become adopted or stay in use supports the premise that an RtD investigation of using icons to improve AI legibility is a productive first step. Barr, Biddle and Noble (2002) state an icons' successfulness' is guaranteed if the user matches the interpretant to the intended object, concept or implication. To determine if the interpretant matches the designer's intention is through performing icon intuitiveness, and usability tests (Ferreira et al., 2006), by asking participants to match AI functions to the icons in a set.

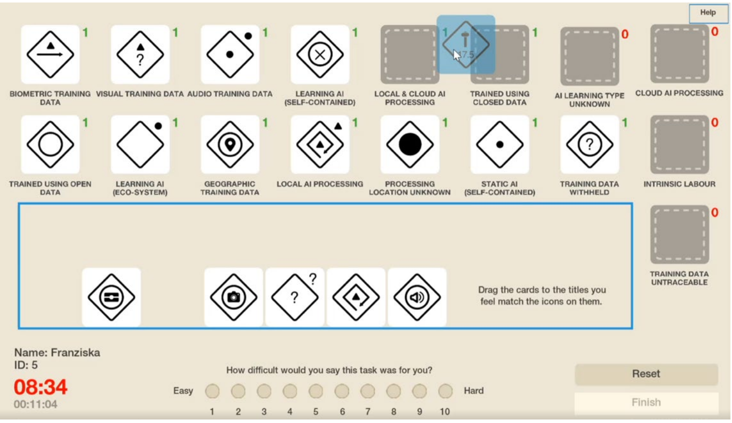
4.1 Uncanny AI to Legible AI Workshops

To test our prototypical AI icons, we presented them in the form of physical cards with separate matching card descriptors. Participants were encouraged to intuitively match and establish connections between the defined AI functions, and their visual representations. The deck of cards as a playful medium, allowed participants to engage with, in a tangible manner, the intangible functions and operations of AI by completing a series of stylised game-like exercises, which were designed to not only test icon intuitiveness but also enter a discourse to question the legibility within AI systems. There were no rules as to what the participants could do with the cards in the first *Making Connections* exercise, leaving it open for the participants to intuitively start to figure out, cluster and pair with their associated descriptor cards.

The workshop was designed to take 'a wide and playful view' (Gaver, 2002) of AI technology and its implications. As such the workshop was consciously designed to instil traces of play through the range of exercises by adopting Gaver's stance on 'playful curiosities', as 'play is … an essential way of engaging with and learning about our world' (Ibid,2002). The second exercise called *What's in my AI*, tasked participants to predict the AI functions that would characteristically occur in randomly presented AI-infused products and generate an icon map using our AI icons. Encouraging an attitude of speculation and ambiguity builds a space for participants to 'intermesh' their own experiences of AI. Fortuitously, the icons unexpectantly developed into a form of pedagogical tool for learning about AI functions, with participants leaving the workshop with a greater critical awareness of AI technology.

4.2 Bespoke Tools for Distance Empirical Testing

Through the pandemic, we adapted the workshop to a digital counterpart. Rather than sourcing an online tool to support what was a face to face workshop; we instead developed an interactive workshop to suit our research medium and replicate the game like mechanics of the physical workshop. To reproduce the playful interactions steered us to use *Godot*, an open-source game-engine, with each task built and coded as a 2D game 'scene' with a Graphical User Interface with flexible components (see figure 5). These components were the icons cards that could be coded with rules in multiple ways, from game physics to movement dependant on the exercise. This transition to digital required an experimental approach in making, while also considering customary design considerations from 'user-onboarding' experience, 'paths of interaction' (Verplank, 2009) and affordances (Norman, 1999). Using the game-engine afforded the opportunity to quickly build, test and make changes and add supplementary exercises that embodied the physical interactions that originally happened spontaneously within exercises, which on reflection were areas of research and further investigation. In particular, blank 'playing' cards allowed participants to design icons if they felt we had missed representing any AI factors. In the digital workshop, this translated into the exercise *Draw Your Own*, where participants were presented with a digital canvas and drawings tools, reminiscent of *Microsoft Paint*, to design and present their icons, which would later be analysed to help develop the second iteration of icons to avoid 'designs becom[ing] purely self-indulgent' (Gaver, 2002).



Drag the cards to the titles you feel match the icons on them.

Name: Franziska
ID: 5
08:34
00:11:04

How difficult would you say this task was for you?

Figure 5. Screenshot from Making Connections exercise. Participants drag and drop the digital icon cards to a matching descriptor.

The digital workshop can collect a large volume of data, creating the possibility of too many data points to pick from. Consequently, we devised a data analyser (see figure 6) through a combination of PHP and MySQL API, a standard implementation for querying and storing information on web-based platforms. The data from the workshop could then be visualised immediately after each exercise. The responses were organised and structured in a predetermined fashion for quickly articulating participants responses, which was used as a tool to analyse the responses together live in the workshops, thereby promoting richer conversations. Moving to an online format has permitted rigorous testing of the icons with a global and divergent audience, perhaps creating a more holistic empirical test for icon intrusiveness.

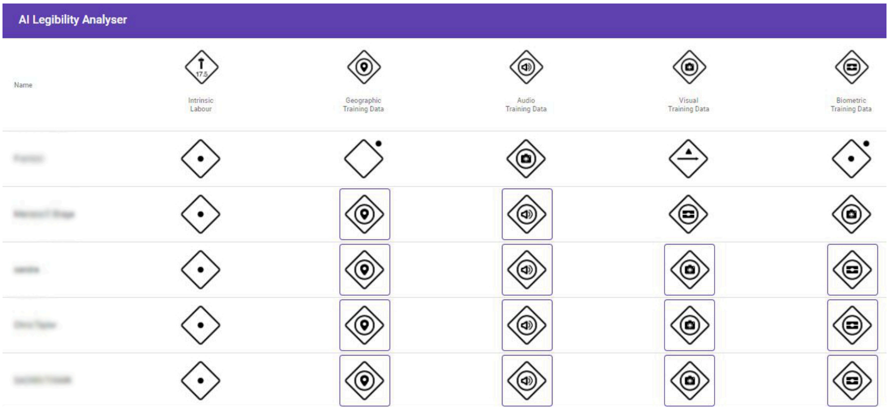
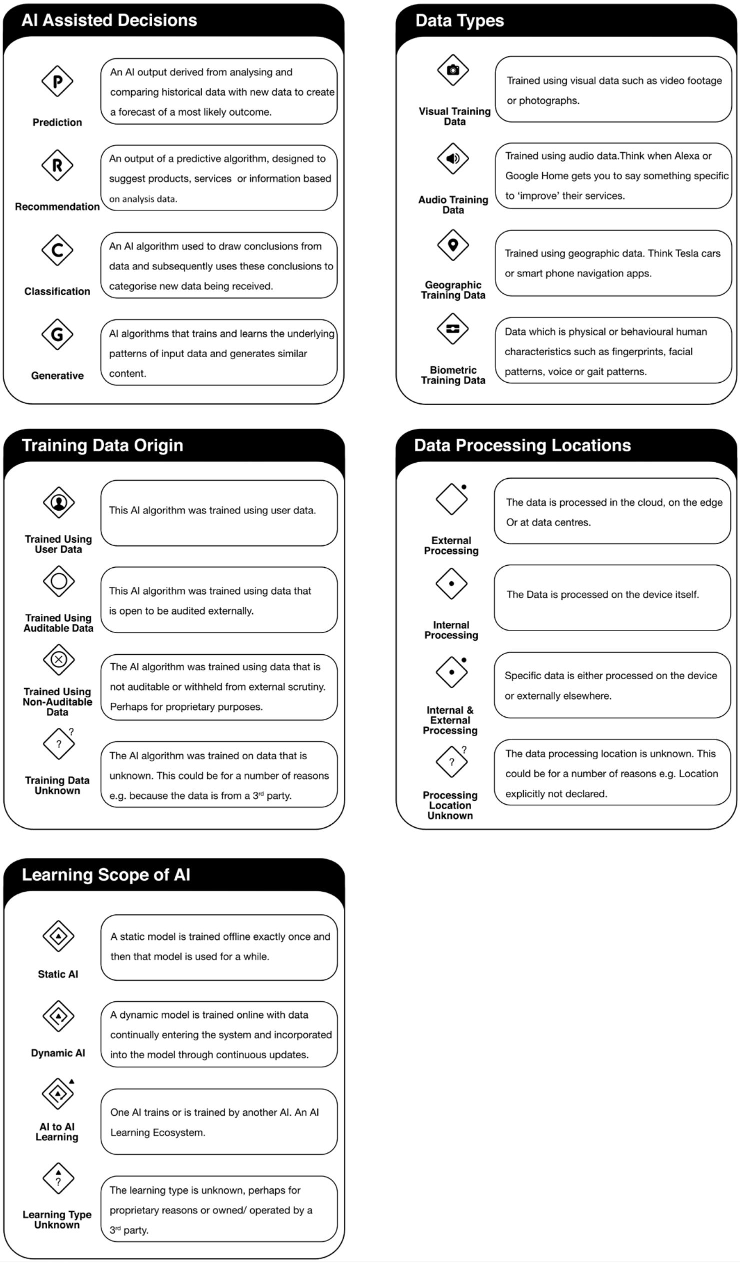


Figure 6. A sample of results from Making Connections. The top row is the icons and descriptor. Below are the participants matches. The bounding box quickly identifies which icons were matched correctly.(Identities intentionally obscured).

5. Analysis, Second iteration and Beyond

It was apparent in the analysis of data that this RtD-based inquiry was successful to some degree at producing an ‘intuitive’ set of AI icons. Though, as predicted, the icons that embodied more symbolic (see figure 4. Learning Scope, triangle to symbolise learning) or indexical (see figure 4. Learning Scope, rotational arrow signifies a round of training) categories were less intuitive. Participants who ‘correctly’ read the icons did it through a systematic process of non-verbal reasoning, grouping clusters of icons together, whereby the first icon had to be correctly placed for the rest to follow suit. Reflecting back, laundry icons are introduced in small digestible additions over time with the arrival of new laundry technology and instructions. The AI icons were introduced simultaneously and are communicating more complex and fluctuating concepts. Though, as identified from the workshop discussions, the icons offered a sign towards ‘more is happening here’, and over time the icons could be ‘learnt like road signs and the highway code’. This might seem like a bias interpretation of our results; however, in keeping with the RtD methodology we adopt, these findings are ‘contingent and aspirational’, manifesting into research which ‘creatively challenge[s] status quo thinking’ (Gaver, 2012). That is not to say that the results are not useful; moreover, they serve as conceptually rich research artefacts subject to ongoing AI developments.



AI Assisted Decisions

Prediction: An AI output derived from analysing and comparing historical data with new data to create a forecast of a most likely outcome.

Recommendation: An output of a predictive algorithm, designed to suggest products, services or information based on analysis data.

Classification: An AI algorithm used to draw conclusions from data and subsequently uses these conclusions to categorise new data being received.

Generative: AI algorithms that trains and learns the underlying patterns of input data and generates similar content.

Data Types

Visual Training Data: Trained using visual data such as video footage or photographs.

Audio Training Data: Trained using audio data.Think when Alexa or Google Home gets you to say something specific to 'improve' their services.

Geographic Training Data: Trained using geographic data. Think Tesla cars or smart phone navigation apps.

Biometric Training Data: Data which is physical or behavioural human characteristics such as fingerprints, facial patterns, voice or gait patterns.

Training Data Origin

Trained Using User Data: This AI algorithm was trained using user data.

Trained Using Auditable Data: This AI algorithm was trained using data that is open to be audited externally.

Trained Using Non-Auditable Data: The AI algorithm was trained using data that is not auditable or withheld from external scrutiny. Perhaps for proprietary purposes.

Training Data Unknown: The AI algorithm was trained on data that is unknown. This could be for a number of reasons e.g. because the data is from a 3rd party.

Data Processing Locations

External Processing: The data is processed in the cloud, on the edge Or at data centres.

Internal Processing: The Data is processed on the device itself.

Internal & External Processing: Specific data is either processed on the device or externally elsewhere.

Processing Location Unknown: The data processing location is unknown. This could be for a number of reasons e.g. Location explicitly not declared.

Learning Scope of AI

Static AI: A static model is trained offline exactly once and then that model is used for a while.

Dynamic AI: A dynamic model is trained online with data continually entering the system and incorporated into the model through continuous updates.

AI to AI Learning: One AI trains or is trained by another AI. An AI Learning Ecosystem.

Learning Type Unknown: The learning type is unknown, perhaps for proprietary reasons or owned/ operated by a 3rd party.

Figure 7. AI icons Version 2 with indepth descriptions. Note the AI Assisted Decisions are textual however, as singular letters which can quickly be read and infer meaning.

Figure 7 introduces the second iteration of the icons developed from the analysis of the first series of workshops. In this version, the significant development is the introduction of a new AI factor – AI Assisted Decisions. While the supplementary AI factors serve more as building blocks of the system, the overall inference and immediate implication of the AI was not accounted for. This notion was deduced from many of the participants expressing that they just wanted the surface level of information – 'why is AI being used?'.Through discussions, this idea was speculated further towards designing a hierarchical system of icons (see figure 8), with 'Presence of AI' at the top collapsing down towards the more 'technical' AI factors. A type of vocabulary logic within the iconography system for users to make their own value-judgements and take into account what's important to them, rather than a proscribed qualitative assessment. Deciding the order of hierarchical system will be an exercise in the next series of workshops to test the new icon iteration where we will ask participants 'what's important to you as a user'.

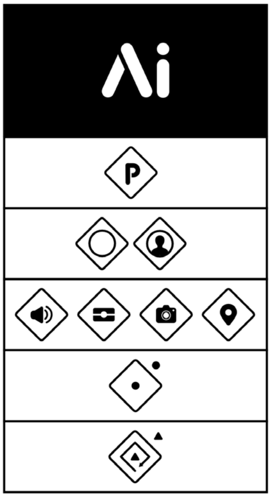
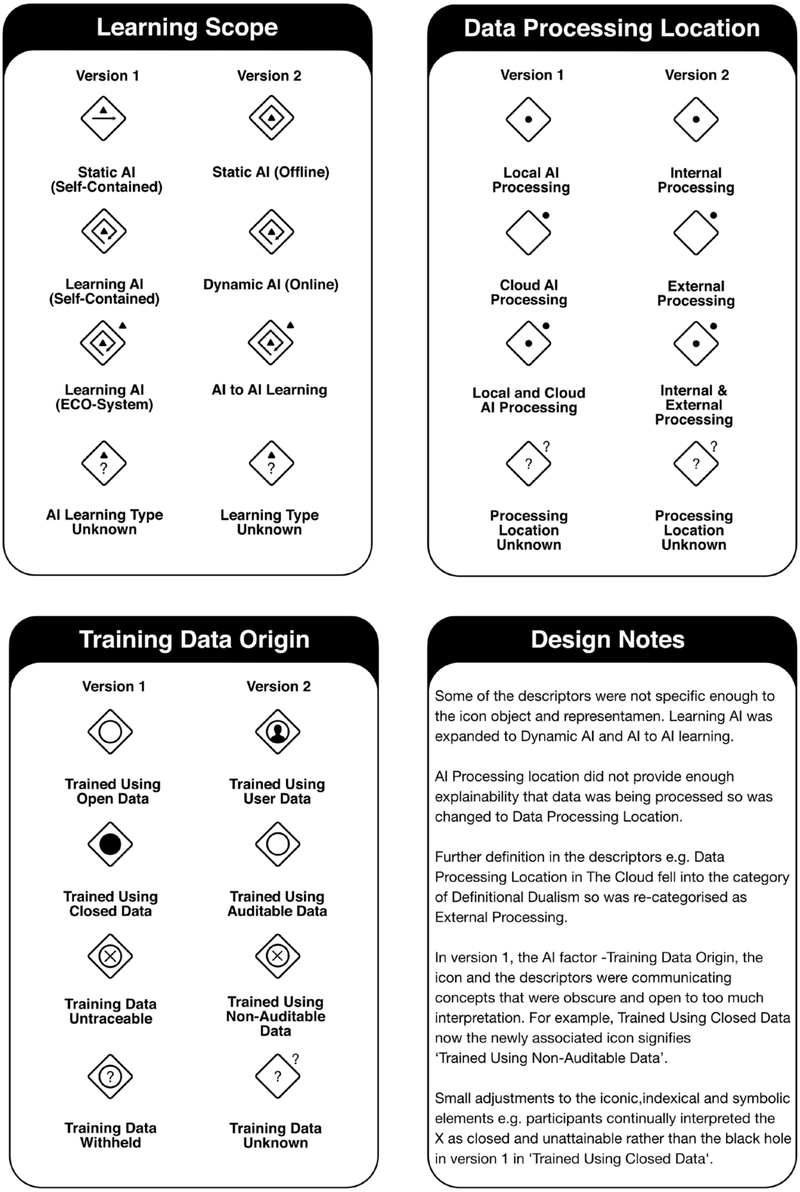


Figure 8. The icons lend themselves to be modularised into a hierarchical structure, forming a AI 'ontological' language.

Overall slight adjustments were made to the iconography set and icons that were not intuitive and therefore not 'successful'. To illustrate Static AI in the first iteration (see figure 9) was presented with a triangle signifying training and an arrow moving in a forward direction, obverse to the notion of static and therefore Static AI, where the model is trained once and used. Small adjustments are further pointed out in figure 9 with accompanying explanations and design thinking. Further empirical testing of the second iteration of icons will continue the generative process of designing intuitive icons for AI legibility, in an attempt of establishing an 'equilibrium' within the icons.



Some of the descriptors were not specific enough to the icon object and representamen. Learning AI was expanded to Dynamic AI and AI to AI learning.

AI Processing location did not provide enough explainability that data was being processed so was changed to Data Processing Location.

Further definition in the descriptors e.g. Data Processing Location in The Cloud fell into the category of Definitional Dualism so was re-categorised as External Processing.

In version 1, the AI factor -Training Data Origin, the icon and the descriptors were communicating concepts that were obscure and open to too much interpretation. For example, Trained Using Closed Data now the newly associated icon signifies 'Trained Using Non-Auditable Data'.

Small adjustments to the iconic,indexical and symbolic elements e.g. participants continually interpreted the X as closed and unattainable rather than the black hole in version 1 in 'Trained Using Closed Data'.

Figure 9. A direct comparison between version 1 and version 2 of the icons. The accompanying design notes highlight some of the thinking behind the changes.

6. Conclusion

The research presented here is not intended to conclude or solve the problem of AI legibility, but rather articulate and triangulate the reality of AI's challenges, the diverse AI research landscape and designs role in improving AI legibility. In summary, the contributions of this paper are as follows. First, we framed AI legibility and the challenge to establish this, in doing so we pinpointed possible factors that aided in obscuring AI. This highlighted the cross-sectoral and interdisciplinary perspectives and research required to face the challenges imposed by AI. Next, we described a design-led response for legible AI by providing a synopsis of our RtD-enquiry, with a reflective account of designing AI iconography and empirically testing their intuitiveness. We also gave an overview of creating a bespoke research tool to actively continue the research during the pandemic, offering another branch of research into building workshops online via game-engines. Although this contribution may seem beyond the scope of the research, we have included it here as it demonstrates the generative and aspirational qualities afforded through an RtD methodology and the freedom of following the research where it takes you. Finally, we pinpointed the next research phase of testing our second iteration of AI icons.

AI adoption is widespread, obscured by its own success, and subsequent lack of knowledge grounded in the reality of these devices, used for socially consequential classifications, which 'valorises some point of view and silences another' (Bowker & Star, 1999). There are currently no supportive and standardised ways of communicating the 'shapeless and faceless, everywhere and nowhere' (Pierce & DiSalvo, 2017) constructs of AI. Often all that users have to work with is metaphors that confuse the reality of AI technology. Advocation for 'interactive explanation systems' (Weld & Bansal, 2019) is in high demand, evidenced in the diverse authored frameworks and guidelines for future AI implementation. Design research similar to ours can strive for accessible and more empowering methods of describing technology and its working parameters for users.

References

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 337–346. https://doi.org///doi.org/10.1145/2702123.2702509

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., & Varshney, K. R.

(2019). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *ArXiv:1808.07261 [Cs]*. http://arxiv.org/abs/1808.07261

Arthur C, C. (1976). *Profiles of The Future*.

Barr, P., Noble, J., & Biddle, R. (2002). *Icons R Icons: User interface icons, metaphor and metonymy* (CS-TR-02/20). Victoria University of Wellington School of Mathematical and Computing Sciences Computer Science.

Bowers, J., & Rodden, T. (1993). Exploding the interface: Experiences of a CSCW network. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 255–262. https://doi.org/10.1145/169059.169205

Bowker, G., & Star, S. L. (1999). *Sorting Things Out Classification and Its Consequences*. The MIT Press.

Brackett, L. (1942). The Sorcerer of Rhiannon. In C. John W (Ed.), *Astounding Science Fiction* (6th ed., Vol. 28, pp. 36–48). Street & Smith Publications, Inc.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

de Souza, C. S., Barbosa, S. D. J., & Prates, R. O. (2001). *A Semiotic Engineering Approach to HCI*. 55–56.

Elish, M. C., & Boyd, D. (2018, November 13). Don't Believe Every AI You See [Research]. *Don't Believe Every AI You See*. https://ai.shorensteincenter.org/ideas/2018/11/12/dont-believe-every-ai-you-see-1

EPRS. (2020). *The ethics of artificial intelligence: Issues and initiatives* (STUDY PE 634.452; Panel for the Future of Science and Technology). Scientific Foresight Unit. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

Ferreira, J., Barr, P., & Noble, J. (2002). The Semiotics of User Interface Redesign. *Proceedings of the Sixth Australasian Conference on User Interface*, *40*, 47–53.

Ferreira, J., Noble, J., & Biddle, R. (2006). *A Case for Iconic Icons. 50*, 87–90.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Frayling, C. (1993). *Research in Art and Design*. *1*.

Gaver, W. (2002). DESIGNING FOR HOMO LUDENS. *13 Magazine*.

Gaver, W. (2012). What should we expect from research through design? *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 937. https://doi.org/10.1145/2207676.2208538

Gittins, D. (1986). Icon-based human-computer interaction. *International Journal of Man-Machine Studies, 24*(6), 519–543. https://doi.org/10.1016/S0020-7373(86)80007-4

Greengard, S. (2018). Weighing the impact of GDPR. *Communications of the ACM*, *61*(11), 16–18. https://doi.org/DOI:https://doi.org/10.1145/3276744

HDI network. (ND). What is Human-Data Interaction? [HDI network]. *HDI Network*. https://hdi-network.org/intro_to_hdi/

Kolko, J. (2018, June). THE DIVISIVENESS OF DESIGN THINKING. *Interactions*, *25*, 28.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.

Lin, H., Hsieh, Y.-C., & Wu, F.-G. (2016). A study on the relationships between different presentation modes of graphical icons and users' attention. *Computers in Human Behavior*, *63*, 218–228. https://doi.org/10.1016/j.chb.2016.05.008

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020a). Researching AI Legibility through Design. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 13. http://doi.acm.org/10.1145/3313831.3376792

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020b). Signs of the Time: Making AI Legible. *Proceedings of Design Research Society Conference 2020*. DRS 2020, Australia. https://doi.org/10.21606/drs.2020.237

Lindley, J., & Coulton, P. (2020). *AHRC Challenges of the Future: AI &amp; Data*. https://doi.org/10.13140/RG.2.2.29569.48481

Lindley, J., Coulton, P., & Sturdee, M. (2017). Implications for Adoption. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 265–277. https://doi.org/10.1145/3025453.3025742

Ma, X., Matta, N., Cahier, J.-P., Qin, C., & Cheng, Y. (2015). From action icon to knowledge icon: Objective-oriented icon taxonomy in computer science. *Displays*, *39*, 68–79. https://doi.org/10.1016/j.displa.2015.08.006

Mortier, R., Haddadi, H., Henderson, T., McAuley, D., & Crowcroft, J. (2015). Human-Data Interaction: The Human Face of the Data-Driven Society. *ArXiv:1412.6159 [Cs]*. http://arxiv.org/abs/1412.6159

Norman, D. (1999). Affordance, conventions, and design. *Interactions*, *6*(3), 38–43. https://doi.org/10.1145/301153

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Peirce, C. S. (1991). On a New List of Categories. In J. Hoopes (Ed.), *Peirce on Signs* (pp. 23–33). University of North Carolina Press; JSTOR. http://www.jstor.org/stable/10.5149/9781469616810_hoopes.7

Pierce, J., & DiSalvo, C. (2017). Dark Clouds, Io&#!+, and [Crystal Ball Emoji]: Projecting Network Anxieties with Alternative Design Metaphors. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 1383–1393. https://doi.org/10.1145/3064663.3064795

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173677

Stahl, W. A. (1995). Venerating the Black Box: Magic in Media Discourse on Technology. *Science, Technology, & Human Values*, *20*(2), 234–258. https://doi.org/10.1177/016224399502000205

Star, S. L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, *35*(5), 601–617.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420.

Suresh, H., & Guttag, J. V. (2020). A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv:1901.10002 [Cs, Stat]*. http://arxiv.org/abs/1901.10002

Turing, A. (1938). *Systems of Logic Based on Orinals*. Princeton University.

Verplank, B. (2009). *Interaction Design Sketchbook*. http://www.billverplank.com/IxDSketchBook.pdf

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, *62*(6), 70–79. https://doi.org/10.1145/3282486

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books Ltd.

About the Authors:

Franziska Pilling is a PhD Design Candidate, funded by the PETRAS IoT hub, researching design's role in making algorithmic intelligence and its associated systems, processes and misconceptions, more legible to users and designers through alternative practices such as Speculative Design with Philosophy.

Haider Akmal is a PhD Candidate and Research Associate at Lancaster University, UK. His PhD thesis discusses the use of Play and Philosophy within Design Research for imagining More-than Human Centred Design approaches through Speculative and Ludic Design.

Adrian Gradinar is a Lecturer in Smart Home Futures focusing on speculative practice-based approaches to exploring ideas around interactivity, personalisation, artificial intelligence, data privacy and transparency, immersion, more-than-human design approaches and better design of Internet of Things objects and spaces.

Joseph Lindley is a Research Fellow interested in how Design Research can contribute towards radical-yet-responsible applications of contemporary technologies including Artificial Intelligence and the Internet of Things.

Paul Coulton is the Chair of Speculative and Game Design in the open and exploratory design-led research studio Imagination Lancaster. He uses a research through design approach to create fictional representations of future worlds in which emerging technologies have become mundane.

Acknowledgements: This research has been supported through the Ucanny AI project funded through the PETRAS National Centre for Excellence in Cybersecurity EPSRC project EP/S035362/1